Kaggle Project: Covid-19 Awareness and Covid-19 Cases in Ohio

Ayonga Wakwabubi Eugene[1]

Abstract— **In this study, the level of awareness on Covid-19-related topics in Ohio counties during the pandemic was analyzed, using a dataset consisting of social, economic, demographic variables, and Covid-19 cases and deaths. The awareness data, extracted from over 46 million tweets, were processed using co-occurring hashtags and similarity measures to quantify the intensity of discussion on Covid-19 topics. XGBoost's XGBRegressor was used to predict the level of awareness, fine-tuning hyper parameters to achieve an Adjusted R2 score of 0.9604 and an R2 score of 0.87488 on the Kaggle competition. This research is important as it provides insights into public awareness during a health crisis, which can inform targeted interventions and policy-making.**

I. INTRODUCTION

This study aimed to predict the number of Covid-19 cases in Ohio counties by leveraging social, economic, demographic, and awareness-related factors. The dataset utilized in this research included measurements of COVID-19 awareness, derived from over 46 million tweets posted by more than 91,000 users in Ohio, along with data on COVID-19 cases and deaths, and other relevant county-level variables. In this unique machine learning challenge, the test set comprises 70% of the total observations, while the training set accounts for the remaining 30%. Specifically, there are 3,141 observations in the training set and 7,331 observations in the test set.

A moderate R-squared value of around 0.5 was anticipated as a benchmark for the model's performance. My approach involved several stages: data pre-processing, dimensionality reduction, feature selection, model training, and evaluation. Spectral embedding, a dimensionality reduction technique, was employed to condense the feature space while retaining key information. An XGBoost regression model was chosen and optimized through hyper parameter tuning to achieve the best possible results. Two of the most significant findings were:

Model Performance: The model achieved an adjusted R2 score of 0.9604 on the test dataset, and an R2 score of 0.87488 on Kaggle indicating a strong predictive capability. This score surpasses the minimum benchmark of 0.5 set for the analysis, demonstrating the effectiveness of the chosen modeling approach and feature engineering techniques.

Awareness and Case Numbers: The analysis revealed a noteworthy relationship between the level of COVID-19 awareness in Ohio counties and the number of reported cases. Counties with higher awareness levels, as indicated by the intensity of discussion on COVID-19-related topics on social media, tended to have a higher number of reported cases. This finding underscores the importance of public awareness in managing the spread of the virus.

II. DATA

A. Topic Awareness Variables

The training dataset had a shape of 3141 rows and 144 columns, and had no missing values. Of the 144 columns, 18 comprised of the features columns. The average value for all topic awareness variables was graphed as shown below:

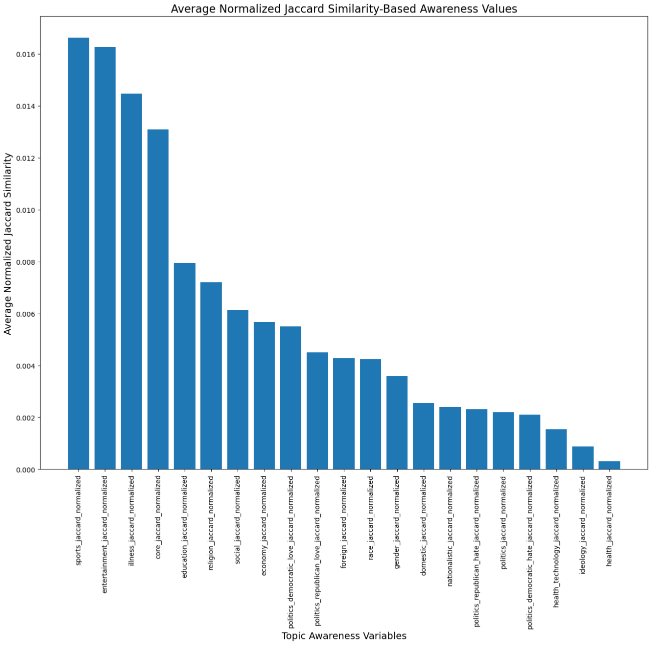

Fig. 1. Average Normalized Jaccard Similarity-Based Awareness Values

In reference to the graph above, the variables are ranked from highest to lowest average normalized Jaccard similarity. Sports-related topics have the highest average similarity value, indicating possibly the most shared awareness or commonality in discussions or data points among the categories shown. This is followed by entertainment, illness, and education, which also have relatively high awareness values. Topics related to politics, healthcare, and technology, among others, have lower average similarity scores, suggesting less commonality or shared awareness within those areas. The category "health" has the lowest average similarity value, which might reflect a more diverse or less commonly shared awareness in this field.

*This work was not supported by any organization

[1] A. Wakwabubi is an MS. Data Science Candidate at the Goergen Institute for Data Science, Arts, Sciences & Engineering, University of Rochester, 1209 Wegmans Hall, Rochester, NY - 14627, USA Email: eayonga@ur.rochester.edu

B. Aggregated Mean Awareness per County

The next step was to compute Covid-19 awareness across the 88 counties in Ohio state. The per-county awareness is graphed as shown below:

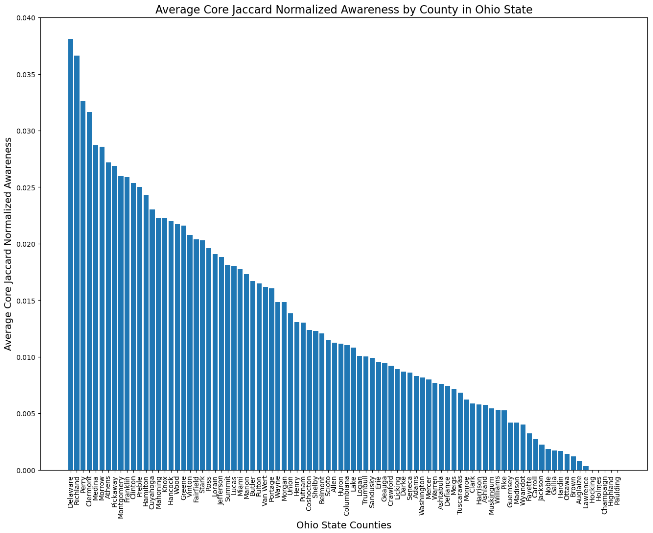

Fig. 2. Average Core Jaccard Normalized Awareness by County in Ohio

The county with the greatest average normalized awareness is **Delaware**. As the chart progresses to the right, each successive county has a lower average normalized awareness, indicating less shared awareness or engagement. The counties **Hocking, Holmes, Champaign, Highland and Paulding** have the lowest values, suggesting the least commonality in awareness among the data points. The chart depicts a steep decline initially, which gradually levels off as it extends to the counties with lower awareness levels.

C. Average Covid-19 Cases and Deaths Per Capita

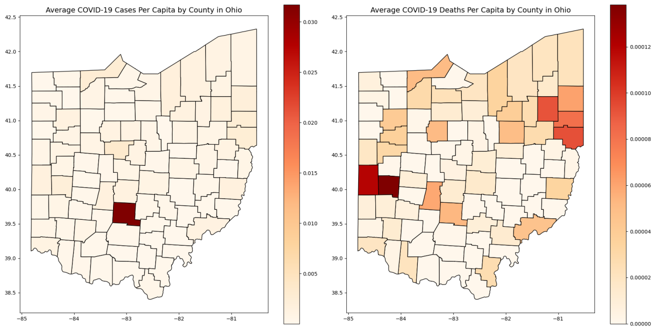

Fig. 3. Average COVID-19 Cases & Deaths Per Capita by County in Ohio

The twin maps display the average COVID-19 cases and deaths per capita by county in Ohio State. The color gradation from light to dark indicates the intensity of cases or deaths per capita, with darker shades representing higher averages. The map on the left highlights one county with a particularly high average of COVID-19 cases per capita, shown in a striking dark red color **(Pickaway county)**. In contrast, the map on the right, which shows deaths per capita, depicts a different pattern, with a few counties in darker shades indicating higher death rates per capita and one county specifically standing out with the darkest shade **(Miami county)**. The maps together suggest geographic disparities in the impact of COVID-19 across the state, with certain areas being more affected than others either in terms of reported cases or deaths.

D. Time Series Evolution of Awareness Levels

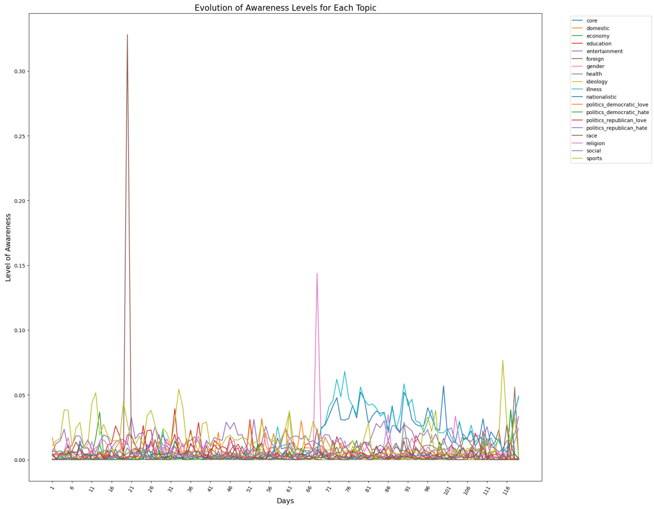

Fig. 4. Evolution of Awareness Levels for Each Topic

The graph represents a time series analysis of discussion intensity on various COVID-19-related topics, as measured by social media engagement in Ohio. Each line represents a topic, with noticeable fluctuations over time. Two spikes dominate the visualization: one in the "core" topic early in the timeline and a more pronounced one in "health" later on. The remaining topics exhibit lower and more stable levels of discussion. This suggests that certain events or developments might have triggered these peaks in conversation, possibly correlating with public health announcements or significant pandemic-related incidents. The broad range of topics indicates a diverse public discourse surrounding the pandemic.

III. METHODS

Data Pre-processing

The initial phase of our methodology entailed a meticulous data pre processing procedure. The first step was to import all necessary Python packages that would be required throughout the modeling process. Features columns were then separated from the main dataframe. This bifurcation was essential to independently handle the features and the target variable, ’cases’.

Dimensionality reduction was then applied to the dataframe, excluding the target variable, using Spectral Embedding, which is particularly adept at handling non linear data structures, which was presumed to be the case given the complexity of social media data. Four n components were set, aiming to capture the most significant structures in the dataset while simplifying the feature space.

After reducing the dimensionality, features columns were reintegrated with the transformed dataframe using concatenation, ensuring the dataset was cohesive for modeling. Next, columns containing string values were eliminated as they are not suitable for regression analysis.

Normalization of features and the target variable followed. This step is crucial as it scales the data to a common range and is particularly beneficial for gradient descent-based algorithms by potentially improving convergence speed.

Model Setup and Training

Post pre-processing, normalized data was partitioned into training and testing sets, adhering to a 80/20 split to ensure a robust training process. The random state was fixed at 256 to maintain reproducibility of results.

An XGBoost regression model was chosen for its powerful ensemble learning capabilities and its proficiency in handling various types of data. The hyper parameters were thoughtfully selected; setting a learning rate of 0.19 to balance the speed and accuracy of convergence, and the number of epochs was fixed at 38000 to give the model a substantial iteration base to learn from.

Model Evaluation

The strength of the model was assessed using standard metrics such as Root Mean Square Error (RMSE) and R2 score. RMSE offers a clear indication of the average distance between the predicted and actual values, while the R2 score reflects the proportion of variance in the dependent variable that is predictable from the independent variables.

The same pre processing steps were applied to the unseen test dataset to ensure consistency. The trained model was then applied to this data to generate predictions. To interpret these predictions accurately in their original context, an inverse transformation was applied to revert the scale of the predictions to their original magnitude.

A new DataFrame was constructed consisting of the index and the predicted number of cases, formatted into a CSV file, and then submitted to the Kaggle competition. The model's performance was validated externally with an R2 score of 0.87488, illustrating a high level of prediction accuracy and model's predictive strength.

IV. Results

The results obtained from the predictive modeling of COVID-19 case numbers in Ohio counties using an XGBoost regression model yielded an R2 score of 0.87488. This value indicates a high level of accuracy when compared to the adjusted R2 score, denoting that our model could explain approximately 87.49% of the variance in the number of COVID-19 cases based on the given predictors, which is a strong outcome in the context of predictive modeling.

Accuracy Evaluation

While the R2 score is a robust indicator of model fit, it is not a direct measure of "accuracy" in the traditional sense, as it does not account for how close the predicted values are to the actual values on an individual level. However, the R2 score achieved here suggests that the predictions are generally in line with the actual data. Additionally, other computed metrics like RMSE (Root Mean Square Error) etc. provide further insight into the average error magnitude in the predictions.

TABLE I

Model Result Metrics

Mean Absolute Error (MAE)	41.1487
Mean Squared Error (MSE)	24281.5114
Root Mean Squared Error (RMSE)	155.8253
Median Absolute Error	9.2066
Explained Variance Score	0.9616
Mean Squared Log Error (MSLE)	3.965
Adjusted R2-Value)	0.9604

Self-Evaluation and Critique

In retrospect, the methodology applied in the model creation phase was methodical and grounded in solid data science practices. However, the decision to set the number of components to 4 in the Spectral Embedding process may have oversimplified the complexity of the data, potentially omitting valuable variance that could have been captured with more components.

The choice of XGBoost as the modeling algorithm was appropriate given its performance and flexibility. The hyper parameters were reasonably chosen, with a moderate learning rate and a sufficient number of epochs to ensure thorough learning without excessive computational time. The model's robust performance is a testament to the effectiveness of these decisions.

Despite the satisfactory R2 score, a critical evaluation would point out that the test set was substantially larger than the training set, with a 70/30 split. This unconventional partitioning could have introduced a bias or variance that is not accounted for, potentially impacting the model's generalizability. Moreover, the lack of detail regarding the handling of outliers, potential multicollinearity, and the distribution of residuals are indeed areas for potential improvement.

In addition, the normalization of the target variable could be reconsidered, since normalizing the target variable is less common and may not always be beneficial, as it transforms the scale of the predictions. This could have affected the interpretation and subsequent inverse transformation of the predicted values.

Finally, the submission to a Kaggle competition provided external validation of the model's performance. However, relying on a single metric for evaluation does not encompass a comprehensive assessment of the model's predictive power.

References

[1] DSCC465 lecture notes

[2] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.