Sex Trafficking Advertisement Classification
With a Support Vector Machine

Rebecca Quay Dragon
Northeastern University

April 26th, 2021

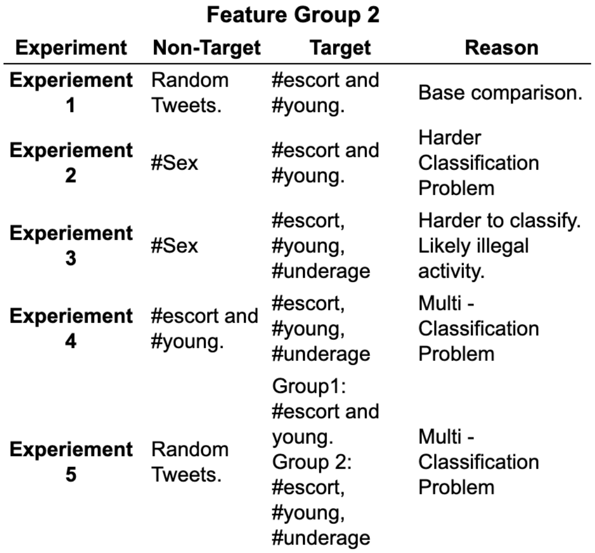
Abstract

This research classifies tweets that are indicative of sex trafficking advertisement language using a support vector machines with linear kernels. The increase of usage of social media by traffickers and victims for advertising illegal sexual transactions is a motivating factor behind pursuing training an algorithm that can correctly classify tweets that are likely advertising sex trafficking. This research capitalizes on an SVM model provided by Sk-learn the machine learning library. Five separate experiments were performed on different tiers of data. The difficulty of the classification problem and importance of the classification increased as the experiments progressed. The results of the models greatly differ in their precision with earlier experiments performing at 99% precision.

Introduction

Every year, 4 million people are sex trafficked globally (USA Today, 2019). The average age of entering “The Life” is approximately 12 to 14 years old (Office of the Assistant Secretary for Planning and Evaluation, 2019). Technology has made it much easier for traffickers and individuals to advertise their illegal services. Social media companies are used by traffickers to groom individuals and then connect them with clients. Thorn, an organization dedicated to fighting human trafficking has states: “One in seven respondents who were advertised on the street reported more than 10 buyers per day. By comparison, one in four respondents who were advertised online reported more than 10 buyers per day” (Thorn, 2019). A first step in this process of stopping online sex trafficking is to classify potential sex trafficking material and then take the material down.

A support vector machine was used to create this model because of its ability to accurately classify multi-classification problems as well as perform a kernel trick which can map a linear to a higher plane. During preliminary research the linear model performed the best out of the linear, polynomial, sigmoid and gaussian kernels on the count vectorized data data. The hinge loss function provided by sklearn is used to compute the loss on the training data. Hinge loss creates the designated boundary points in which data stops contributing to the loss function. As previously stated, 5 separate experiments were run for this research. The experiments became progressively harder for the SVM to classify but are important to understanding why it will be hard for social media companies to differential between sexually explicit tweets, sex trafficking advertising tweets, and the sex trafficking advertising tweets of minors. The latter has the highest priority for accurate classification and performs well compared to both regular tweet data and sexually explicit tweet data. A detailed description of these experiments can be seen below. The non-target and target columns identify the type of tweets that were queried for each of the experiments.



Feature Group 2

Experiment	Non-Target	Target	Reason
Experiment 1	Random Tweets.	#escort and #young.	Base comparison.
Experiment 2	#Sex	#escort and #young.	Harder Classification Problem
Experiment 3	#Sex	#escort, #young, #underage	Harder to classify. Likely illegal activity.
Experiment 4	#escort and #young.	#escort, #young, #underage	Multi - Classification Problem
Experiment 5	Random Tweets.	Group1: #escort and young. Group 2: #escort, #young, #underage	Multi - Classification Problem

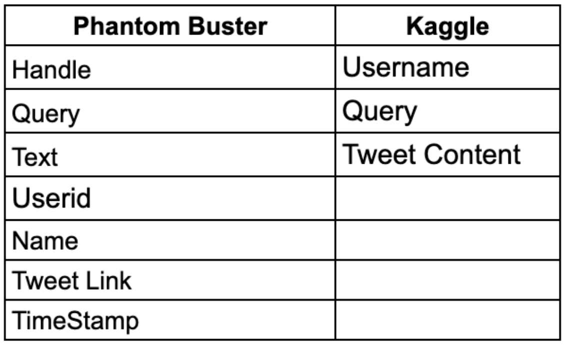
Experiments 1, 2, and 3 yielded an average precision of 90% while using count vectorized features. As expected, the SVM struggled the most in classifying the regular trafficking tweets with the trafficking tweets specifically attributed to advertising someone who is underage. Experiment 5 demonstrates the closeness between these two categories and the ability to separate them from regular tweets with a precision of 99% and recall of 100% when classifying the regular tweets against the two other trafficking classifications.

Background

The dataset needed to be created by scraping Twitter for each individual experiment's tweet queries. This data was then cleaned and processed into the count vector that was used during modeling. Prior to modeling the hashtags within the tweets were removed as they were used to search the tweets and may cause a large bias.

The data used in the modeling was collected from two separate sources. The first being a Kaggle data set of 16 million tweets from 2019 (P, Kaggle, 2019). The second source is from scraping twitter for the different hashtags using the website Phantom Buster (Phantom Buster, 2021). From the data scraping, three separate target data sets were created.

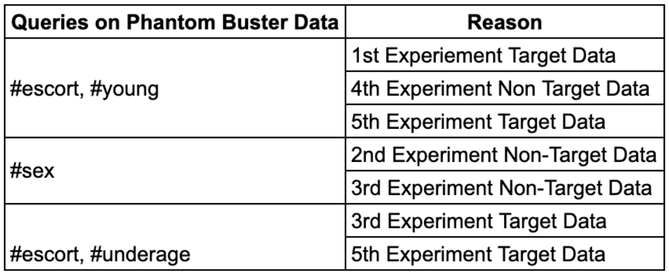
Data gathered from each scrape:



Phantom Buster	Kaggle
Handle	Username
Query	Query
Text	Tweet Content
Userid	
Name	
Tweet Link	
TimeStamp	

The following queries were run from the Phantom Buster website. Queries were run on different hashtags searches. The experiment the data was used in is listed in the second column.

Queries:



Queries on Phantom Buster Data	Reason
#escort, #young	1st Experiement Target Data
	4th Experiment Non Target Data
	5th Experiment Target Data
#sex	2nd Experiment Non-Target Data
	3rd Experiment Non-Target Data
#escort, #underage	3rd Experiment Target Data
	5th Experiment Target Data

Preprocessing needed to be done to create both the features for the SVM and to map the features from the Kaggle data set to the scraped data sets.

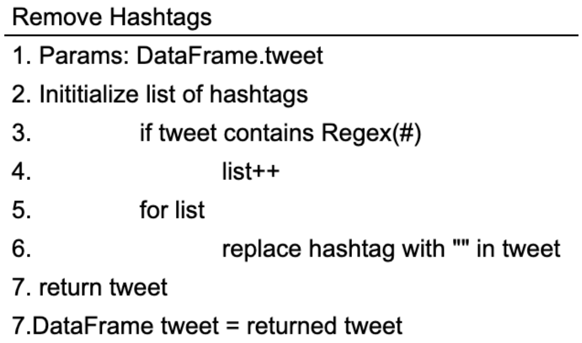
Preprocessing algorithms:

Different types of features were tried when experimenting with the SVM. The count vectorizer performed better than all of the boolean or limit count features created by smaller preprocessing algorithms. The following outlines the pseudocode and methodology used to clean the data and create the features.

Data Cleaning

Hashtag removal:

Hashtags needed to be removed in order because they are similar across all sex trafficking classes would cause an excessive amount of weight to be put on few features. Although it cannot be explicitly proven from this SVM algorithm seeks to classify tweets that would use similar language as those selling sexual acts with hashtags while not using hashtags in order to circumvent being banned from Twitter if caught.



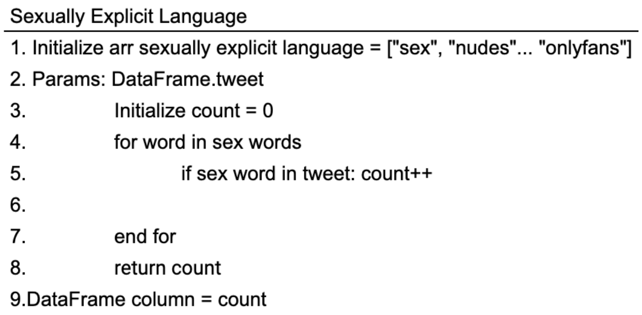
Remove Hashtags

```
Params: DataFrame.tweet
Inititialize list of hashtags
    if tweet contains Regex(#)
        list++
    for list
        replace hashtag with "" in tweet
return tweet
DataFrame tweet = returned tweet
```

Features

Sexually Explicit Language:

This feature counts the number of sexually explicit words used in the tweet. The sexually explicit words are generic terms which indicate sexual interactions.



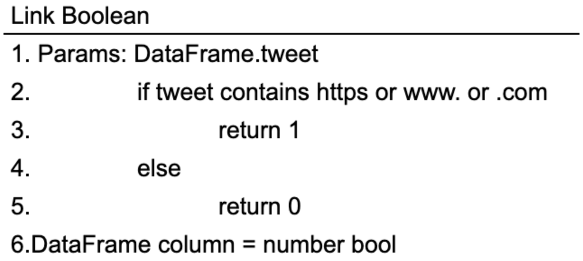
Sexually Explicit Language

```
Initialize arr sexually explicit language = ["sex", "nudes"... "onlyfans"]
Params: DataFrame.tweet
    Initialize count = 0
    for word in sex words
        if sex word in tweet: count++

    end for
    return count
DataFrame column = count
```

Link boolean:

Links are often used to indicate pages such as onlyfans.com, backpage.com, and other websites where there are more lenient rules and even encouraged sexual interactions between a buyer and seller. However, selling pornography is not a illegal but is indicative that the individual could be selling sex.

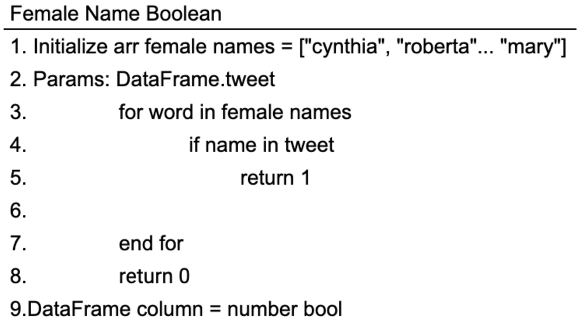


Link Boolean

```
Params: DataFrame.tweet
    if tweet contains https or www. or .com
        return 1
    else
        return 0
DataFrame column = number bool
```

Female name boolean:

In early exploratory data analysis, female names were noticed in many of the tweets from the same accounts when advertising different females. For this reason there is a female name boolean represented as either a 0 for none and a 1 for a female name existing.



Female Name Boolean

```
Initialize arr female names = ["cynthia", "roberta"... "mary"]
Params: DataFrame.tweet
    for word in female names
        if name in tweet
            return 1

    end for
    return 0
DataFrame column = number bool
```

Count Vectorization:
A count vector represents the frequency of a given word within the text as a feature within the text. When a count vector is created the words within the document provided to the vector are made into in this case a large numpy array that can be given to the SVM for processing. The model uses sci-kit learn’s CountVectorizer to fit and transform the tweet column into a large numpy array for each of the rows within a pandas DataFrame. A count vectorizer tokenizes the text it receives and assigns a unique number to each of the words within the document passed to the count vectorizer. For the modeling the size of the features was not limited.

Count Vectorizer - Bag of Words Pseudo Code

```
Initialize feature vector bg feature = [0,0,....,0]
for token in text.tokenzie() do
        if token in dict then
                token idx=getindex(dict, token)
                bg feature[token idx]++
        else
                continue
        end if
end for
return bg feature
```

Related Work

Previously, a similar study was done across South American countries at the Departamento de Informática y Ciencias de la Computación in Ecuador. This study used both a Naive Bayes algorithm and a Support Vector Machine. Their features included identifying hashtags which were used previously by sex traffickers, flagging links to sites known to be used by traffickers, and flagging if the tweet contained a photograph. In the first feature group, a feature counts sexually explicit language and creates a boolean value for the links within the tweets. With their research their SVM was able to achieve 85% precision (Hernandez, 2019).
The first group of features extends their successful work with a support vector machine while choosing not to pursue the naive bayesian network.

Project Description

The sex trafficking data is classified using a support vector machine. Modeling revealed the linear kernel performed the best when classifying the data. The hinge loss function is implemented by the SVM during training in order to create adequate decision boundaries for classification.

A basic SVM can be represented as:

$$\mathrm{sign}(w^T \phi(x) + b)$$

Where:

$$x_i \in \mathbb{R}^p$$

$$y \in \{1, -1\}^n$$

$$w \in \mathbb{R}^p|$$

$$b \in \mathbb{R}$$

Represents the state space.

This basic SVM represents the decision boundary between the different classification groups.

A linear SVM does not take advantage of the gaussian kernel trick where the classification can easily be represented in higher planes. The primal form of a linear SVM can be represented with the following equation (Sklearn, 2021):

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1} \max(0, y_i(w^T \phi(x_i) + b))$$

An SVM calculates the distance from the hyperplane to each point. If the dot product of the predicted value and the actual value is positive the SVM has correctly classified the data and if the dot product is negative the data is incorrectly classified.

The first term of the primal SVM shown above represents the support vector for perfectly separable data. Minimizing this first term represents identifying the closest point to the decision boundary.

The final term is making use of hinge loss to calculate the error within the SVM being trained. C adjusts the strength the penalty has on updating the weights.

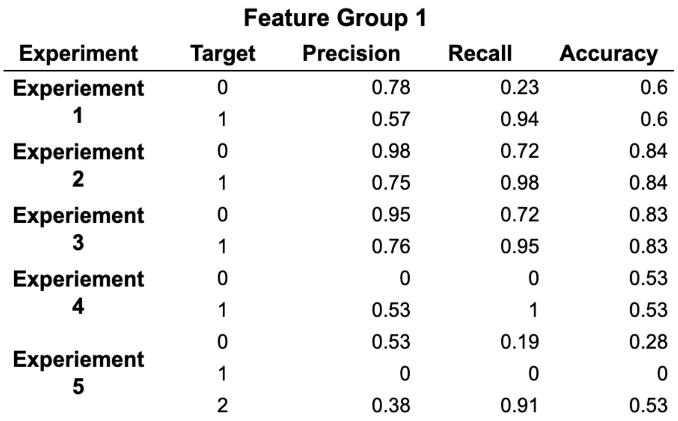
Experiments

There are ten separate experiments on five different tiers of data pairings to see how well the SVM was performing on classifying very different groups of data and very similar groups of data. The first tier contains random tweets compared to tweets containing the hashtags #escort and #young. The tiers became progressively more challenging for the SVM to classify. The second tier compared the #escort and #young data set with a twitter query for #sex. This was to ensure the SVM could classify against those tweets which were similar but not illegal. Tier 3 used a dataset with the hashtags #escort and #underage. This is likely the most illegal group because most states have laws against sex with a minor. Tier 4, compared #escort and #young with #escort and #underage. The final experiment was a multi classification problem comparing the random tweets with #escort and #underage as well as #escort and #young.

Data Tiers	Non-Target	Target	Reason
Tier 1	Random Tweets.	#escort and #young.	Base comparison.
Tier 2	#Sex	#escort and #young.	Harder Classification Problem
Tier 3	#Sex	#escort, #young, #underage	Harder to classify. Likely illegal activity.
Tier 4	#escort and #young.	#escort, #young, #underage	Multi - Classification Problem
Tier 5	Random Tweets.	Group1: #escort and young. Group 2: #escort, #young, #underage	Multi - Classification Problem

The ten experiments separated into two different feature groups. The first feature group contained the female name boolean, sexually explicit language count, as well as a link boolean. The results of these experiments can be seen below. However, this group did not yield the desired results due to a lack of features compared to the count vectorizer feature group. The results from feature group 1 are displayed below. The number 1 always represents the targeted sex trafficking data the model is seeking on classifyign. The number 0 represents random tweets, sexually explicit tweets, or similar tweets that have a broader scope than the tweets labeled as 1.

Feature Group 1 Results :

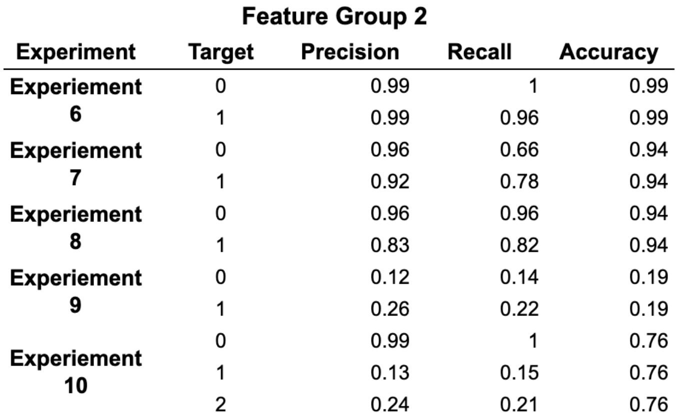


Feature Group 1

Experiment	Target	Precision	Recall	Accuracy
Experiement 1	0	0.78	0.23	0.6
	1	0.57	0.94	0.6
Experiement 2	0	0.98	0.72	0.84
	1	0.75	0.98	0.84
Experiement 3	0	0.95	0.72	0.83
	1	0.76	0.95	0.83
Experiement 4	0	0	0	0.53
	1	0.53	1	0.53
Experiement 5	0	0.53	0.19	0.28
	1	0	0	0
	2	0.38	0.91	0.53

Feature Group 2:

As expected, the first experiments performed much better than the latter. The first experiment performed the best with a near perfect precision and recall. This result is expected because the known differences between the datasets are greater. Experiment 2 performed worse than experiment 1 because of the similarities between the datasets. Experiment 3, had a similarly hard time classifying the #underage tags to the #sex tag. However, both results indicate the ability to classify against similar data. The following experiments showed how hard it is to identify between #underage & #escort tweets and those tweets that are just for #escorts. The results of feature group 2 can be seen below:



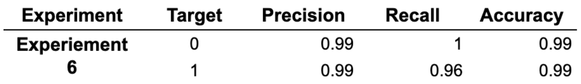
Feature Group 2

Experiment	Target	Precision	Recall	Accuracy
Experiement 6	0	0.99	1	0.99
	1	0.99	0.96	0.99
Experiement 7	0	0.96	0.66	0.94
	1	0.92	0.78	0.94
Experiement 8	0	0.96	0.96	0.94
	1	0.83	0.82	0.94
Experiement 9	0	0.12	0.14	0.19
	1	0.26	0.22	0.19
Experiement 10	0	0.99	1	0.76
	1	0.13	0.15	0.76
	2	0.24	0.21	0.76

As the feature data is analyzed the following graphs represent the probabilities that the feature is either of the different classified categories. It was impossible to graph all of the thousands of count vectorized features to show the clustering. However, the following graphs show the distribution of the data representing the trafficking tweet target in red and the comparison data in blue.

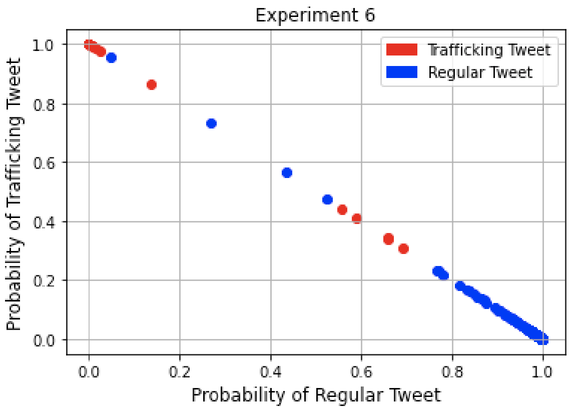
Experiment 6

Results:



Experiment	Target	Precision	Recall	Accuracy
Experiement 6	0	0.99	1	0.99
	1	0.99	0.96	0.99

Graphical Representation:

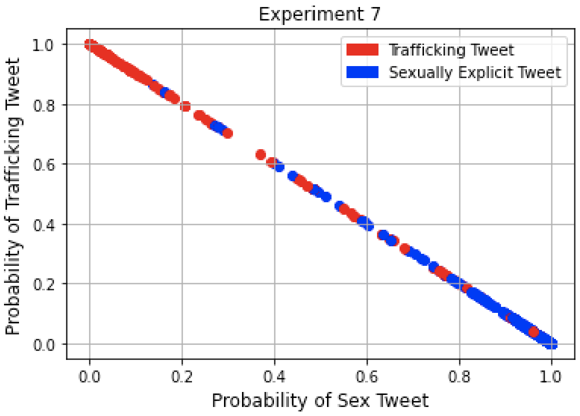


The graphical representation infers that the features created a wide support vector as seen in the major gap between the majority of the red scatter points observed in the upper left corner of the graph to the majority of the blue scatter points. This was the most successful classification of the group as it compared regular tweets with the first tier of trafficking tweets.

Experiment 7
Results:

Experiment	Target	Precision	Recall	Accuracy
Experiement 7	0	0.96	0.66	0.94
	1	0.92	0.78	0.94

Graphical Representation:

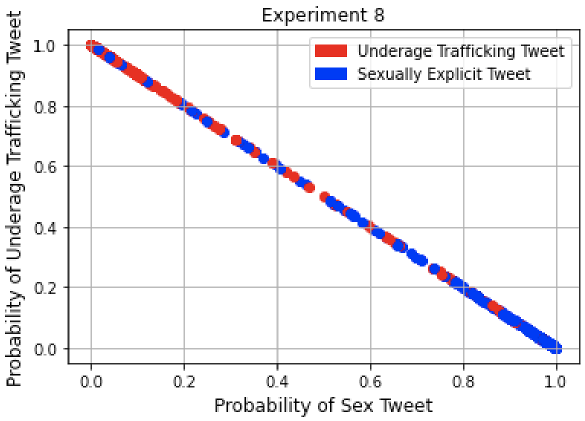


As seen in both the performance scores graphical representation the different classified scatter points are much closer together. Since the data contained the sex trafficking data as the 1 category target and tweets containing the #sex in the other there is likely overlap in the sexually explicit language. The language within the tweets is still unique enough for high precision. However, the recall which represents the correct positive predictions made is lower for both classification targets due to the similarities in language.

Experiment 8
Result:

Experiment	Target	Precision	Recall	Accuracy
Experiement 8	0	0.96	0.96	0.94
	1	0.83	0.82	0.94

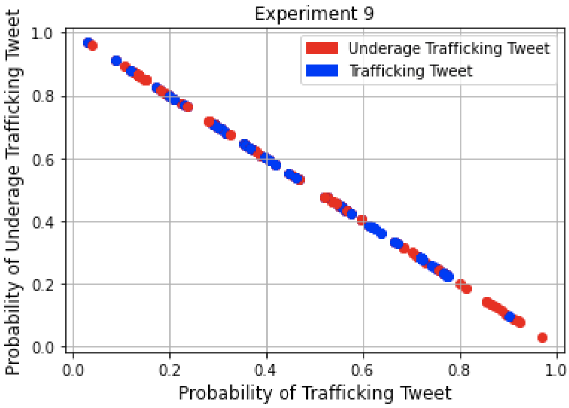
Graphical Representation:



Surprisingly, experiment 8 surpassed the results of experiment 7 in recall. This difference is likely due to differences in the datasets. Those advertising underage individuals may choose less sexually explicit language but still use unique language in their tweets.

Experiment 9

Experiment	Target	Precision	Recall	Accuracy
Experiement 9	0	0.12	0.14	0.19
	1	0.26	0.22	0.19

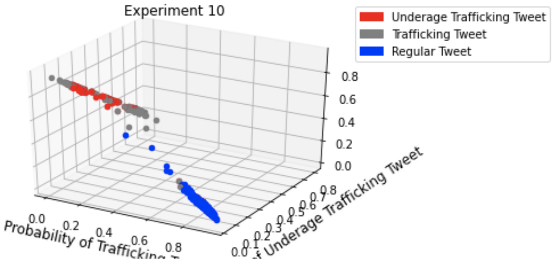


Experiment 9 demonstrates the lack of difference between the underage trafficking tweets and those that are not denoted to be underage. This is likely due to younger people being more likely to have a controller advertising for them.

Experiment 10
Results:

Experiment	Target	Precision	Recall	Accuracy
Experiement 10	0	0.99	1	0.76
	1	0.13	0.15	0.76
	2	0.24	0.21	0.76

Graphical Representation:



Experiment 10 compares the ability for the SVM to classify 3 different groups: underage trafficking tweets, a trafficking tweet, and regular tweets. The overlap in language similarities in both trafficking groups is demonstrated in the graph. The model as in experiment 1 will correctly classify regular tweets against trafficking tweets but cannot classify the sub-categories of trafficking tweets.

Conclusion

The ten experiments performed using a the linear kernel support vector machine demonstrate that tweets can be classified and flagged for their potential sex trafficking content. There is no reason why a query should have been able to be pulled that returns #escort and #underaged on twitter or any other social media websites. With the increase of trafficking through social media websites social media companies should be more vigilant in protecting and providing resources to their users when potentially harmful material is posted by that user. Two thirds of victims of trafficking report they never saw a helpline (Thorn, 2019). Further research to be done would be either applying or replicating this algorithm on websites like Facebook, Instagram, Reddit, and Craigslist. Alerting these companies of their ability to classify this data or providing them with the algorithm to do so is a promising way of limiting the amount of sex trafficking taking place on social media.

References

1.4. Support Vector Machines — scikit-learn 0.24.1 documentation. (n.d.). Sk-Learn. Retrieved April 27, 2021, from https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation

Hernández-Álvarez, M. H. A. (2019). Detection of possible human trafficking in Twitter. Departamento de Informática y Ciencias de la Computación - DICC Escuela Politécnica Nacional, Ecuador. https://orcid.org

"Human Trafficking Within and Into The United States: A Review of the Literature." Office of the Assistant Secretary for Planning and Evaluation. Accessed July 31, 2019, https://aspe.hhs.gov/report/human-trafficking-and-within-united-states-review-literature#Trafficking.

Kelly, C. U. T. (2019, July 30). 13 sex trafficking statistics that explain the enormity of the global sex trade. *USA TODAY*. https://eu.usatoday.com/story/news/investigations/2019/07/29/12-trafficking-statistics-enormity-global-sex-trade/1755192001/

P. (2019, January 2). *Twitter Sentiment Analysis*. Kaggle. https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis

Phantombuster. (2021). PHANTOM BUSTER. https://phantombuster.com/

Thorn Survivor Insights. (2019). https://www.thorn.org/survivor-insights/