Capstone 2: Project Proposal

Topic Modeling of Text-Based Ad Feedback

Problem Statement:
How can we identify 5-10 topics in text-based ad feedback from surveys among a campaign's target audience?

Context:
Annual evaluation surveys were distributed among an advertising brand's target audience. As part of the surveys, participants were shown three video ads that were flighted prior to the campaign evaluation. In response to each video ad shown in the surveys, participants were asked to explain what they think the main message of the video ad is. This data is used to inform campaign messaging comprehension and to inform creative updates for future message package flighting.

Qualitative feedback/text-based responses from an advertising campaign's target audience can be invaluable in determining ad message comprehension, and effectiveness as well as in making decisions for future rounds of campaign flighting. However, human analysis of this data requires reading through thousands of responses and manually coding them into themes and topics which is time-consuming and subjective to the reviewer. Utilizing a machine learning algorithm to identify topics based on the textual responses can cut the amount of manual labor hours spent in analysis and minimize some bias due to subjectivity.

Criteria for Success:
Develop a machine learning model to identify the top 5-10 topics in text-based ad feedback among a campaign's target audience.

Scope of Solution Space:
Conduct natural language processing (NLP) analysis of about 1,448 text-based survey responses to generate a topic model identifying key topics in participant feedback to viewing an advertising campaign's video ads. This will require first cleaning text feedback, merging text responses with sample data, processing the text for analysis, developing initial topic models and then conducting hyper-parameter tuning to land on a final topic model.

Constraints:
- Data is based on survey participants who opt-into participating in research, so there is a level of sample selection bias involved in the algorithm results.
- Defining the model's topic clusters still involves human interpretation, so this model requires some domain knowledge by the reviewer and cannot completely eliminate reviewer bias.
- The sample size is relatively small for modeling, with 1,448 text responses (prior to cleaning) among 724 survey participants.

Stakeholders:
Potential clients for this problem include market and advertising researchers, advertising campaign evaluators, and marketing analytics teams.

Data Sources:

- See sample data here
    - This data was collected as part of surveys evaluating a nicotine vape prevention brand across two US states. The same three video ads were tested in each survey.
    - The data includes two data files:
        - Ad_Feedback_Text (response level - stacked):
            - ID: Unique identifier for response
            - Text: Includes all text responses
            - Ad: Indicates the specific video ad the response was for (encoded)
            - Question: Indicates the question asked in the survey
        - Ad_Feedback_Demos (respondent level - flat):
            - ID: Unique identifier for response
            - CalculatedAge: Participant age calculated from birthdate
            - Race: Participant self-reported race
            - Gender: Participant self-reported gender
            - Segment: Audience segment the participant belongs to
            - Region: State where survey was distributed (encoded for client privacy)
            - Urban_Rural: Indicates whether participant indicated they live in an urban or rural area
    - Data sourced from Rescue Agency Public Benefit LLC campaign tracking online surveys.
    - Data has been de-identified to protect participant privacy.
    - Data not publicly available for compliance with IRB protocol requirements, participant and client privacy.