ECPAT-USA Social Media Research for Albany, NY

Advanced Analytics and Practicum: Final Report

**Sara Osowski** | sara.osowski@rutgers.edu

**Nivashini Muthuvel** | nm822@scarletmail.rutgers.edu

[**Introduction**](#_heading=h.jac5ugwwo3l2) **2**

[**Background**](#_heading=h.627k30azr69z) **3**

[Company and Industry - Sara](#_heading=h.pjavd35kq3f) 3

[**Literature Review**](#_heading=h.opxu9iji0rr) **3**

[An LSTM Approach to Short Text Sentiment Classification with Word Embeddings - Sara](#_heading=h.f1outjo1bbr0) 4

[Combatting Human Trafficking with Deep Multimodal Models - Nivashini](#_heading=h.hwomkirgnten) 4

[Detection of Human Trafficking Ads in Twitter Using Natural Language Processing and Image Processing - Nivashini](#_heading=h.b6ar6fcljpvi) 6

[Detection of Possible Human Trafficking in Twitter - Sara](#_heading=h.guuhq84fgloz) 7

[Analysis of Twitter Messages for Sentiment - Sara](#_heading=h.ye9dfdu3w98j) 7

[Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation - Sara](#_heading=h.8txqa4qpmqbl) 8

[**Data Overview**](#_heading=h.fh6lj9f2ra6s) **8**

[Twitter API & Cleaning Tweets](#_heading=h.wa8uhvjdrw50) 8

[Sugaring Dataset - Sara](#_heading=h.mvvbb8l62ajn) 9

[Escorting Dataset - Nivashini](#_heading=h.qu38rj5swqls) 11

[**Project Methods**](#_heading=h.9etzr3hv9fw2) **15**

[Determining Keywords and Data Scraping - Sara](#_heading=h.q8m0tnqcnuk2) 15

[Data Cleaning and Processing - Nivashini](#_heading=h.62537t1l2ecc) 18

[Modeling Approach for Escorting-Related Tweets - Nivashini](#_heading=h.955x626ynel7) 19

[**Results: Escorting Modeling - Nivashini**](#_heading=h.cb3lq5xvc3uj) **20**

[**Interpretation of Results**](#_heading=h.cpnfj18ticr3) **22**

[Model Using Escorting Data - Nivashini](#_heading=h.55an53lfhybe) 22

[Applying Model to Sugaring - Sara](#_heading=h.hjjo6d8b4m58) 23

[**Recommendations for Future Work**](#_heading=h.mdy8qczh6c35) **23**

[**Conclusion**](#_heading=h.owvdl5et4lxv) **24**

[**Acknowledgements**](#_heading=h.6nug0ot4sbog) **24**

# 

# Introduction

*Entire team*

Human Trafficking is a massive global problem, driven by the unfair demand for cheap and exploitable labor. Human trafficking can be described as the modern form of slavery as women, men and children are recruited and obtained from extreme labor exploitation. These people who have experienced trafficking can be forced or totured through force, corercion, and threats for labor and/or sex. On an international level, ECPAT-USA states that “40.3 million people were victims of modern slavery.”[[1]](#footnote-0) It also mentions that over one million children are forced into sexual exploitation. In New York City, 2480 children and youth self reported or were determined to have been trafficked and sexually exploited in 2016.

Trafficking victims can experience severe trauma and go through “physical and psychological abuse, including beatings, sexual abuse, food and sleep deprivation, threats to themselves and their family members, and isolation from the outside world.”[[2]](#footnote-1) It is estimated by the International Labor Organization that at least 2.4 million individuals are being utilized for labor due to human trafficking at any given moment. 80% of the human trafficking victims are women and children, and the average age of the victims is 20.

ACLU estimates that human trafficking makes the second largest criminal industry, making about $44.3 billion worldwide annually. Social media platforms aid in human trafficking. Information from social media, such as Twitter, must be collected and analyzed to gain more data-driven insights to predicting human trafficking and identifying methods to prevent human trafficking.

ECPAT-USA (End Child Prostitution and Trafficking - USA) was created “to end the commercial sexual exploitation of children” around the world.[[3]](#footnote-2) Since 2020, ECPAT-USA has collaborated with Rutgers University’s Professional Science Master’s Externship Exchange Program and help teams find data-driven insights to identify and prevent human trafficking. This project aims to build upon previous work to collect data from Twitter and identify trafficked youth being sold through social media. This project specifically focused on uncovering potential trafficking-related social media posts in Albany, NY.

# Background

## Company and Industry - Sara

Previous analytics practicum groups have identified a wide range of keywords, emojis, or hashtags used on social media platforms that are commonly used or associated with human trafficking. The purpose of this project is to scrape human trafficking related data from Twitter to use supervised machine learning on it, allowing us to identify potential sexual exploitation advertisement tweets in addition to trends in data scraped from Twitter for potential human trafficking and sexual exploitation activities. Trafficking can be done for sex, labor, and domestic servitude, and people who have experienced trafficking can be abused physically, psychologically, and through legal process. This project aims to gather and analyze substantial information from social media to be used in transit communication in Albany to bring awareness of the issue.

The project’s goals are to create new keywords/hashtags and scrape data specific to sugaring in the Albany area and to find insights related to escorting using previous data and sugaring-related tweets as sex trafficking advertisements. We aimed to scrape tweets and use data that was already gathered to create a model that detects sexually explicit tweets to be flagged for potential human trafficking.

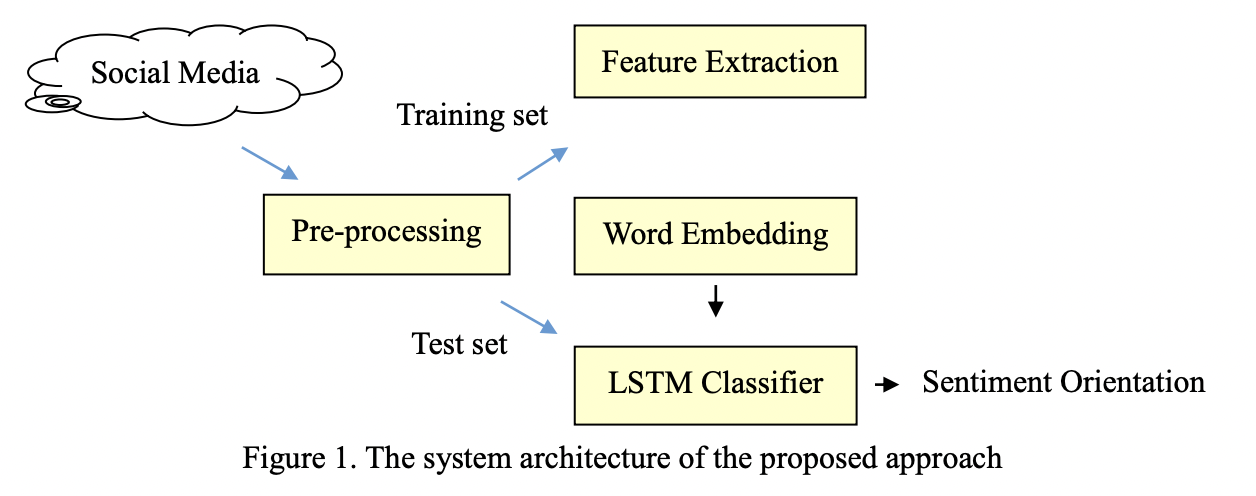
# **Literature Review**

The 6 related publications we analyzed were –

1. An LSTM Approach to Short Text Sentiment Classification with Word Embeddings.[[4]](#footnote-3)
2. Combating Human Trafficking with Deep Multimodal Models.[[5]](#footnote-4)
3. Detection of Human Trafficking Ads in Twitter Using Natural Language Processing and Image Processing.[[6]](#footnote-5)
4. Detection of Possible Human Trafficking in Twitter.[[7]](#footnote-6)
5. Analysis of Twitter Messages for Sentiment.[[8]](#footnote-7)
6. Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation.[[9]](#footnote-8)

## An LSTM Approach to Short Text Sentiment Classification with Word Embeddings - Sara

This paper discusses an approach to sentiment classification by using word-embedding techniques. Since social media posts are usually very short, there’s a lack of features for effective classification. Word embedding models can be used to learn different word usages in various contexts. This paper discusses word embedding and long short-term memory (LSTM) for sentiment classification in social media data. Text words are converted into vectors using word embedding models and the word sequence in sentences are input to LSTM to learn the long distance contextual dependency among words. The following image shows the system of this approach.

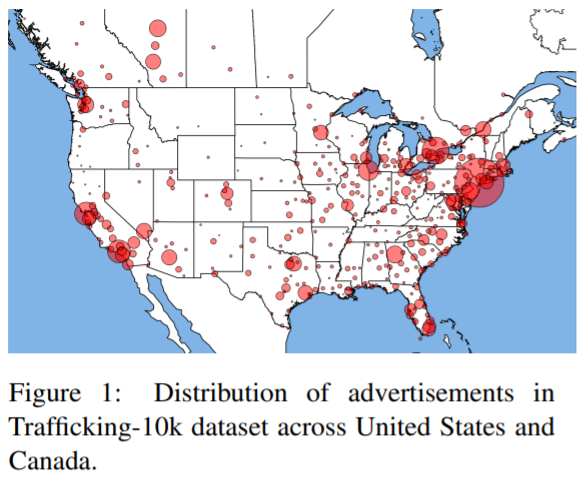


As you can see above, this approach involved pre-processing social media data. Then Word2Vec word embedding model is used to learn word representations as vectors, and finally LSTM is adopted for sequence prediction among words in a sentence. The project compared the performance of LSTM with Naïve Bayes and Extreme Learning Machine and showed that this LTSM approach can achieve better performance than conventional probabilistic models and neural networks. However, it is key to note that it requires good quality and a large quantity of data to do so.

## Combatting Human Trafficking with Deep Multimodal Models - Nivashini

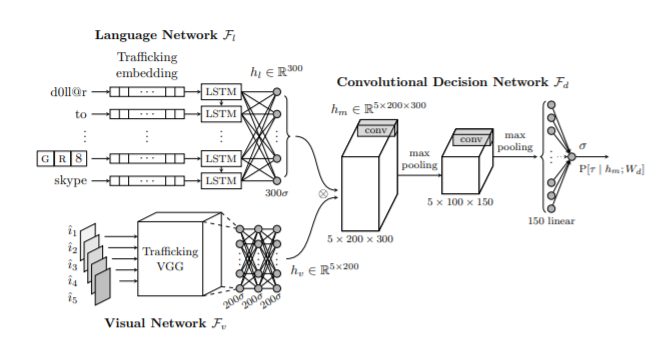
In this paper, researchers from Carnegie Mellon University present the Trafficking-10k dataset, which is a novel dataset that has more than 10,000 human trafficking advertisements annotated for this task. The dataset had two information sources for each advertisement - texts and images. The researchers also designed and developed a deep multimodal model called the Human Trafficking Deep Network (HTDN) to accurately detect human trafficking adverstiments.

The researchers collected 10,000 ads randomly from various escort ads to create the Trafficking-10k dataset, and the dataset was distributed throughout the United States and Canada. The dataset consisted of plain text format and no/some images. The advertisements were annotated as suspicious or non-suspicious by three expert annotators, which had years of experience in detecting human trafficking. Two law enforcement officers also annotated subsets of the dataset to ensure that the annotations are generalizable across all parties involved. The advertisements were quantized into seven levels of various degrees of suspiciousness.



Followed by analysis of the dataset, the researchers developed the Human Trafficking Deep Network (HTDN). It is a multimodal network that consisted of language and vision components, which was perfect for the dataset created. The input data contained the ad, text and images.

The first step was to train the HTDN pipeline on “word vectors based on the skip-gram model.” The Skip-gram model is chosen as it is able to capture the context of the words without relying on the characters or word order. The word embedding was trained using 1 million unlabeled ads from a different dataset. The final trained word vectors were able to cover 94.9% unigrams in the Trafficking-10k dataset. Second, the researchers created a language network to address the irrelevant information in trafficking ads and constituency violations. In addition to the language network, the researchers created a vision network to work in parallel and the image data was inputted. “A deep convolutional neural network called Trafficking-VGG (T-VGG), a fine tuned instance of the well-known VGG network” was used to learn more about the contextual and abstract information from the images inputted. Then a multimodal fusion was created to analyze the interactions between the text and images, by utilizing a joint multimodal tensor.



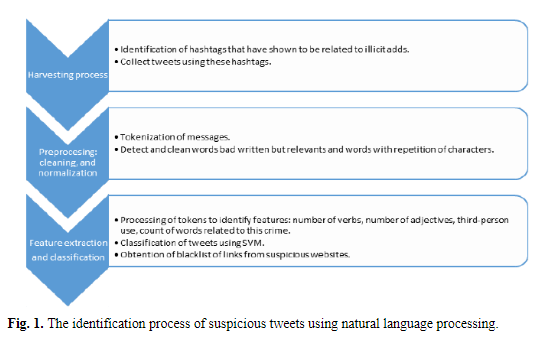
*Figure 3: Overview of the HTDN model*

Lastly, the model was compared with other previously used algorithms for detecting trafficking ads by comparing the accuracy, weight accuracy and f1- scores. HTDN was shown to outperform all the other models, including Random Forest, Logistic Regression and Linear SVM.

## Detection of Human Trafficking Ads in Twitter Using Natural Language Processing and Image Processing - Nivashini

In this paper, the researchers utlize natural language processing and image processing to detect and identity human trafficking advertisements in Twitter. First, the data and messages from Twitter that are considered to be suspicious of human trafficking are collected. This is done by using hashtags that are specific and normalized. Followed by this, Haar filters and Support Vector Machine (SVM) algorithm are used to classify the collected information from the tweets, especially the age and gender groups.

The paper’s approach for detecting human trafficking ads in Twitter is divided into 2 phases. Phase 1 of the approach is using natural language processing for the treatment, analysis, and classification of the tweets using specific hashtags. The researchers conducted preliminary analysis in tweets and Facebook posts and identified common words and phrases used for human trafficking. After collecting the tweets, the data was normalized with NLP techniques such as lexical normalization algorithms and the messages were cleaned following a specific criteria.



*Figure 4: Image from the paper showing the workflow of identifying suspicious tweets using NLP*

Following this process, the researchers extracted features for classification by using a Support Vector Algorithm with a “semi-supervised approach for classifying 55123 recent tweets, all with the chosen target hashtags.” The researchers used 10% of annotated data to evaluate the accuracy and performance of the algorithms. The SVM classification model’s precision was 90.7%, recall was 87.3% and F-measurement was 89.9%.

## Detection of Possible Human Trafficking in Twitter - Sara

This paper discusses the relevance of human trafficking in Twitter. Twitter has become, it says, a “criminal-friendly tool[s] used to contact and deceive their preys and also, for making covert advertising of their illicit activities.” This project does what we did - tweets are collected and processed to detect tweets that may be relevant to trafficking. This project uses a semi-supervised learning method with Naîve Bayes and SVM algorithms to classify the tweets as "suspicious” or “not - suspicious" and tests the validity of predefined features used for their classification against the classification made by experts and was able to effectively detect tweets that may be related to trafficking of underage girls specifically.

This project was very relevant to our work and their methods of classification were able to be applied to our project’s work. We evaluated the techniques used here to be considered in our work due to the effective performance of this model.

## Analysis of Twitter Messages for Sentiment - Sara

In this dissertation, Eric Brown has the end goal of extracting sentiment from Twitter and tying this to stock prices as a tool for market decision making. While the Twitter portion is new, he notes that techniques and prediction methods for stock prices are not new, and are the basis for many types of analysis within the Finance sector. In the late twentieth century, the Random Walk Theory was popular, and effectively stated that future stock prices cannot be predicted based on previous prices. As the age of tech and data emerged, data soon became available to have this theory lose favorability. Now, coming up with such a method is possible, while not easy, and it is important in uncovering the underpinnings of the complex stock market. Though our project is not discussing the stock market, his concepts and methods can be applied to sentiment scraping for Twitter data related to trafficking. Rather than analyzing positive or negative language from the Twitter data, sentiment analysis is applied to our project for a scoring of sexually explicit language.

He takes a Bayesian classification approach where 3 nominal sentiment classes are assigned to tweets for the training data, also using some user reputation score as an input. He uses regression to model sentiment to stock prices and goes on to provide actual strategy recommendations as a next step, making this a prescriptive approach. While the initial results he found were not promising, after re-working the models (namely the sentiment data), he was able to uncover more promising results.

## Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation - Sara

Zhang conducted a literature review of similar studies to the previous article which will be helpful for our team to find further information on our topic of sentiment analysis and the importance of the topic. Zhang used Twitter API through Python’s Twython. They collected 500,000 tweets using the search feature of the API and saved relevant information including hashtags and keywords into MangoDB. This thesis explained n-grams and stated that tagging parts of speech and including a “neutral” category was not helpful. They used Python’s Natural Language Toolkit and Chi-Squared tests to find the best features. Zhang used Naive Bayes, Maximum Entropy, and Support Vector Machines to classify tweets into “positive” and “negative'' categories.

Some challenges Zhang encountered were negations as well as slang, which was noted when starting our project. We collected keywords that were used by previous groups and studied potential other slang keywords that can be used to supplement our sugaring terms.

# Data Overview

## Twitter API & Cleaning Tweets

The Twitter academic developer account gives API key and API Secret key. After applying, we created an application which creates an Access token and Access token Secret. The four keys are used to extract Twitter data in Python. The tweepy package in Python will be used to access Twitter API. Tweepy streaming functions are used to stream tweets and their corresponding data into a json file. We filter tweets which have our keywords and are in English language into the file.

Our next task was to create data from this raw unstructured data and clean it to include the parts which are useful for our analysis. The json is flattened into a pandas dataframe which has numerous data points.

For confirming locations of tweets, we used text extraction from the tweet as most tweets do not use geolocation and previous research suggests that traffickers are not known to use geolocation. For location extraction from text, we load three packages on the Python notebook: NLTK (natural language toolkit), spacy, and location tagger. These three packages are used to read the user tweet text and detect any place names that are mentioned in the text. New columns are created with cities and states. Data is saved as a .csv file.

## Sugaring Dataset - Sara

The sugaring dataset resulted in 4,406 rows and 3 columns. The process of scraping data will be discussed along with the project methods in a later section. However, the following is a screenshot of the sugaring data that was collected:



*Figure 5: Screenshot of the Sugaring Data, Before Cleaning*

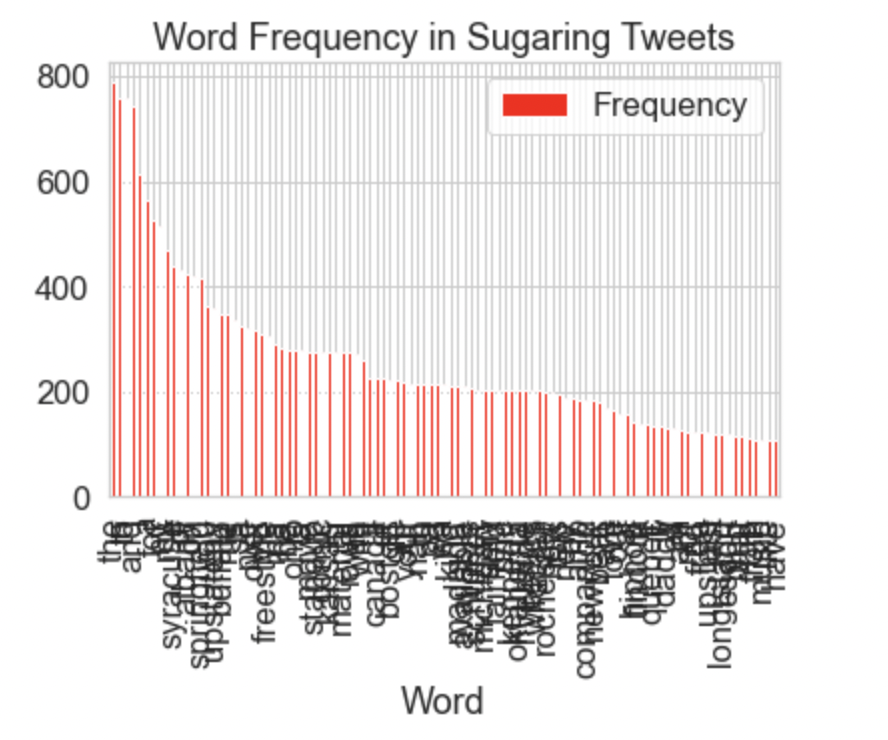
2,000 of the collected tweets were scraped manually for the span of one day (4/5/2022) using our old code (inefficient method) only with keywords correlating “Bing” & “Toga” with “SA”. Therefore, there are more tweets with these keywords than other keywords. The rest of the data was scraped using the partially-automated, hashtag-only scraping method from 1/1/17 to 4/1/22. The data scraped from this method includes information from all 2,800 pairs of hashtags. The following table shows further details about each data type that is collected from every tweet.

| **Column** | **Data Type** | **Range** | **Example** | **Count** |
| --- | --- | --- | --- | --- |
| Index | Numeric (Float 64) | 0 - 4406 | 0 | 4406 |
| Text | String (Object) | N/A | Certified Dabi simp üëπ 18 years of life under my belt üòé She/Her | 4406 |
| Date/Time | Date Time (Object) | 2017-01-01 04:47:01+00:00  - 2022-04-05 18:06:20+00:00 | 2022-04-05 18:06:20+00:00 | 4406 |

*Table 1: Sugaring Data Information*

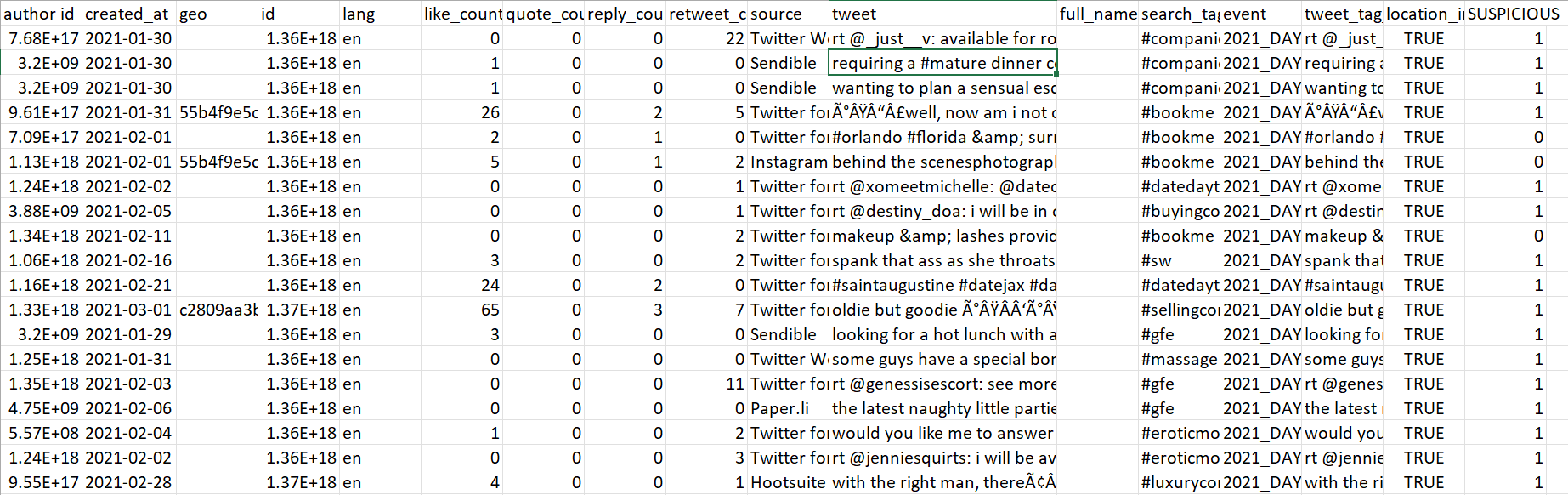
**Exploratory Data Analysis** - Sara

The sugaring dataset was evaluated to find patterns, beginning with the frequency of each word. The following figure shows this word frequency of the 100 most frequently occurring keywords with how many tweets they occur in. Using our code, you can additionally see a list of all of these keywords in order of most to least frequent.



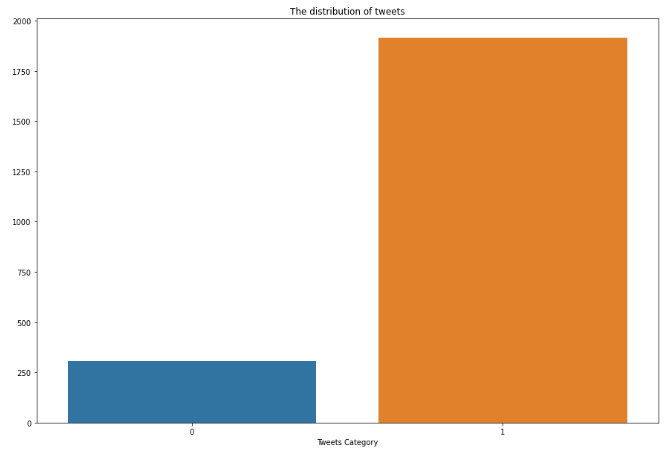
*Figure 5: Sugaring Word Frequency*

## Escorting Dataset - Nivashini



*Figure 6: Screenshot of the labeled escorting related tweets in the dataset*

The escorting data was collected from the previous semester dataset, which was collected from scraping tweets from Twitter with specific hashtags. It has data from Janurary 2019 to November 2021 and is captured using various search tags. In this dataset, the tweets were labeled as sexually explicit and non-sexually explicit manually. The dataset contains 1915 suspicious tweets (labeled as 1) and 306 non-suspicious tweets (labeled as 0).



*Figure 7 showing the distribution of tweets as sexually explicit (1) or non-sexually explicit (0)*

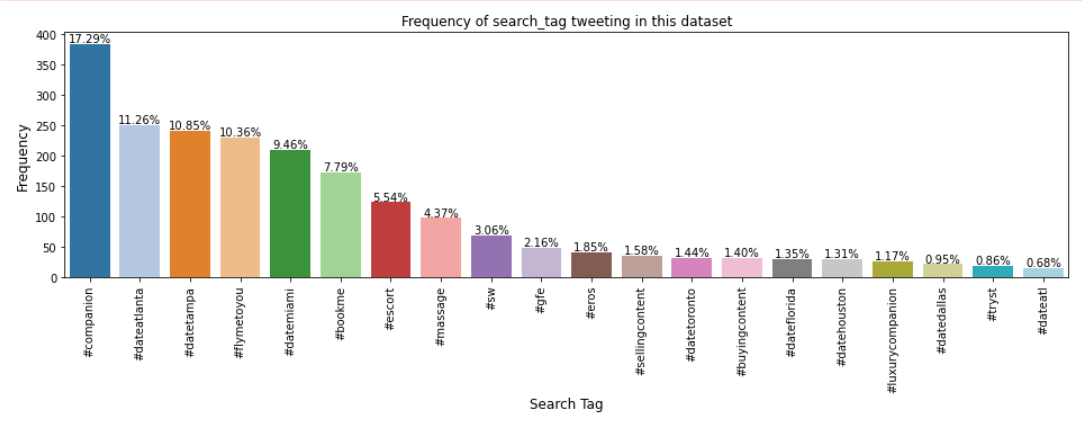
Each datapoint has information on several intrinsic and contextual categories - author ID, tweet ID, date, time, tweet text, search tag, etc. Approximately 2000 scraped tweets were labeled and this is the dataset used to train the machine learning models. The final escorting dataset had 2221 rows and 17 columns.

| **Column** | **Data Type** | **Range** | **Example** | **Count** |
| --- | --- | --- | --- | --- |
| author id | Numeric (Float 64) | N/A | 9.55E+17 | 2221 |
| created\_at | Date Time (Object) | Jan 2019 - Nov 2021 | 2021-01-30 17:50:11+00:00 | 2221 |
| geo | String (Object) | N/A | 55b4f9e5c516e0b6 | 661 |
| id | Numeric (Float 64) | N/A | 1.36E+18 | 2221 |
| lang | String (Object) | en | en | 2221 |
| like\_count | Numeric (Float 64) | 0 - 1319 | 9 | 2221 |
| quote\_count | Numeric (Float 64) | 0 - 8 | 1 | 2221 |
| reply\_count | Numeric (Float 64) | 0 -102 | 2 | 2221 |
| retweet\_count | Numeric (Float 64) | 0 - 491 | 10 | 2221 |
| source | String (Object) | N/A | Twitter Web App | 2221 |
| tweet | String (Object) | N/A | requiring a #mature dinner companion in #orlando? #companion #gfe~ @crystaheart https://t.co/ylgmzcti1z https://t.co/qc4crzp4bi | 2221 |
| search\_tag | String (Object) | N/A | #companion | 2221 |
| event | String (Object) | N/A | 2021\_DAYTONA\_500 | 2221 |
| tweet\_tag\_combined | String (Object) | N/A | rt @\_just\_\_v: available for romantic appointments in  Ã°ÂŸÂ“ÂŒ#orlando #tampa &amp; surrounded till the end on june  Ã°ÂŸÂšÂ—#naples &amp; #miami - by request ¢ÂœÂˆÃ¯Â¸ÂÃ¢Â€Â¦ #companion | 1379 |
| location\_ind | Boolean | True, False | TRUE | 2221 |
| SUSPICIOUS | Numeric (Integer) | 1, 0 | 1 or 0 | 2221 |

*Table 2 with attributes and their characteristics*

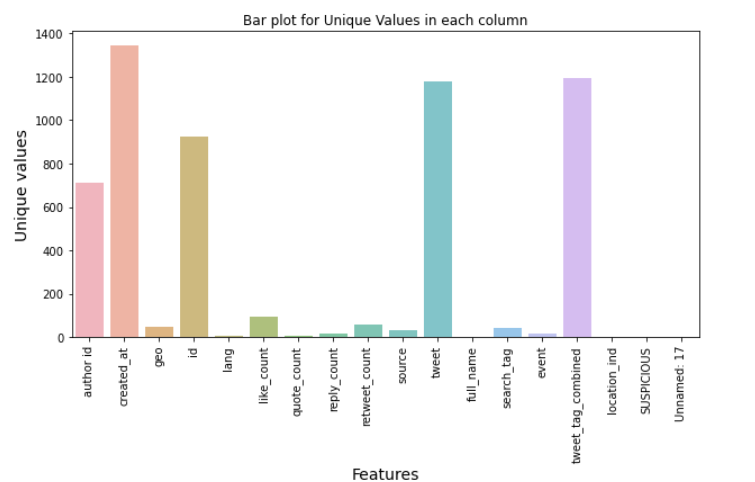
**Exploratory Data Analysis** *- Nivashini*

After analyzing the dataset, exploratory data analysis was performed to check for assumptions, find patterns and conduct graphical visualizations to observe trends. The exploratory data analysis included finding the most frequent search tags in which the tweets were posted. It was founds that search tags such as #companion, #dateto(location), and #flytome were the most popular hashtags used for sexually explicit tweets, making up for more than 60% of the sexually explicity tweets.



*Figure 8 showing the frequency of tweets by each author as percentage of the tweets collected in the escorting dataset*

To further analyze the dataset, a bar plot was created to understand the frequency of unique values in each column and seeing if there are any interesting patterns. It was found that the number of unique tweets ID was much higher than the number of unique twitter user IDs. This meant that multiple tweets related to escorting hashtags were created by the same user profile.



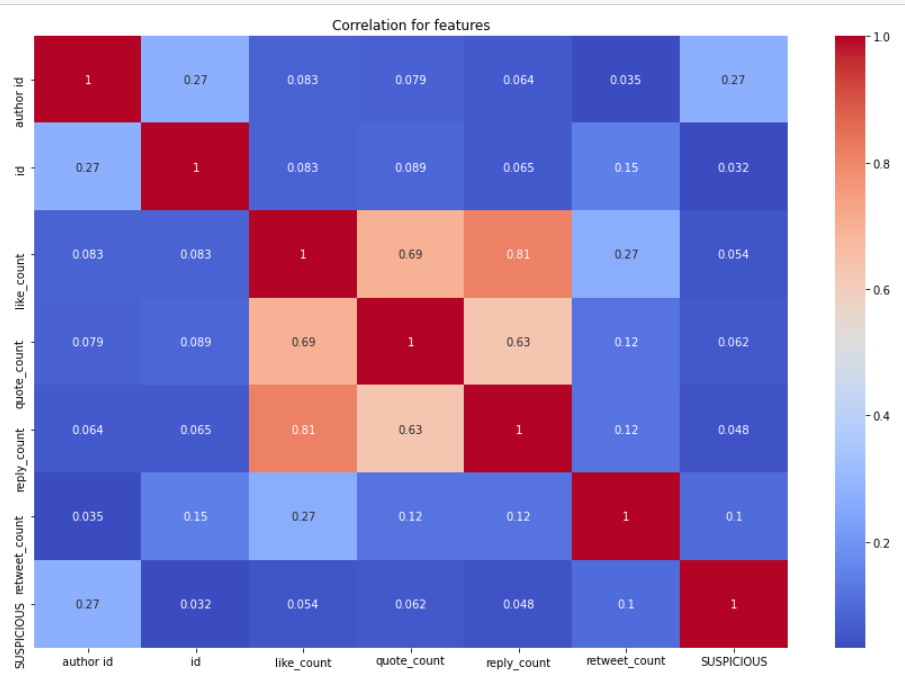
*Figure 9 showing the frequency of unique values in each attribute in the escorting dataset*

To further analyze the author ID and see if there were any major patterns, a bar plot was created showing the frequency of tweets by each unique author ID as percentage of the tweets collected in the escorting dataset. The largest number of tweets by a single author ID was equal to 2.8% of the dataset, which meant that the author had posted roughly 60 tweets relating to escorting.



*Figure 10 showing the frequency of tweets by each author as percentage of the tweets collected in the escorting dataset*

A feature correlation analysis was also performed to see the relationship between the features and further understand the data. From the matrix, it was found that author ID is weakly correlated (0.27) with the sexually explicit tweets. The author ID attribute is found to be the strongest correlation when compared with other attributes in the dataset.



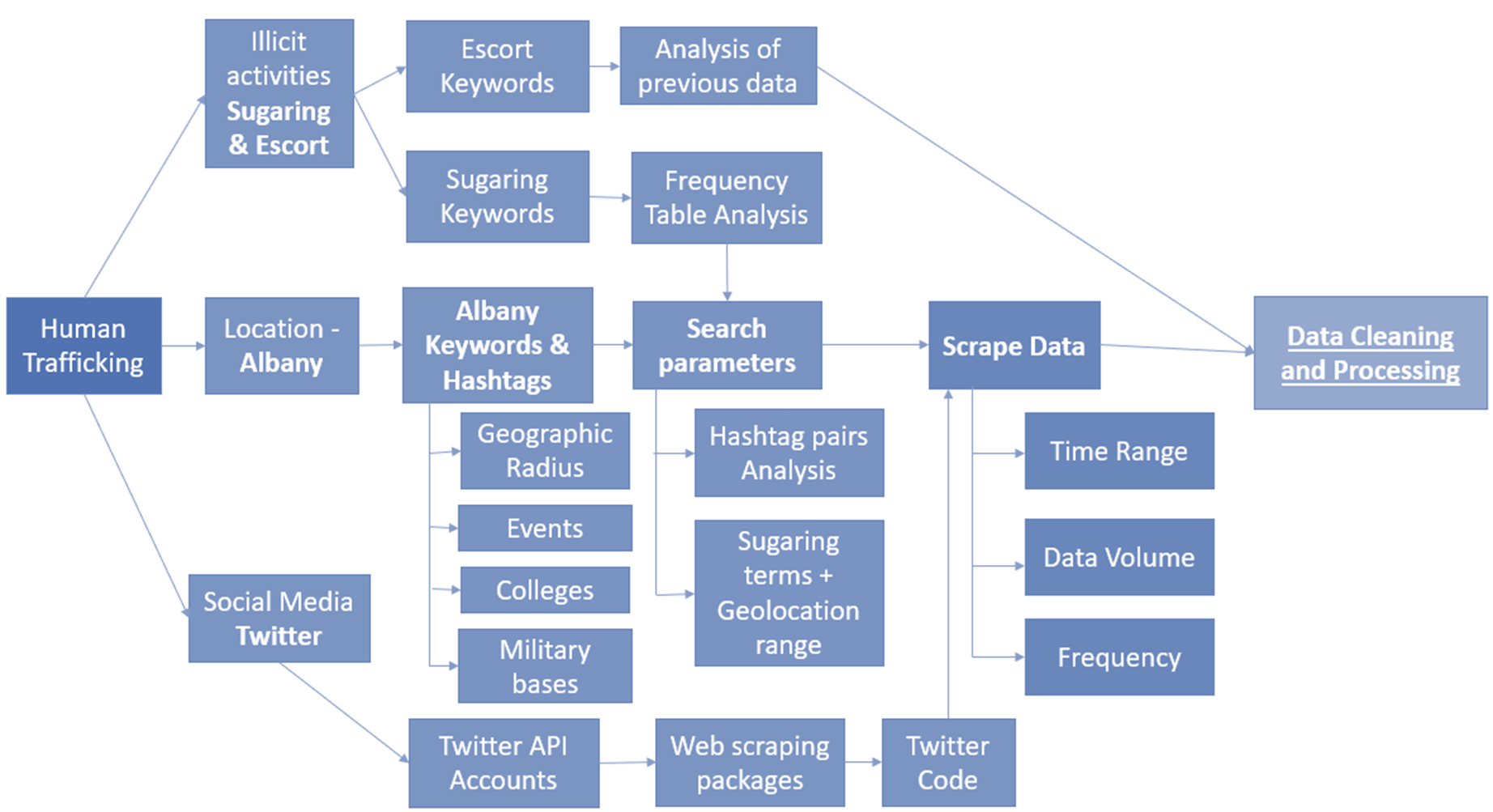
*Figure 11 showing the correlation matrix between the features of the escorting dataset*

Then, the dataset was checked in the sexually suspicious and non-suspicious category to see if the how much of the data is balanced. The dataset was extremely unbalanced as the number of tweets for suspicious were much higher than the number of tweets for non-suspicious categories. This was expected to as the data was scraped using hashtags that were chosen for being most likely to be used to promote sex trafficing content. The dataset was chosen to be inputted into the model as an unbalanced and balanced dataset to ensure the most accurate results.

# Project Methods

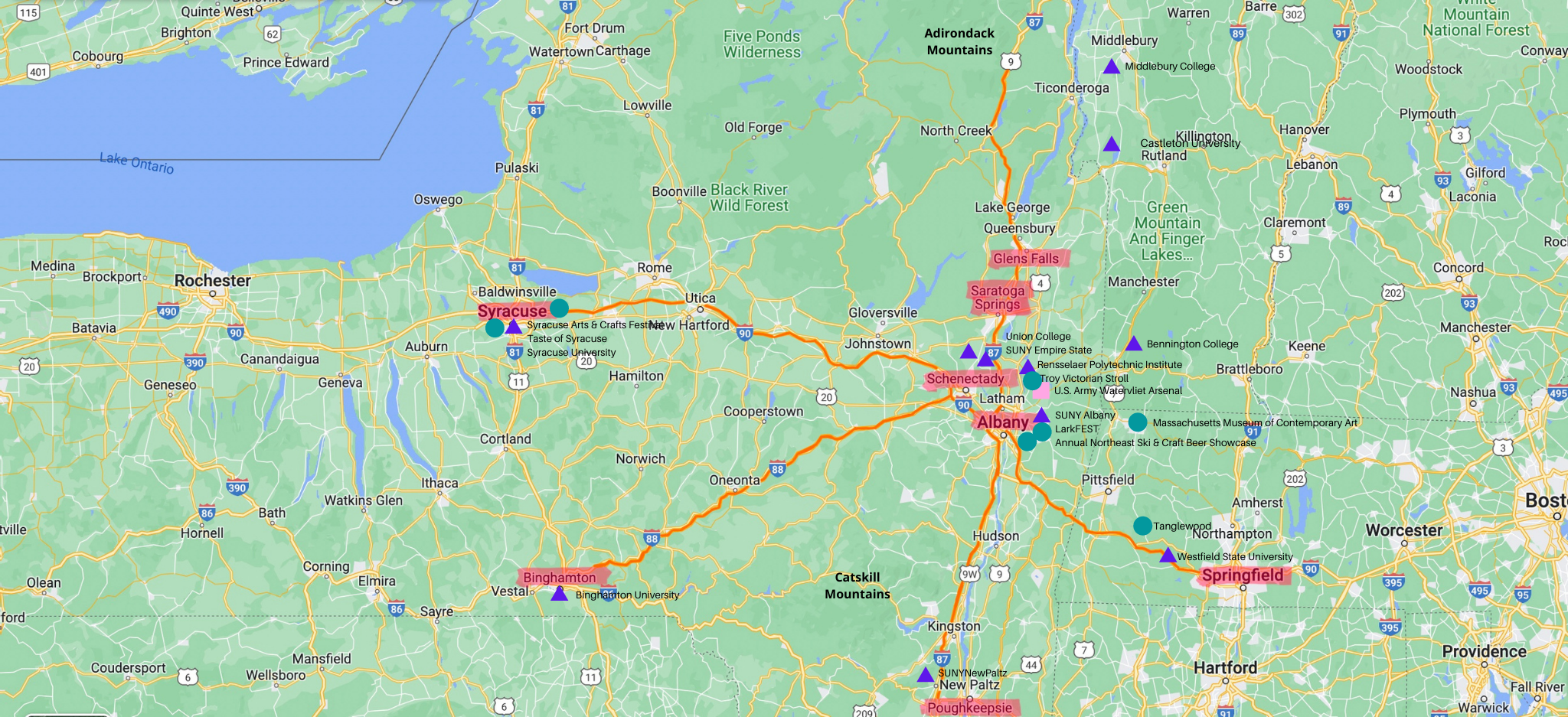
## Determining Keywords and Data Scraping - Sara

The data collection process flow was guided by the following chart:



*Figure 12: Data Collection Flow*

We created a Twitter API academic developer account and observed the intended location for our search to determine what keywords can be created. The following map indicates the geographic area that was decided to be studied. The map indicates interstates and highways outlined in orange, major cities highlighted red, colleges indicated with purple triangles, events with teal circles, and military bases with pink squares. All locations and events that were searched are labeled in the map.

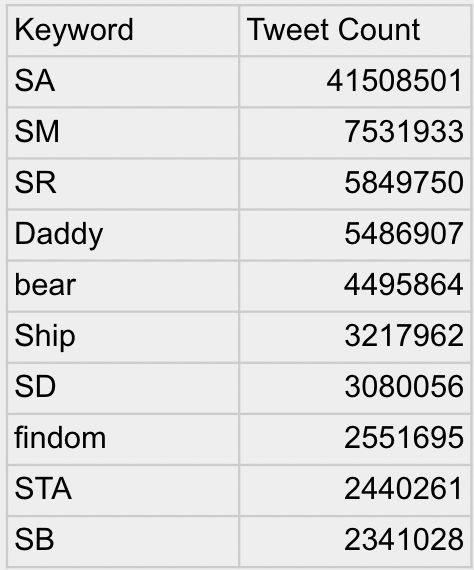


*Figure 13: Map of Area Being Studied*

We created new location/event keywords from research and added sugaring-related keywords to the provided sugaring keyword list, resulting in 57 location/event and 50 sugaring keywords.

To decide the order of keyword scraping, we used Python to extract the counts of each keyword in all Twitter history to understand which keywords would pull the most results. This helped to know which words to start with. As we initially intended to search these keywords as just words rather than hashtags, this method showed a very large amount of results, which is why we aimed to focus on scraping pairs of the top 10 occurring keywords for sugaring and the top 10 for location/events in the Albany area. After the original scraping method was affected by automation issues, we moved onto a partially-automated scraper that only scrapes hashtags, which resulted in a much smaller occurrence of each keyword in hashtag form. This tweet count method was still used and was effective in deciding the ordering of our scraping, but did not actually affect the end result as we scraped every combination of the 57 location/event and 50 sugaring keywords that have been created to get a substantial enough resulting dataset.

The tweet count code generates the number of tweets of a specific hashtag between a mentioned date and time range by accessing the entire archive of public tweets. With it, we searched for each of our keywords from 1/1/22 to 3/22/22. This showed every occurrence of tweets that include the keyword being searched. This code was only able to search for one keyword, rather than a pair of keywords, which is why the tweet count is so high.

**

*Table 3: Sugaring Tweet Count (1/1/22 to 3/22/22)*



*Table 4: Location/Event Tweet Count (1/1/22 to 3/22/22)*

Additionally, the tweet count code is only able to be used with Academic Research Twitter accounts, and it paginates at 31 days.

The initial data scraper uses a manual method to scrape tweets. This method can scrape for a pair of hashtags. It requires an input of a date to start scraping at, and it also requires a number of tweets to scrape for. This scraper was built upon for automation. This scraper was used to gather 2,000 of the tweets that we included in our dataset.

The improved scraper was intended to scrape multiple features of the tweets included like count and username and to scrape larger amounts of data including keywords. We encountered numerous errors while doing this, so this method was fully scrapped.

The final scraper was completed using the methods that the Albuquerque group will discuss, which was run for all 56 location/event and 50 sugaring-related hashtags. This scraper prints errors each run. However, it is able to effectively scrape tweets under a certain unknown amount of tweets, which is why we determined to only scrape hashtags, which are less frequently occurring. This method requires choosing a specific chunk of data to be scraped, scraping it, saving it as a file, and then going on to the next small chunk of data. I recommend scraping in groups of roughly 40 pairs of keywords. As we scraped from most to least frequently occurring keywords, we were able to scrape in larger chunks of keyword combinations towards the end.

## Data Cleaning and Processing - Nivashini

After analyzing the dataset, the next step was data cleaning. The dataset tweets were converted to lowercase. Next step was to generate the words in the tweet into sentences while removing the hashtags and other punctuations. Special characters, emojis and stop words were removed from the tweets using the stopwords package from nltk.corpus, with the language set to English. Stopwords are everyday words that would not be necessary for detecting sexually implict tweets and are removed from the data as they can obstructe the algorithms. Additionally, HTML entities like “&lt” and “&amp” will be removed from data with a HTML parser in cleaning.

The tweet texts were then lemmatized using the WordNetLemmatizer from nltk.stem. The last step was converting the text into vectorized text to input into the machine learning models. This step was done by CountVectorizer from the sklearn feature extraction package. Text vectorization converts the text data into numerical vectors to be inputted into various natural language classification models. Looking at the previous semester work, the punctuation of the text was chosen to be left in the text as previous models seemed to have shown better performance in predicting the sexually implict tweets with punctuation. Only the escorting dataset was then split into 80% train and 20% test for modeling after this.

## Modeling **Approach for Escorting**-**Related Tweets -** *Nivashini*

The labeled dataset was split into two categories for testing purposes - Balanced and Unbalanced dataset. The balanced dataset was created using random undersampling, which is removal of random data points from the majority class. The majority class in the escorting dataset was sexually suspicious tweets. Baseline model and 2 classification models were created to train and test with this dataset in balanced and unbalanced format. The dataset was split into 80% train and 20% test.

| **Model Used** | **Dataset Used** |
| --- | --- |
| Baseline Model - Dummy Classifier | Both - Unbalanced and Balanced |
| Classification Model - OneVsRestClassifier | Balanced |
| Classification Model - MultinomialNB | Unbalanced |

*Table 5*

The baseline models were applied by using the Dummy Classifier library in Python. For a balanced dataset, the baseline model was set to the strategy ‘stratified’ as it generates random predictions according to the training data label distribution. For the unbalanced dataset, the baseline model was set to the strategy ‘most frequency’ as the classifier predicts the most frequency class label in the training dataset. The classification models chosen were OneVsRestClassifier and MultinomialNB.

OneVsRestClassifier is a multi-class classification, and Support Vector Classification model with kernel =’liner’ was used with this classifier. This multiclass support is handled considering the one-us-one scheme within the multiclass labeling[[10]](#footnote-9). Even though this classifier works well with this dataset, it can be an issue for large datasets. For larger datasets, LinearSVC or SGDClassifiers need to be used instead. Therefore, a balanced dataset was used for OneVsRestClassifier.

MultinomialNB “implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification.”[[11]](#footnote-10) This model is extremely efficient when used for much larger datasets and has quicker fit time for the model. Both the models require word vector counts to be inputted. Therefore, an unbalanced dataset was used for the MultinomialNB.

Each model’s parameters were researched and chosen to ensure the best parameters for each model. The performance of each model will be shown with the precision, recall and accuracy score metrics.

*Precision = true positive/(true positive + false positive)*

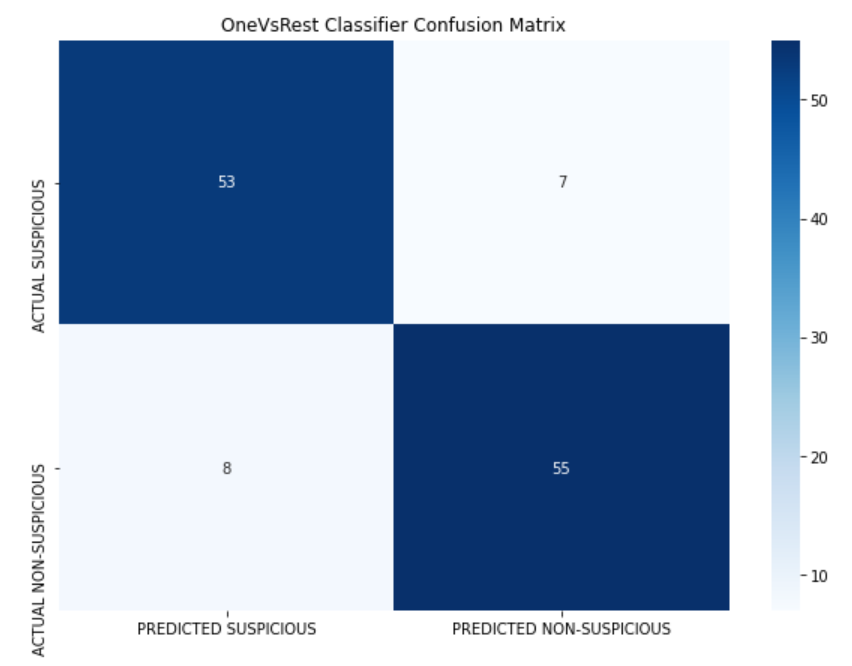
*Recall = true positive / (true positive + false negative)*

*Accuracy = Number of correct prediction/Total number of predictions*

# Results: **Escorting Modeling** *- Nivashini*

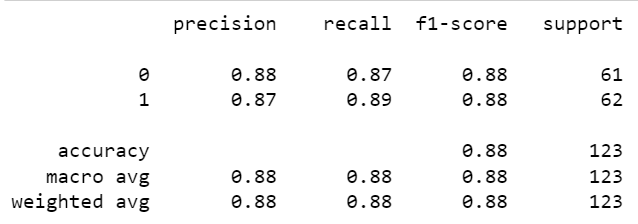
Baseline model was created using a dummy classifier from sklearn.dummy in Python. Strategy 'stratified’ was given and the most was fitted with the escorting dataset. The baseline model using a balanced dataset had an accuracy score of 0.47. The classification model using onevsrest classifier was used with the Support Vector Machine. When the model was fitted to the training dataset and tested, the model had an accuracy score of 0.88. This was significantly better than the baseline model score.

A confusion matrix was plotted for the OneVsRestClassifier. The model accurately predicted the majority of the tweets as sexually explicit or non-sexually explicit and performed the best in predicting non-sexually explicit tweets. The result of the confusion matrix is shown below.



*Figure 14 showing the OneVsRestClassifier Confusion Matrix*

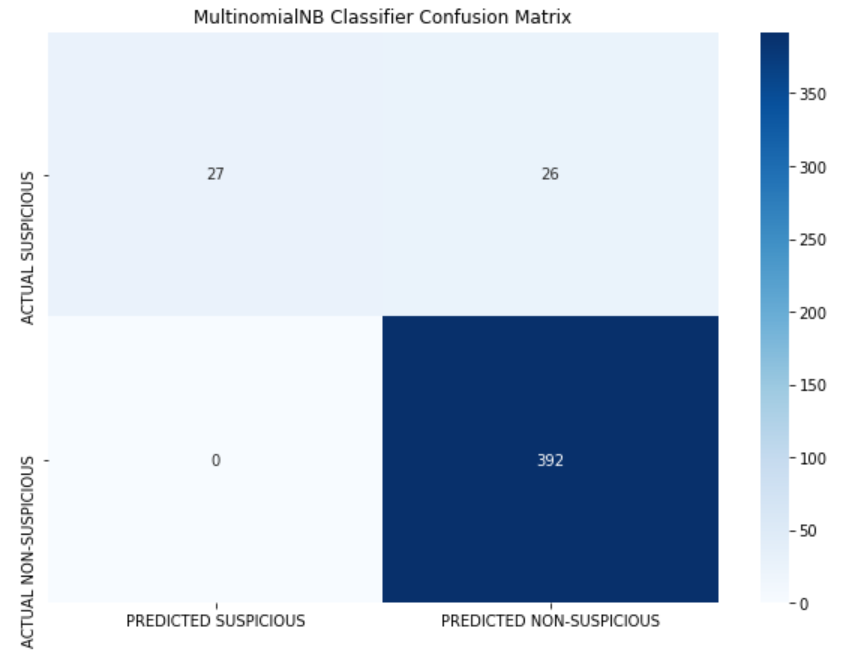
Furthermore, the precision, recall and f1-scores of the OneVsRestClassifier classifier were analyzed using the classification scores. It was found that the precision, recall and f1-scores were all 0.88 in average.



*Figure 15 showing the OneVsRestClassifier classification report results*

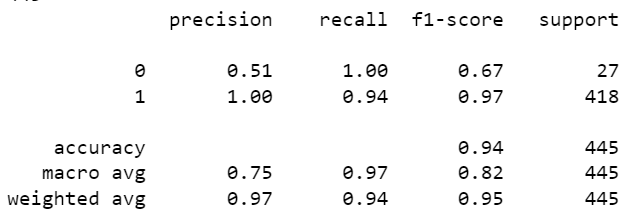
For the unbalanced dataset of escorting data, the baseline model was created using a dummy classifier with the strategy ‘stratified’. The baseline model accuracy score for the unbalanced data was 0.74. The classification model was created using the MultinomialNB classifier after using countvectorizer for text processing. which converts the strings into token integer counts, and TfidfTransformer to convert the integer counts to weighted TF-IDF scores. The model was able to predict the sexually implict tweets in the test dataset with the accuracy score of 0.94.

Just like the previous model, a confusion matrix was plotted for the MultinomialNB classifier. The model accurately predicted the majority of the tweets as non-sexually explicit but performed badly when predicting the sexually suspicious tweets. The result of the confusion matrix is shown below.

**

*Figure 16 showing the MultinomialNB Confusion Matrix*

Furthermore, the precision, recall and f1-scores of the MultinomialNB classifier were analyzed using the classification scores. It was found that the precision, recall and f1-scores were found to be 0.75, 0.97 and 0.82 in average.

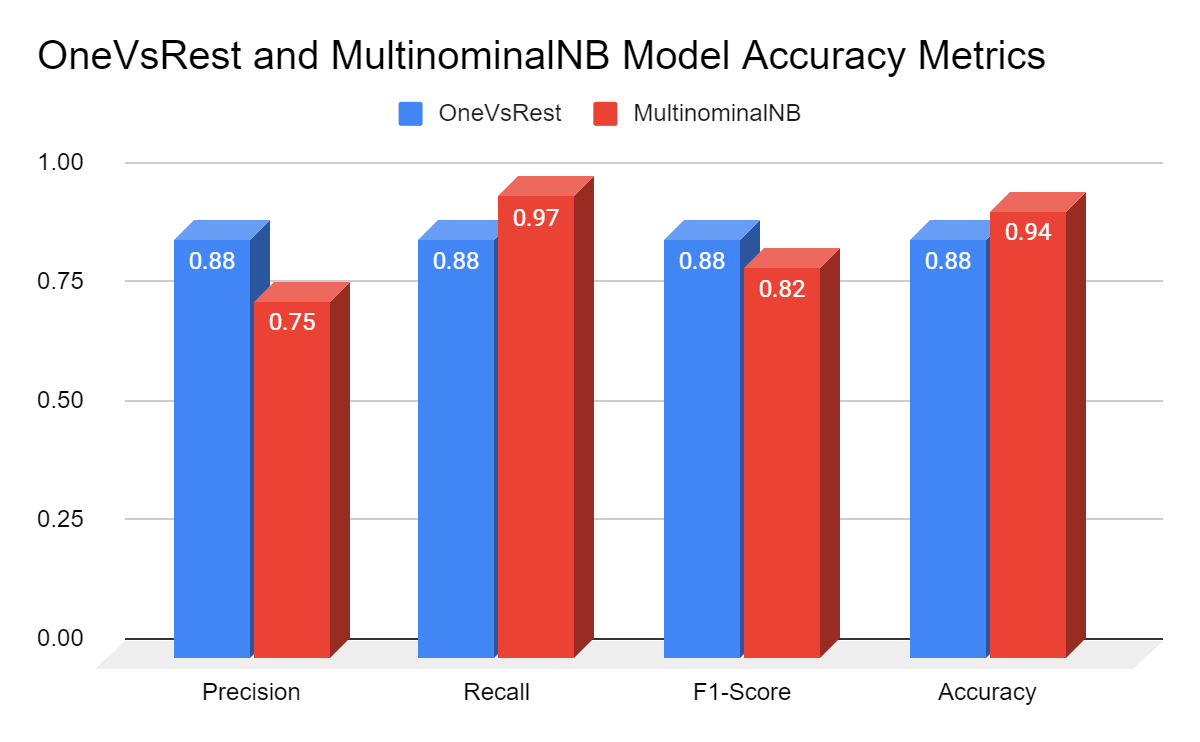


*Figure 17 showing the MultinominalNB classification report results*

# Interpretation of Results

## Model Using Escorting Data - Nivashini

It was found that search tags such as #companion, #dateto(location), and #flytome were the most popular hashtags used for sexually explicit tweets, making up for more than 60% of the sexually explicity tweets. When plotted for the frequency of tweets with each user, it was found that the maximum number of tweets by one user is 60 tweets. This can be useful in flagging certain users on the social media platform or tracking these user activities. The dummy classifier strategies were chosen to ensure that the baseline models were given the appropriate parameters to produce the baseline accuracy score values. When comparing the accuracy scores alone, the models were successfully able to classify the tweets as sexually explicity or not sexually explicit and both the models performed better than the baseline models.



*Figure 18 showing the accuracy metric scores of each model in Precision, Recall, F1-Score and Accuracy*

Even though the MultinominalNB model performed significantly better than the baseline model, the model doesn’t seem to perform better than the OneVsRestClassifier. This can be explained when looking at the confusion matrix and the classification score results. Therefore, the best model for classifying tweets as sexually explicit or non-sexually explicit is the OneVsRest Classifier model.

## Applying Model to Sugaring - Sara

We applied the model that was effective in finding sexually explicit tweets from the escorting data to the sugaring data. This proved ineffective; only four tweets were flagged. This is due to the model only searching for sexually explicit words that are realted to escorting. The model can be further improved if the project were extended and we were able to label data from the sugaring dataset and create another model that is trained to flag more sexually explicit data.

# Recommendations for Future Work

*Sara*

We recommend better documenting code for other ECPAT-USA groups to use in the future. We used escorting data from last semester, but we only had access to the group’s models, meaning that we had to re-do some work that was previously done. To see if the model from escorting data can further improve, a future group can label more data or label with a different method.

Future work can also become more accurate through translating strings to emojis in data cleaning steps in addition to the tokenization and word processing that we used.

Groups can perform a temporal analysis by utilizing location and geographic coordinate data in the model as a feature rather than only in scraping of tweets. Geographic coordinates are not very common in collected data, but it would be helpful to do this to see if any insight is gained.

We lastly recommend that future groups build a fully automated scraper with more features off of the current scraping method as the current scraper only collects the tweet text and the date/time information. Future groups can also build on our work by further exploring the sugaring dataset as more insight can be gained from this data.

# Conclusion

Throughout this project, we learned how severe the issue of commercial sexual exploitation of children is and how ECPAT-USA is working to eliminate the human trafficking. We gained insight on how tweets can be scraped, analyzed, and encoded using Python, and we created models to classify suspicious tweets using the vectorized data and had accuracy scores over 80%.

# Acknowledgements

We would like to express our gratitude to Dr. DiRienzo, Dr. Nelson, and Dr. Misra for helping us throughout this project. We would also like to thank John Betak for being an amazing program mentor with a plentitude of knowledge on trafficking and Ryan Maiorano for the initial tips on scraping tweets. Additionally, we thank Yvonne Chen and Katisha Andrew for teaching us about ECPAT-USA and the language behind trafficking. Lastly, we thank Evan for helping us through our scraping challenges as well as the rest of the Albuquerque practicum group for all of their hard work. Nivashini and I are extremely grateful for such great advisors and such a great team that we were able to learn so much from.

1. “STATISTICS ON TRAFFICKING AND EXPLOITATION.” ECPAT-USA, <https://www.ecpatusa.org/statistics>. [↑](#footnote-ref-0)
2. “Human Trafficking: Modern Enslavement of Immigrant Women in the United States.” American Civil Liberties Union, https://www.aclu.org/other/human-trafficking-modern-enslavement-immigrant-women-united-states. [↑](#footnote-ref-1)
3. “Community Education.” ECPAT-USA, https://www.ecpatusa.org/communityeducation. [↑](#footnote-ref-2)
4. *An LSTM approach to short text sentiment classification ...* (n.d.). Retrieved February 17, 2022, from https://aclanthology.org/O18-1021.pdf [↑](#footnote-ref-3)
5. Tong, E., Zadeh, A., Jones, C., & Morency, L.-P. (2017, May 8). *Combating human trafficking with deep multimodal models*. arXiv.org. Retrieved February 17, 2022, from https://arxiv.org/abs/1705.02735 [↑](#footnote-ref-4)
6. *(PDF) detection of human trafficking ads in Twitter using ...* (n.d.). Retrieved February 18, 2022, from https://www.researchgate.net/publication/342691738\_Detection\_of\_Human\_Trafficking\_Ads\_in\_Twitter\_Using\_Natural\_Language\_Processing\_and\_Image\_Processing [↑](#footnote-ref-5)
7. *Detection of possible human trafficking in Twitter*. (n.d.). Retrieved February 18, 2022, from https://www.researchgate.net/profile/Myriam-Alvarez/publication/338451205\_Detection\_of\_Possible\_Human\_Trafficking\_in\_Twitter/links/5fb2c2e3299bf10c3685e0a0/Detection-of-Possible-Human-Trafficking-in-Twitter.pdf [↑](#footnote-ref-6)
8. *Brown, Eric D., “Analysis of Twitter Messages for Sentiment" (2014). Masters Theses & Doctoral Dissertations. 291. https://scholar.dsu.edu/theses/291* [↑](#footnote-ref-7)
9. Zhang, Linhao. “Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation.” The University of Texas at Austin, 2013, repositories.lib.utexas.edu/handle/2152/20057. [↑](#footnote-ref-8)
10. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC> [↑](#footnote-ref-9)
11. <https://scikit-learn.org/stable/modules/naive_bayes.html> [↑](#footnote-ref-10)