Migrant Crisis Perception Clustering & Sentiment Analysis Report

Cleaning & Preparing the Data

There were two csv files to look at, newspaper headlines and tweets. Originally, clustering was going to happen on the categorical variables regarding things that were mentioned in the headline/tweet (variables like mentioning political affiliation, taxes, etc.). This resulted in a split between the unstructured textual data, otherwise known as the corpus, and the rest of the clustering data. The newspaper headlines originally had N=999 records and the tweets dataset had N=1630. For the newspaper dataset, 864 NA records were dropped in the clustering dataset, resulting in a final N=135. For the newspaper corpus, we have the same number of records. For the tweets, 1256 NA records and 17 duplicate records were dropped for the clustering dataset, resulting in a final N=357. As for the tweet corpus, 39 NA records were dropped, resulting in an N=1591 sample size of raw tweets.

Preparing the data for clustering and sentiment analysis consisted of transforming variables that had big scales and skewed distributions. Later on, clustering on the categorical variables did not yield any information of value, so from here on in, the corpus datasets are the only ones that matter. The corpuses for both the newspaper headlines and the tweets were processed by removing special characters and digits (essentially anything that was not a-z). We also further normalized all the text by making everything lowercase. After this, stop words were removed to improve the quality of the text (Note: stop words are words such as “as”, “like”, “if”, etc.). After developing these processes to process the text, all the textual variables were put together into one and processed all at once. In other words, the variables “Headlines,” “Hit Sentence,” and “Key Phrases” were put together into one variable and processed (“Headlines” does not exist for tweets so just the other two for the tweet dataset).

Clustering

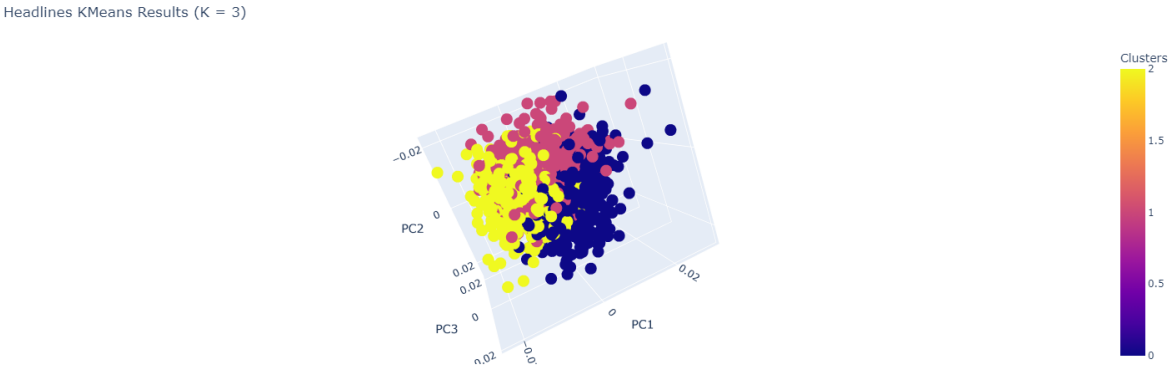
As mentioned previously, clustering was initially done on the categorical variables using KModes clustering. This algorithm uses the mode of a variable as opposed to the mean to calculate the clusters. Using Silhouette Scores, we can evaluate the quality of the clusters. This metric essentially looks at how closely related points within a cluster are while also looking at how unrelated a point in one cluster is to another point in a different cluster. The closer to 1, the better the cluster. The silhouette score for the newspaper headlines using KModes with a cluster size of K=5 (most optimal) yielded a silhouette score of about .20. Similar results were yielded for the tweets dataset. Going with regular KMeans raises the silhouette score a little bit but becomes hard to evaluate due to the amount of categorical variables vs continuous variables in the dataset. This resulted in a shift of direction from clustering the entire dataset with the variables using KModes to get a look at different groups to clustering the actual text using KMeans to achieve the same (and better) results. This was done by vectorizing words using word embeddings with the Word2Vec model. This is necessary since Word2Vec works to essentially quantify the words in a way where context and meaning are preserved for the model to find relationships. Although when clustering it was ambiguous what was the best cluster size, this method yielded great results. For the newspaper headlines, a cluster size of K=3 yielded a silhouette score of about .79. The three groups consisted of:

1. Headlines surrounding the domestic political climate in relation to migrants
2. Domestic treatment/perception of migrants
3. Domestic and geopolitical government approaches towards migrants

(Topics were decided by sampling 10 records from each cluster multiple times to see what interesting words come out of the cluster)

ChatGPT Generated Topic Names (Respectively, using GPT-3.5):

1. US Migrant Crisis with New York City Demands for Action
2. US Migrant Crisis amid Gov Policies and Shelter Shortage
3. US Migrant Crisis and Refugee Assistance Programs



For the tweets, even more in-depth results were yielded. Our elbow plots suggested that an optimal size for clusters was K=7 and this yielded an even better silhouette score of about .83. The clusters revolved around these topics:

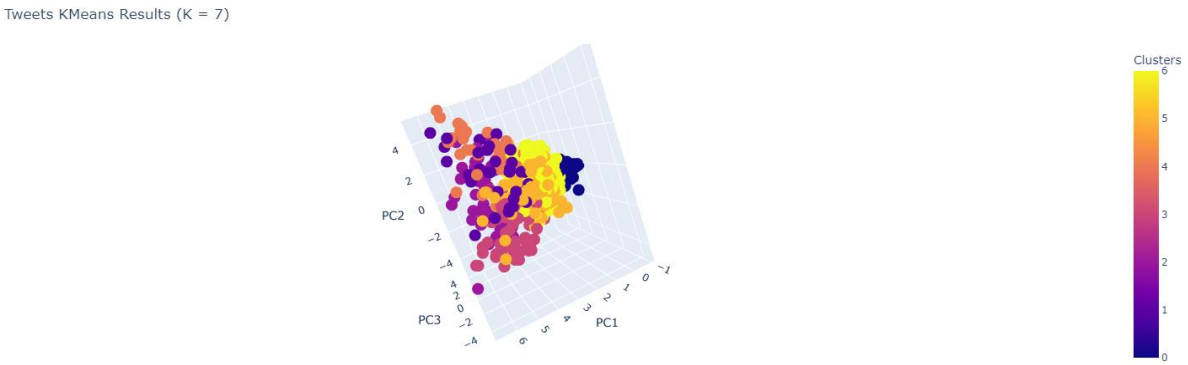
1. Domestic treatment/perception of migrants
2. Migrant perception and criticism of how migrants may be treated, not just in the US (included a lot of words such as “suffering,” “abandonment,”, “nazi,” “holocaust,” etc.)
3. Possible reasons for migrant influxes/discourse in support of offering asylum to migrants
4. Migrant living after reaching the United States
5. Patriotic/nationalist conversations regarding migrants

6. Geopolitics and environmental issues in relation to migrants
7. American treatment towards migrants, most likely coming from both sides (in support vs in opposition)

(Topics were decided by sampling 10 records from each cluster multiple times to see what interesting words come out of the cluster)

ChatGPT Generated Topic Names (Respectively, using GPT-3.5):

1. US Migrant Crisis
2. Migrant Crisis on US-Mexico Border
3. US Migrant Crisis and Urban Support
4. US Migrant Crisis: Challenges and Responses
5. Israeli government's genocide of Palestinian civilians and the US government's support
6. US Migrant Crisis and Global Response
7. Migrant Crisis in the United States

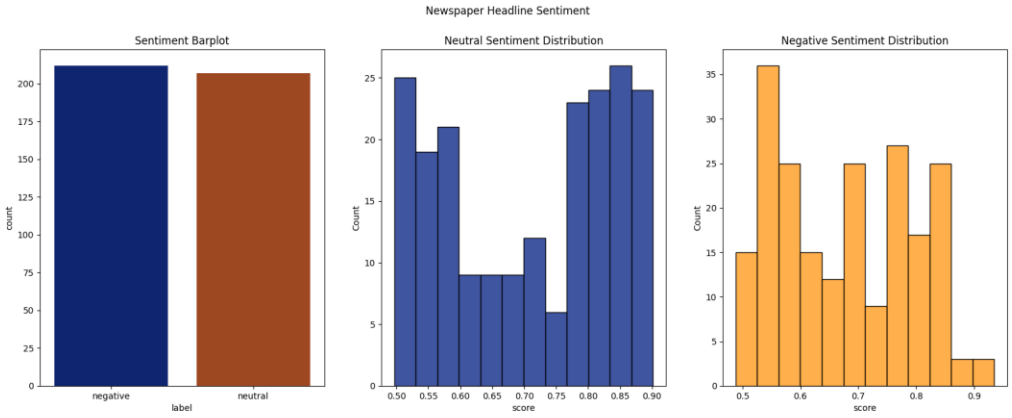


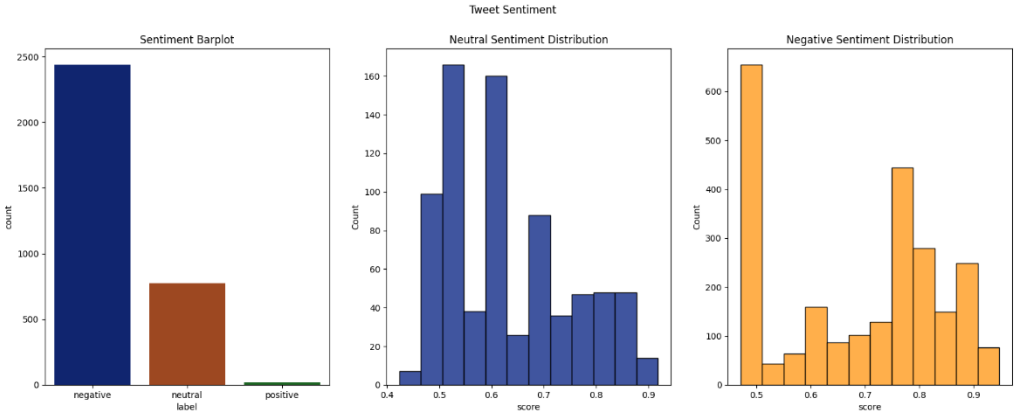
Overall, clustering headlines and tweets directly served as a way to data mine for the different topics that are being talked about in relation to migrants. Our high silhouette scores with the

different clustering results for both the headlines and the tweets give us confidence that these are meaningful groups that make up the conversation. Utilizing ChatGPT to fully analyze all the words within a cluster for a topic name also gives us a bit more insight into what the words in the clusters fully consist of and mean.

Sentiment Analysis

For sentiment analysis, we used the Twitter-RoBERTa-Base sentiment model sourced from Hugging Face. This is an LLM that is finetuned specifically for sentiment analysis and is trained on 124 million tweets from January 2018 to December 2021. Overall, both the headlines and the tweets had mostly negative, if not neutral sentiment regarding migrants. The newspaper headlines were more balanced between negative and neutral, but the tweets were very clearly leaning towards negative sentiment. For the headlines, if a headline was rated neutrally, it tended to be very neutral. The same could not be said for the negative distribution, which shows that a good number of the headlines were not as negative as they could be, with most of the headlines not getting past .8 negative sentiment rating. This trend is also visible with the tweets, with even more entries rating well below .8. The neutral sentiment for the tweets though were ambiguous as a good amount of them rated between .4 and .7 neutral sentiment.




Tweet Sentiment
Sentiment Barplot
count
label
negative
neutral
positive
Neutral Sentiment Distribution
Count
score
Negative Sentiment Distribution
Count
score