Reyta Akdeniz

790824002

Advanced Statistics

Midterm Report

**THE RESEARCH QUESTION**

What factors predict the level of ethical consideration, regarding the recent boycotting movement, in individuals' consumption behaviour, as measured by the number of boycotting-related tweets they post?

**THE DATA**

This project intends to use the twitter data on boycotting-related tweets, by Politus Analytics.

**PROJECT DESCRIPTION**

The ultimate purpose of the project is to apply a machine learning model to predict the number of boycotting-related tweets an individual posts as an indicator of their level of ethical consideration in their consumption behaviour, based on the factors like demographic variables (age, gender, education), and their primary sentimental approach(emotion) to boycotting movement. The general level of activity needs also to be taken into consideration at this point.

As a preliminary task of this project, one of the ideas is to try and create an education variable based on people’s tweets. Another model, therefore, to infer the education of people needs to be applied and utilized, based on variables like language usage, certain grammar rules whether followed or not followed in tweets and the count of the unique words used. Although, the appropriate model to generate the education year value is not decided yet, it is planned to be a unsupervised model since there is not any existing data of education in the data frame to train the model with.

These variables that are proposed to be used to predict one’s education level are all planned to be aggregated to user(individual) level as like the total number of unique words, for instance, in all tweets of a user or the language usage index score average for all tweets of the user. For language usage, indices like Coleman-Liau index can be utilized or a similar idea can be developed for better applicability to Turkish language, or it can be replaced by a simpler measure like average sentence length, according to functionality of it. Thus, a collection of functions intended to be worked out for creating those variables to be indexed into one’s education level, which needs to be tested for reliability.

Once these variables created and education variable is predicted and inserted to the data frame, it is proposed to use regression based –since the dependent variable is a continuous variable—machine learning model. It is going to be a supervised model since the plan is to train the model based on the existing data, where I already have the total number of tweets by individuals regarding the boycotting. The python libraries like Scikit-learn are going to be utilized for ML models.