Unstructured Data Analysis - Final Project

Sentiment Analysis of Stop, Question and Frisk Police Reports in New York City (2021)

00922719 Li-Vern Teo

This report was completed independantly by the author, without external input.

A supplementary Jupyter Notebook, containing the code used for this assessment, and the dataset used are provided as part of the submission, and in a GitHub repository (link here).

1 Introduction and Motivation

This report analyses a simplified version of the Stop Question and Frisk dataset provided by the New York Police Department (NYPD) in 2021. It seeks to understand if there is a link between an officer's description of a suspect's behaviour and the suspect's race and the officer's treatment of the suspect.

According to the New York Criminal Procedure Law 140.50, police officers may temporarily stop a member of the public if they "reasonably suspects that such person is committing, has committed or is about to commit either (a) a felony or (b) a misdemeanor". During the stop, the officers are allowed to question the individual and obtain private information about them (ie name, address etc), search the body of the individual if the officer "reasonably suspects that he is in danger of physical injury", and finally, issue an arrest without a warrant if any laws were found to be broken either before, or as a result of the stop (FindLaw 2021). This practice is commonly known as Stop, Question and Frisk in New York City. An equivalent practice in the UK is known as Stop and Search.

Stop, Question and Frisk has been widely criticised as a means of perpetuating systemic oppression towards African-Americans and Hispanic-Americans, with various lawsuits being raised by civil defence organisations against its usage (Devereaux 2012).

1.1 Stop, Question and Frisk dataset

The dataset used in this report was provided by the NYPD itself via its open data portal (Department 2021). The raw dataset from 2021 consists of 83 fields and 8,947 records. For the purpose of this report, the dataset was amended to keep only 34 columns of interest.

The following table details the columns that were kept and their descriptions

Column Name	Description
STOP_ID	Unique identifier code for stop
STOP_FRISK_DATE	Date of stop
ISSUING_OFFICER_RANK	Rank of issuing officer
SUPERVISING_OFFICER_RANK	Rank of supervising officer
SUSPECTED_CRIME_DESCRIPTION	Crime that the officer suspects is being committed
SUSPECT_ARRESTED_FLAG	Flag (Y/N) if the suspect was arrested
SUSPECT_ARREST_OFFENSE	Alleged crime that resulted in an arrest, if an arrest was made
FRISKED_FLAG	Flag (Y/N) if the suspect was frisked
SEARCHED_FLAG	Flag (Y/N) if the suspect was searched (e.g. car, house etc)

Column Name	Description
ASK_FOR_CONSENT_FLG	Flag (Y/N and null) if the officers asked for consent during frisk/search
CONSENT_GIVEN_FLG	Flag (Y/N and null) if consent was given by the persons stopped
OTHER_CONTRABAND_FLAG	Flag (Y/N) if some other contaband was found
FIREARM_FLAG	Flag (Y and null) if a firearm was found
KNIFE_CUTTER_FLAG	Flag (Y and null) if a knife was found
OTHER_WEAPON_FLAG	Flag (Y and null) if other weapons were found
WEAPON_FOUND_FLAG	Flag (Y and null) if a weapon was found
PHYSICAL_FORCE_CEW_FLAG	Flag (Y and null) if a taser was fired
PHYSICAL_FORCE_DRAW_POINT_FIREARM_FLAG	Flag (Y and null) if a firearm was drawn
PHYSICAL_FORCE_HANDCUFF_SUSPECT_FLAG	Flag (Y and null) if handcuffs were used
PHYSICAL_FORCE_OC_SPRAY_USED_FLAG	Flag (Y and null) if pepper spray was used
PHYSICAL_FORCE_OTHER_FLAG	Flag (Y and null) if other physical forced was used
PHYSICAL_FORCE_RESTRAINT_USED_FLAG	Flag (Y and null) if force was used to restrain the suspect
PHYSICAL_FORCE_VERBAL_INSTRUCTION_FLAG	Flag (Y and null) if a verbal instruction was issued
PHYSICAL_FORCE_WEAPON_IMPACT_FLAG	Flag (Y and null) if a weapon was used
DEMEANOR_OF_PERSON_STOPPED	Description of demeanour of the person stopped
SUSPECT_REPORTED_AGE	Reported age of suspect
SUSPECT_SEX	Sex of suspect
SUSPECT_RACE_DESCRIPTION	Race of suspect
SUSPECT_HEIGHT	Height of suspect
SUSPECT_WEIGHT	Weight of suspect
SUSPECT_BODY_BUILD_TYPE	Build tyope of suspect
SUSPECT_EYE_COLOR	Eye colour of suspect
SUSPECT_HAIR_COLOR	Hair colour of suspect
STOP_LOCATION_BORO_NAME	Name of borough where stop was initiated

The columns that will be used to determine the officer's description of the suspect is the `DEMEANOR_OF_PERSON_STOPPED` column. The suspect's race is reported by the `SUSPECT_RACE_DESCRIPTION` column (it is not clear from documentation if this description was provided by the officer or the suspect), with the physical treatment of the suspect described by boolean columns with the `PHYSICAL_FORCE` prefix.

1.2 Report Strucuture

The report begins with some exploratory data analysis that seeks to understand:

- The breakdown of stops by race
- The likelihood of being stopped by officers, normalised to the suspect's race
- The propensity of officers to use 'severe physical force'
- The most common words used to describe suspects

Three different natural language processing techniques were then tested as a means of determining a link between the officer's description of the suspect, and the suspect's race. The techniques used were:

- Self-trained word embedding model
- Pre-trained word embedding model
- VADER sentiment analysis tool and Feature Importance testing.

The report concludes with a description of the effectiveness of each technique, and provides closing arguments regarding the outcome of the analysis.

Complete details of all code used in this report can be found on a supplementary notebook in this github repository

2 Exploratory Data Analysis

Following data cleaning and preprocessing, the dataset was reduced to 8,215 usable entries in 2021. It showed that only 37.6% of stops led to an arrest, indicating that the majority of suspects stopped did not possess any contraband, nor was there sufficient proof of illegal activity in progress.

Figure 1(a) shows the racial breakdown of suspects stopped. Black and (Black/White) Hispanic suspects make up the vast majority of stops (~87%), and this was true after normalising for the racial breakdown in New York City. (Figure 1(b))[1]

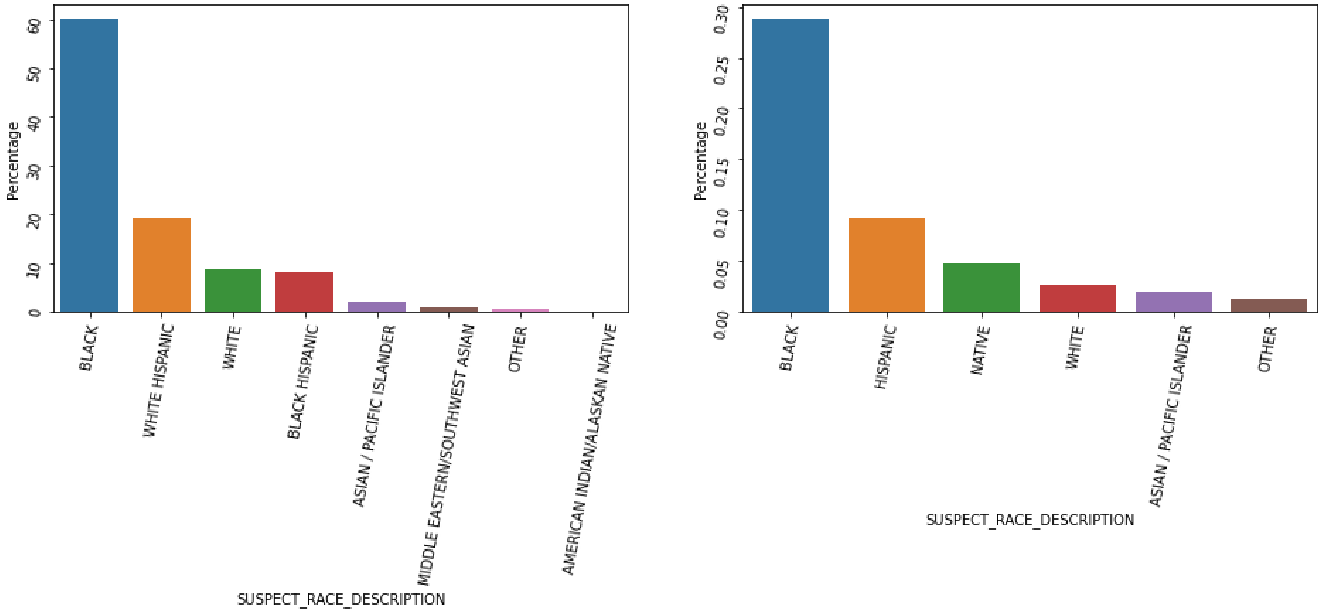


Figure 1: (a) Percentage breakdown of 2021 stops by race, (b) 2021 stops as a percentage of the demographic breakdown of New York City (as described by the 2021 Census)

Where severe physical force was used in the stop (defined by the author as the use of a taser, pepper spray, the drawing of a firearm and/or a weapon of impact by the officer), Figure 2 shows that Black suspects were slightly more likely to experience severe physical force, followed by White Hispanic, White and Black Hispanic suspects. Although, the use of severe physical force was only limited to about 7% of all stops.

Figure 3(a) shows the top 17 most common words used to by officers to describe suspects. With the complete corpus being represented in the word cloud in Figure 3(b).

The most common words showed a mixture of positive or neutral sentiment such as: 'calm', 'cooperative', 'compliant', 'understanding', and negative sentiments such as: 'nervous', 'angry', 'annoyed', 'irate', 'aggressive' etc.

The analysis thus far supports the claims by civil rights groups that Black and Hispanic citizens are disproportionately stopped by officers, while Black and White Hispanic suspects were at a slightly higher risk of experiencing severe physical force than White suspects. These statistics, alongside the fact that the majority of these stops do not lead to an arrest, calls into question the effectiveness of the program itself, and the risk it poses to the discrimination of Black and Hispanic citizens.

[1] Demographics of New York City were obtained from the 2021 US Census. The census reports race differently from the NYPD's database and so care has been taken to group up race descriptions where applicable (e.g. "Black Hispanic" and "White Hispanic" have been grouped to "Hispanic")

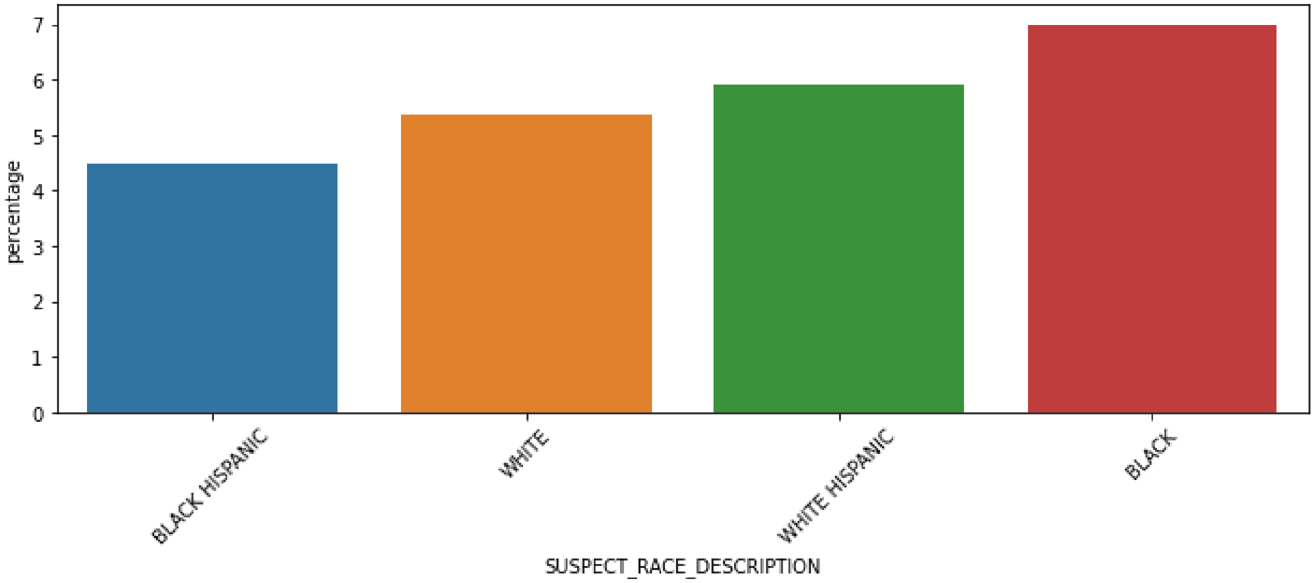
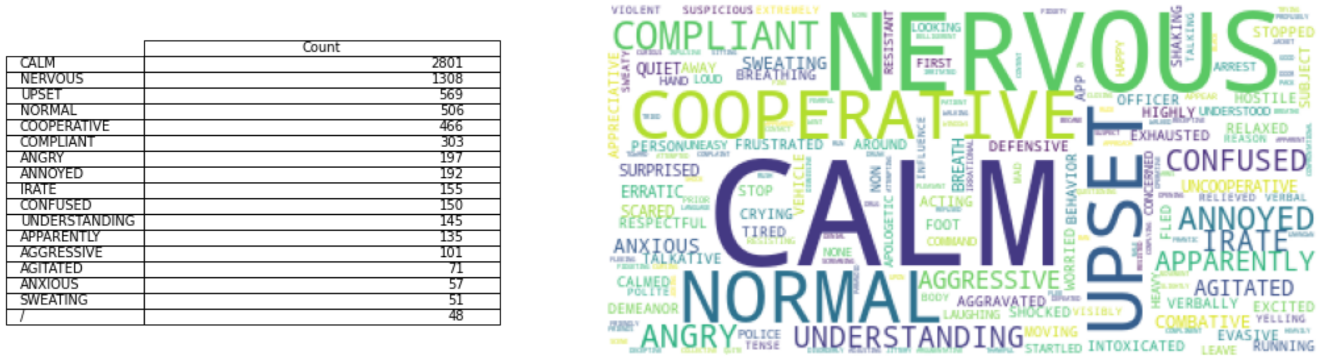


Figure 2: Percentage of stops where severe physical force was used, by suspect race



	Count
CALM	2801
NERVOUS	1308
UPSET	569
NORMAL	506
COOPERATIVE	466
COMPLIANT	303
ANGRY	197
ANNOYED	192
IRATE	155
CONFUSED	150
UNDERSTANDING	145
APPARENTLY	135
AGGRESSIVE	101
AGITATED	71
ANXIOUS	57
SWEATING	51
/	48

CALM
NERVOUS
COOPERATIVE
NORMAL
UPSET
COMPLIANT
CONFUSED
ANNOYED
IRATE
APPARENTLY
AGGRESSIVE
UNDERSTANDING
ANGRY
AGITATED
ANXIOUS
VIOLENT
SUSPICIOUS
EXTREMELY
QUIET
AWAY
SWEATING
BREATHING
HAND
LOUD
LOOKING
FIRST
RESISTANT
SHAKING
TALKING
STOPPED
ARREST
HAPPY
SUBJECT
OFFICER
HOSTILE
HIGHLY
UNDERSTOOD
RELAXED
EXHAUSTED
REASON
APPRECIATIVE
PERSON
UNEASY
FRUSTRATED
AROUND
DEFENSIVE
SURPRISED
ERRATIC
STOP
VEHICLE
INFLUENCE
BREATH
BEHAVIOR
CONCERNED
UNCOOPERATIVE
RELIEVED
VERBAL
SCARED
RESPECTFUL
CRYING
TIRED
ACTING
FOOT
COMMAND
NON
APOLOGETIC
WORRIED
FLED
TALKATIVE
NONE
CALMED
POLITE
DEMEANOR
POLICE
TENSE
AGGRAVATED
BODY
LAUGHING
SHOCKED
VISIBLY
HEAVY
VERBALLY
EXCITED
COMBATIVE
YELLING
EVASIVE
MOVING
STARTLED
INTOXICATED
LEAVE
RUNNING

Figure 3: (a) Top 17 words in the demeanor column and their word count (b) Word cloud of corpus in demeanor column

Meanwhile the descriptions of suspects showed a wide range of sentiments in the Top 17 words used to describe suspects during a stop. At this point, no one sentiment or word stands out from the word cloud or the table, and further NLP techniques will be used to further investigate this field.

3 Word Embeddings - Self trained model

3.1 Introduction to Word Embeddings

Word embeddings is a Natural Language Processing technique that represents words as vectors in a vector space. The most common Word Embeddings applications use a neural network to determine word vectors by either guessing the context that a word is used (Skip-Gram) or predicting a word to be used in a given context (Continuous Bag of Words).

In a vector space, sentences with similar vocabulary and meaning are expected to fall in and around the same area. The cosine-similarity metric is often used to measure the similarity between 2 word-vectors.

For sections with Word Embedding applications, the `Word2Vec` algorithm will be used.

3.2 Application

Word Embedding models are data hungry, and required datasets that provide sentences with rich meanings to determine the context in which words are used. For this section, the `20-newsgroups` benchmark dataset was used to supplement this need and train the model. This dataset was chosen because

- It consists of articles from a wide range of topics
- The vast corpus reduces the risk of the Stop, Question and Frisk dataset containing out of vocabulary words
- It was hoped that the nature of the dataset (ie articles written by various authors of numerous topics) would be suited to better encoding the sentiment of vocabulary words.

Pre-processing was done by standardising the dataset to be all lowercase, then punctuations, single character words and stop words were removed, and finally the remaining corpus was lemmatised[2] and documents were returned in the form of tokens. The `Word2Vec` model was then trained with the preprocessed output with the following parameteres

```
Word2Vec(corpus, min_count=5, vector_size=2)
```

instructing the model to only use words which occured at least 5 times in the corpus, and to produce vectors with 2 dimensions.

The Stop, Question and Frisk dataset was pre-processed with the same procedure as in `20-newsgroup`, with the addition of a spell-check step prior. Spell-check was completed using the `SpellChecker` package and is an important step as the `DEMEANOR_OF_PERSON_STOPPED` field was liable to human error and mistakes in spelling will result in out of vocabulary words.

With Word vectors were obtained by using the `Word2Vec.wv['word']` method. The following processes were considered when calculating document vectors

- Where the `'word'` was out-of-vocabulary (ie it did not exist in the `20-newsgroups` dataset) no vector could be returned for that word and the observation was allowed to be `np.NaN`
- Where there was more than one word used to describe a suspect (e.g. `appeared normal` or `nervous and sweating`) the average of the vectors of each word (after preprocessing) was used.

[2] Lemming was chosen instead of stemming in this case as lemming keeps a recognisable form of the word (e.g. the lemma swimming is swim, while the stem of swimming is swi). Which simplifies data analysis and supports readability of this report

Figure 4(a) shows the outcome of vectors as determined by `Word2Vec`. Each point represents a document vector, with different point colours representing the suspect's race. There was no clustering behaviour present. Figure 4(b) shows the same plot, overlayed with the positions of the Top 10 words in the dataset. It is seen that most words fall into the central cluster, regardless of the sentiment of the words, which might be the cause of the lack of identifiable clustering in Figure 4(a)

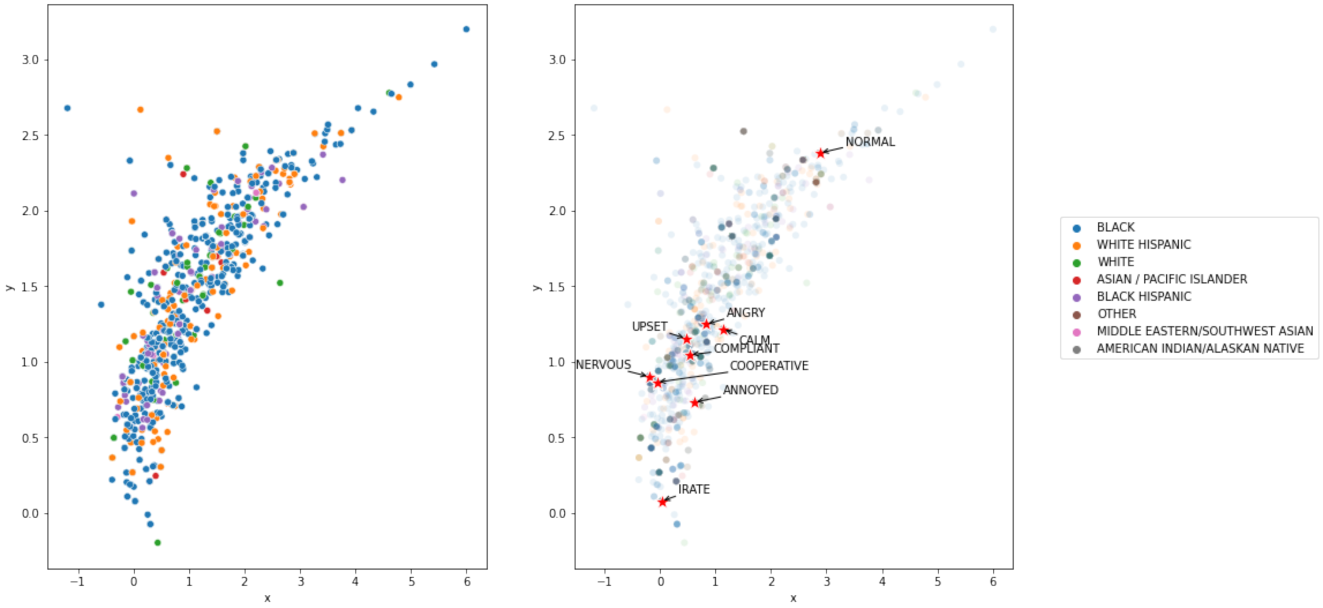


Figure 4: (a) Document vectors for each demeanor record by race (b) Same as in (a) but with Top 10 words overlay

4 Word Embeddings - Pre-trained model

4.1 Introduction to Pre-trained models

As mentioned in Section 3.1, word embedding models require a significant amount of data and preprocessing for optimal performance. It is common practice to download and use vectors from pre-trained models instead which provide the following benefits:

- it is memory efficient - the user won't need the original dataset used, just the output (ie vectors)
- it is computationally efficient - the model is already trained and all is required by the user is to perform a lookup on the model output.

This section uses the `glove-wiki-gigaword-100` (Gensim 2018) provided by Gensim. The original model was trained on a 2014 extract of Wikipedia and Gigaword, amounting to 5.6 billion tokens and a 400k word vocabulary. The model itself consists of pre-trained vectors of 100 dimensions and is only 128MB in memory.

4.2 Application

As before, the individual word vectors were determined using the outcome of the pre-trained models, and document vectors were calculated by the mean of all word vectors in an observation. It should be noted that this method only resulted in 19 unusable rows due to out of vocabulary words, where the self-trained model resulted in 657 rows being dropped from the dataset.

The vectors produced had 100 dimensions, and as such, patterns cannot be easily identified through visual inspection. Instead, density based spatial clustering was used to identify clusers.

Density based clustering is a method for finding clusters by looking for nearest neighbours. The two parameters which determine the size and density of a cluster are `eps` which correspond to the minimum acceptable distance between 2 points for them to be considered neighbours, and `min_samples` which correspond to the minimum number of samples needed for a core point to be determined.

Density based clustering was applied to our vector space using the `DBSCAN` class from `sklearn` with the following parameters

```
clustering = DBSCAN(eps=2.5, min_samples=50).fit(X)
```

which grouped the data points into 6 clusters, numbered `[-1, 0, 1, 2, 3, 4]`

Heatmaps were then used to identify the makeup of each cluster and to identify patterns within them

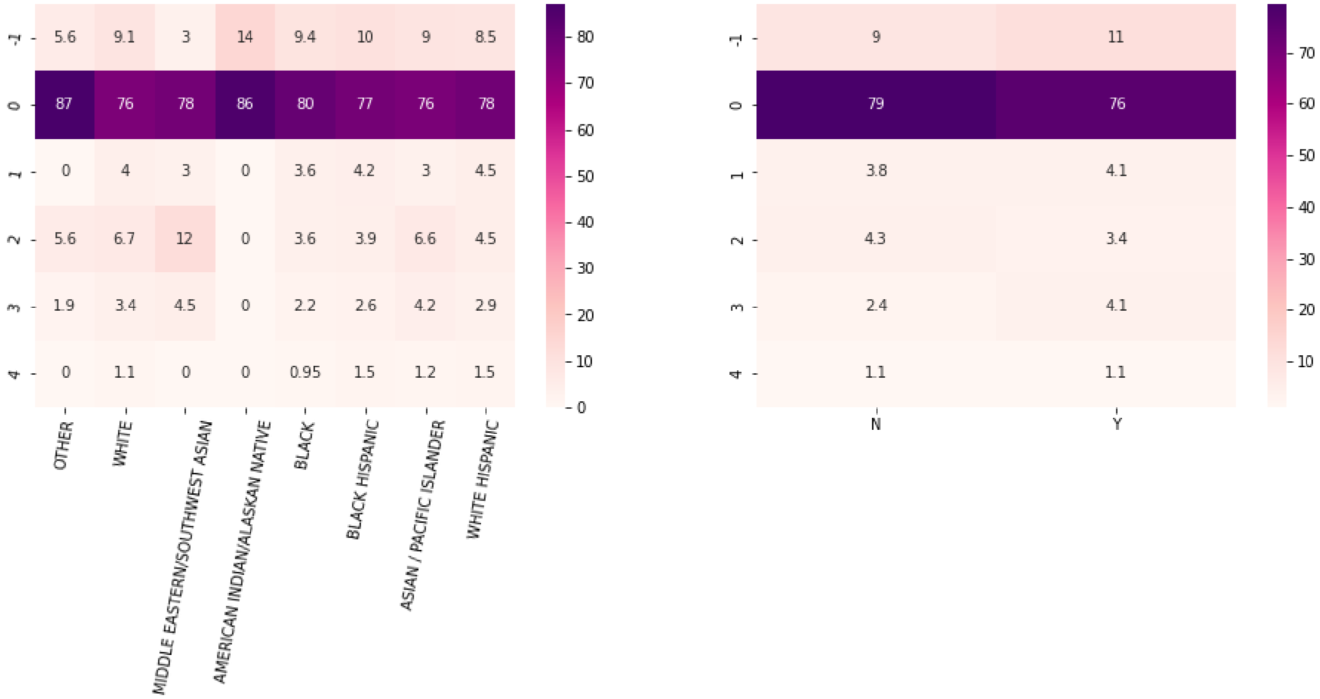


Figure 5: (a) Racial makeup of each cluster, represented as a percentage of the racial makeup of the dataset (b) Use of severe physical force in each cluster, represented as a percentage of instances of severe physical force in dataset

Figure 5 shows that the majority of all data points were allocated to Cluster 0, followed by Cluster -1. The percentage distribution of points in the other clusters were too small for any noticeable patterns to be inferred.

One might infer that the behaviours observed in the lower dimensional space (Figure 4) are being extrapolated to this high dimensional space. A possible explanation is that the word embedding models are grouping words by their type rather than the sentiment (ie the words in this vocabulary are overwhelmingly about 'feelings', and the models are not separating vectors by the meaning of the 'feelings'). Although the nature of the model (shallow neural network) and high dimensional space make it impossible to definitively conclude if this is true.

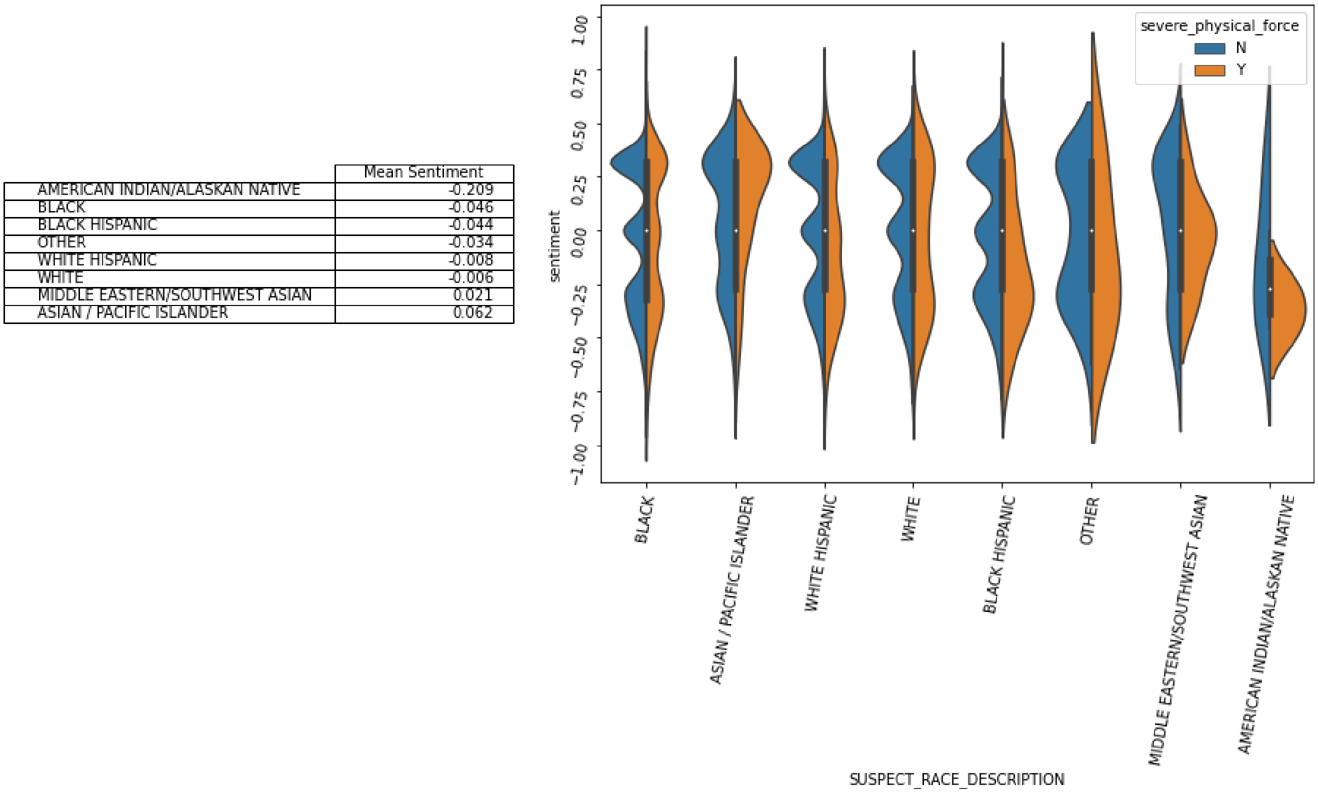
5 VADER Sentiment Analysis and Feature Importance

5.1 VADER

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a corpus with pre-labelled sentiments. It uses a rule based system to assign a sentiment score (ranging from -4 to +4) to complete sentences. VADER is also able to use compounding words (e.g. 'very', 'too' etc) and sentiment changing lexicon (e.g. 'not', 'non', etc) to give deeper meaning to the sentiment of the sentence as a whole, rather than just analysing tokens on their own.

5.2 Application

In order to preserve stop words and sentence structure, only the spell check preprocessing step was applied to the dataset. The distribution of sentiment scores produced by VADER is found in Figure 6



	Mean Sentiment
AMERICAN INDIAN/ALASKAN NATIVE	-0.209
BLACK	-0.046
BLACK HISPANIC	-0.044
OTHER	-0.034
WHITE HISPANIC	-0.008
WHITE	-0.006
MIDDLE EASTERN/SOUTHWEST ASIAN	0.021
ASIAN / PACIFIC ISLANDER	0.062

Figure 6: (a) Mean sentiment score by suspect race (b) Distribution of sentiment scores by race and use of severe physical force in stops

Figure 6 shows sentiment was skewed significantly negative for suspects that were American Indian/Alaskan Native, and slightly negative for Black, Black & White Hispanic, Other and White suspects. The only suspects that showed positive sentiment skews appeared to be those of Asian/Pacific Islander and Middle Eastern/Southwest Asian descent.

There was a lack of significant variation in sentiment distribution in cases where severe physical force was used and not. Notably it is seen that for stop of Asian/Pacific Islander suspects, sentiment still skewed positive, even when severe physical force was used.

5.3 Determining feature importance

A random forest model was then used to determine the factors that contribute to the sentiment expressed by officers. The model uses only 26 features from the dataset, with the following features removed:

Dropped Column	Justification
STOP_ID	Perceived to lack relevance to sentiment
STOP_FRISK_DATE	Perceived to lack relevance to sentiment
SUSPECT_ARREST_OFFENSE	Mostly null values
DEMEANOR_OF_PERSON_STOPPED	Encoded as part of sentiment
SUSPECT_REPORTED_AGE	Perceived to lack relevance to sentiment; encoded as part of build
SUSPECT_HEIGHT	Encoded as part of body build type

Dropped Column	Justification
SUSPECT_WEIGHT	Encoded as part of body build type
SUSPECT_EYE_COLOR	Perceived to lack relevance to sentiment
SUSPECT_HAIR_COLOR	Perceived to lack relevance to sentiment

Additionally, sentiment was also encoded into 5 equally distributed bins, representing 'Very Negative', 'Negative', 'Neutral', 'Positive' and 'Very Positive' sentiment respectively. This was due to the fact that the features used where categorical in nature, and as such, categorical labels was more appropriate.

The trained model had an accuracy of 89% on the dataset, which gives us some confidence to the reported feature importance scores plotted in Figure 7

Figure 7 shows that the most important feature in predicting sentiment was `SUSPECTED_CRIME_DESCRIPTION`, followed by `STOP_LOCATION_BORO_NAME`, `SUSPECT_BODY_BUILD_TYPE` and `SUSPECT_RACE_DESCRIPTION`. It also showed the ranks of the issuing and supervising officer to be of importance.

This suggests that the suspect's race was a significant factor in the officer's sentiment towards to suspect, however, it was not as important as the suspected crime, the borough in which the stop was made and the build of the suspect themself.

6 Conclusion

The statistics of the dataset showed that Black, (Black/White) Hispanic and Native Americans where more likely than White, Asian/Pacific Islander or Other citizens to be targeted by the police in a Stop, Question and Frisk stop. This validates concerns by civil rights activities and provides proof of systemic inequality in policing activity. The hypothesis of this study focused on determining if there was a link between the demeanor of the suspect, as reported by the officer(s) who made the stop, and the suspect's race.

Three techniques where tested as part of this analysis: (i) word embeddings on a self-trained model, (ii) pre-trained word embedding model, (iii) VADER sentiment analysis and feature importance. Both word embeddings models proved to be poor at identifying patterns in the dataset. Most datapoints in the vector space were grouped into a single cluster, making it difficult to identify any separating qualities of the data.

The sentiment analysis tool showed more promise. Analysis of distributions showed the sentiment expressed towards suspects that were American Indian/Native, Black, Black Hispanic, Other, White Hispanic and White skewed slightly to the negative, while Middle Eastern/Southwest Asian and Asian/Pacific Islanders had slightly more positive sentiment expressed. An additional observation from the distribution plot was that there was not much difference between the distribution of sentiment and the use of severe physical force. With some exceptions for American Indian/Alaskan Native, Black Hispanic and Middle Eastern/Southwest Asian suspects, where the sentiment distributions where more heavily skewed negative in cases where severe physical force was used.

A Random Forest model was then trained to predict sentiment using 26 other features provided in the dataset. The model produced had an accuracy score of about 89%, and showed that the top 5 most important features in predicting sentiment were:

1. The suspected crime being committed (as reported by the officers)
2. The New York borough in which the stop was initiated
3. The build of the suspect
4. The suspect's race
5. The rank of the supervising officer

This shows that while race is not the single most important factor in the expressed sentiment, it's significance in the model's decision making is notable, and is likely to be a compounding factor in addition to the top 3 features in the dataset.

The feature importance exercise also revealed that the Top 5 most important features were unrelated to the outcome of the stop itself (e.g. if weapons or contraband was found, if physical force was used by the officer or even the

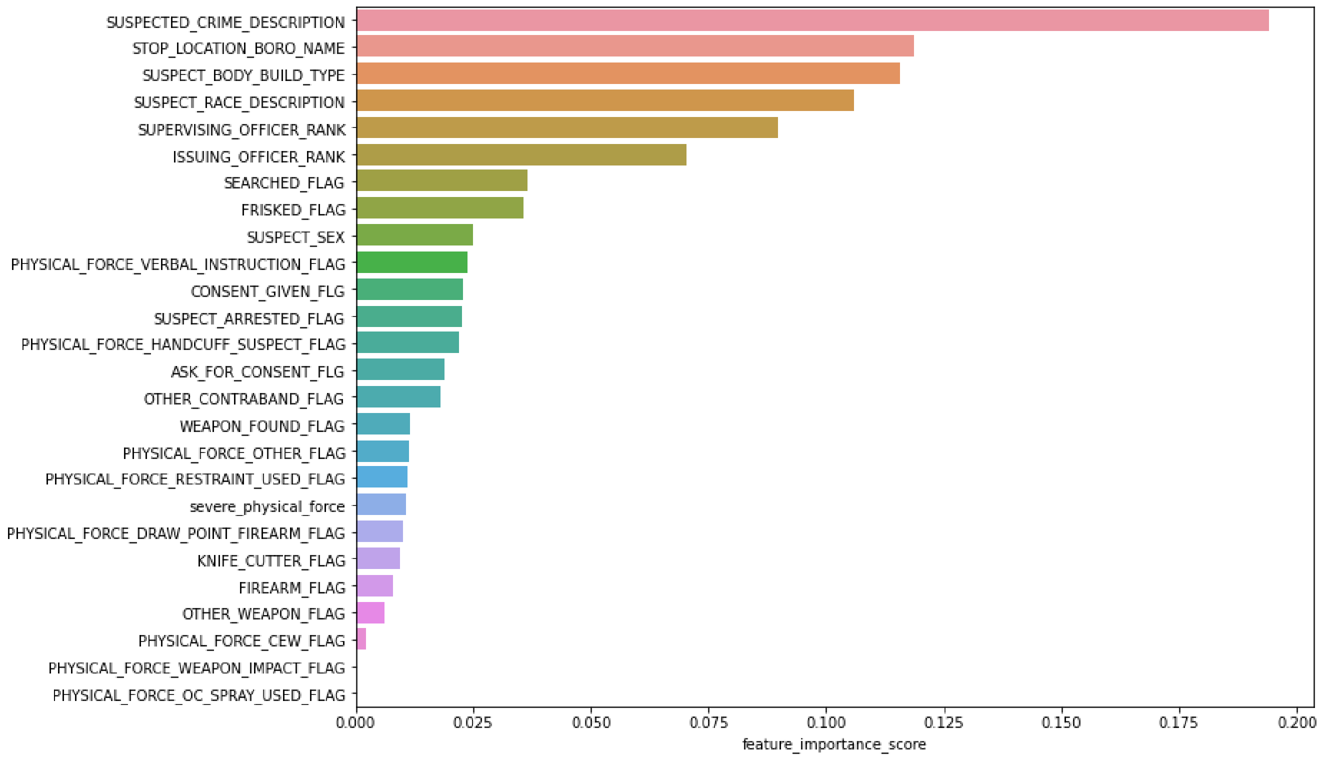


Figure 7: Feature importance scores of features used to predict sentiment

guilt/innocence of the suspect), and were instead all factors relating to the officers themselves, their perception of the suspect, and the physical characteristics of the suspect, all of which is out of the suspect's control. The outcome of this study suggests sentiment analysis has the potential to uncover unconscious or hidden biases in an officer's decision making when conducting a Stop, Question and Frisk stop.

References

Department, New York Police. 2021. "Stop, Question and Frisk Data." https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page.

Devereaux, Ryan. 2012. "Lawsuit Alleges NYPD Violated Civil Rights by Entering Private Buildings." *The Guardian*. https://www.theguardian.com/world/2012/mar/28/nypd-lawsuit-civil-rights-bronx.

FindLaw. 2021. "New York Consolidated Laws, Criminal Procedure Law - CPL § 140.50 Temporary Questioning of Persons in Public Places; Search for Weapons." https://codes.findlaw.com/ny/criminal-procedure-law/cpl-sect-140-50.html.

Gensim. 2018. "RaRe-Technologies/Gensim-Data." https://github.com/RaRe-Technologies/gensim-data.