Sentiment Analysis of Twitter Data to Determine the Accuracy of the "Karen" Archetype

CIS 400 - Spring 2021

Janet Lee

Raffi Mendoza

Gregory Morneault

Maxwell Reed

Maricel Vicente

Introduction - Karen Mythology:

On social media, a common sight to see is a woman named Karen that strikes fear into the hearts of every manager as she walks in. And now, in the era of coronavirus, “Karen” has made a resurgence, with a slight twist. Karen’s are mostly known now as those who refuse to get the vaccine, as well as being, “adopted as a shorthand to call out a vocal minority of middle-aged white women who are opposed to social distancing, out of either ignorance or ruthless self-interest,” wrote the Atlantic’s Kaitlyn Tiffany. Here, we are asking the question if there is a relationship between having the name Karen and possessing the characteristics of this archetype.

What does it mean to call a woman a “Karen”? The origins of any meme are hard to pin down, and this one has spread with the same intensity as the coronavirus, and often in parallel with it. Karens are “the police women of all human behavior.” Karens don’t believe in vaccines. Karens have short hair. Karens are selfish. Confusingly, Karens are both the kind of petty enforcers who patrol other people’s failures at social distancing, and the kind of entitled women who refuse to wear a mask because it’s a “muzzle.” They are likely to be a woman at the age of 40 or over and hold microaggressions against racial groups.

Approach:

To approach our project, we developed a twitter data mining bot to grab tweets with given characteristics such as the name, location, and contents of the tweet itself to gather our data and see if we can find evidence towards or against the claim of a relationship in negative sentiment and one’s name. For us to tackle this feat, we had to understand the characteristics of Karens and current events. We know that the Karen meme was famously used in the beginning of the COVID-19 pandemic, so we evaluated the current events since then. We knew we wanted to take the approach of gathering tweets of Karens about topical current events. We started with the generalization that most Karens are likely to be against the mask mandate, COVID-19 vaccines and social distancing. We took this generalization to generate our keywords “mask” and “vaccine”. In return, these keywords yielded tremendous results with tweets from Karens with negative sentiment.

Our technique for approaching a large problem started by breaking it down into mini sections, so it was easier for us to understand. For our other keywords, we used the same approach of starting with a generalization and hypothesizing keywords in that segment to yield the most accurate results. Each team member applied this method towards finding significant keywords, which would give us the most balanced data. Early on, we knew that if we collected negative sentiment using negative keywords, our data would be biased. In order to combat that conflict, we chose keywords which were impartial towards any topic positive or negative. To prove our hypothesis, we wanted the tweets mined to be a mixture of positive and negative sentiment, so we could have a general idea of the sentiment for that topic. The beginning of this approach was vital for us to continue our development of the project and practice good techniques to gather data.

In order to gather this data, we took advantage of the free Twitter Python API where we could specifically call the data we are seeking. This also had the added feature of us being able to filter accordingly to minimize the amount of noisy data that would complicate our work.

The following from our approach segment of this report is a breakdown of the code that we used. We will address here what the given lines of code displayed are intended for and how they relate to the project in large.

```
# Looking for Twitter users that have that name
userName= "Mary"
# This prints out tweets and tweet statistics to a text file
sys.stdout = open('tweets.txt', 'w')
# This makes a clean text file each time its run so save the data if you need it later
sys.stdout.close()
```

To start with our mining script, we set up the basic aspects of the code. We started with determining the name for which we will look for among our Twitter users (in this case, we were looking for women named Mary).

Following the declaration of the name is the generation of a text file for which we would store the data we collect for future analysis and closing the tweets.txt file so we can append to the written file later. By starting with writing the file, it would automatically delete old versions for us so we can save some time by not having to change to a new file every time.

```
# Getting users with userName in their twitter handle/name
# Each iteration of results gets the Maximum number of 20 users, hence why we call it 5 times
results = twitter_api.users.search(q=userName, page=1)
results2 = twitter_api.users.search(q=userName, page=2)
results3 = twitter_api.users.search(q=userName, page=3)
results4 = twitter_api.users.search(q=userName, page=4)
results5 = twitter_api.users.search(q=userName, page=5)

# From the accounts that we just mined, we grab their screen_name and location
tweets1 = [(r['screen_name'], r['location']) for r in results]
tweets2 = [(r['screen_name'], r['location']) for r in results2]
tweets3 = [(r['screen_name'], r['location']) for r in results3]
tweets4 = [(r['screen_name'], r['location']) for r in results4]
tweets5 = [(r['screen_name'], r['location']) for r in results5]

# Combining this information into one giant List
tweets1.extend(tweets2)
tweets1.extend(tweets3)
tweets1.extend(tweets4)
tweets1.extend(tweets5)
```

For the first segment of the snippet above, we are making five separate searches for the username to give us a total of 100 separate users to grab data from. In the second segment we cycle through each of the results we received in the form of a line comprehension to obtain the respective screen_name and location values held within the results. The third segment was put together to have all the data collected in one big list so as to have a centralized resource for the user names and locations.

```python
# These variables will track total tweets, positive and negative tweets, as well as the confidence score
tweetCount = 0
posCount = 0
negCount = 0
confidenceScore = 0

sentimentanalyzer = pipeline("sentiment-analysis")
```

In the snippet above, we have the last stages of initialization for values to determine the estimated net sentiment of all tweets received, to keep track of the number of tweets that met our criteria, and to calculate the confidence score of the data obtained. The *sentimentanalyzer* variable is pipelined from the sentiment analysis tool of the imported module *Transformers*.

```python
for (name,location) in tweets1:
            karenTweets = harvest_user_timeline(twitter_api, screen_name=name, max_results=500)
            #Gathering the 500 most recent tweets from the given screen_name
            sys.stdout = open('tweets.txt', 'a') #appending information to the text file
            print("////////////////////////////////////////NEXT USER////////////////////////////////")
            print(name)
            print(location.encode('utf8'))
            for i, t in enumerate(karenTweets): # Parsing through the user's tweets
                if ('Trump' in t['text'] or 'Biden' in t['text']): 
                    # This is where we insert keywords to filter the tweets through, in this case for out Presidents category
                    # if (not t['retweeted'] and 'RT @' not in t['text']): # We had this line for when we wanted to see only original tweets
                        try:
                                tweetCount = tweetCount + 1
                                d = sentimentanalyzer(t['text']) # Transformers package performing sentiment analysis on the tweet
                                tweetLabel = d[0].get('label') # Gets the label
                                tweetScore = d[0].get('score') # Gets the confidence score
                                if (tweetLabel=='POSITIVE'):
                                    posCount = posCount + 1
                                else:
                                    negCount = negCount + 1
                                confidenceScore = confidenceScore + float(tweetScore)
                                print("\n")
                                print(i, t['text'].encode('utf8')) # prints out the tweet
                                print(d) # prints out the label and confidence score
                        except:
                                pass
```

In the snippet above, we have the for loop to test all of the data we have obtained for these users working from the name and location that we were able to receive. We then open the sys file *tweets.txt* backup to append items we find which include printing the username and encoding 'utf8' in the location which converts the location into the form of Unicode Transformation Format which can allow us to compensate for special characters into the string that may otherwise not be include in ASCII standard formatting.

Getting into the inner for loop, we begin to iterate through the user tweets available on file looking for specific keywords that we set up to filter the tweets to specific topics (in this case relating to presidents Biden and Trump). If we have one of our keywords found within a tweet, we start with setting up a try-catch exception in order to keep the script from crashing and allowing us to grab the rest of the viable information from our users having to start all over again. Within our try statement, we take on the three risky tasks of getting the sentiment of the text with the Transformers package, get the label associated with the tweet, and to get the confidence score of the tweet.

If any of these three were to fail, we would be okay as we have made the iteration for the tweet count and the error indicating to us that one of three desired pieces of information is missing, we can leave it behind knowing we are not missing anything important for the project.

Otherwise, if we grab all three pieces of information without crashing, the score achieved will be allocated accordingly in the posCount or negCount variables and we can print the tweet that met our criteria alongside the label and confidence scores we found.

```
averageScore = confidenceScore / float(tweetCount) # calculates average confidence score for all tweets from harvested users
positiveScore = (posCount / tweetCount) * 100 # calculates the percentage of tweets that were positive
print(f"This program read {tweetCount} tweets")
print(f"Number of Positive Tweets is:{posCount}")
print(f"Number of Number of Tweets is:{negCount}")
print(f"{positiveScore}% of the tweets are positive")
print(f"The average confidence score is:{averageScore}")
sys.stdout.close()
```

At this end of the script, we calculated the average confidence score of all the tweets we harvested, calculated the percent that achieved a net positive score, and presented the remaining bits of data from what we were able to find in the group. We can then finally close the sys *tweets.txt* file so as to allow us to store it in the event of further research/checking of data collected.

Data:

Once we had the twitter data mining tool developed, we had to perform a bit of research with it leading up to our actual mining. We had to see which topics of conversation would typically produce negative sentiment from both sides of the debate so that we could minimize the risk of unintended sampling bias in our data. We also needed to find areas that matched the Karen archetypes fields of conversation which led us to the topics of 1) the vaccines for COVID-19 and mask wearing to minimize the spread 2) police and black lives matter and 3) presidents Trump and Biden.

To find likely twitter user candidates from these areas of conversation, we grabbed a sample of 24 female names alongside Karen from a list of the most popular female names in the 1970's to see how the data we produce for Karen compares to women with other names and to try to get as close as we can to the age group the meme indicates our Karen archetype would be in. We chose to seek women in order to minimize the influence of other lurking variables outside of the scope of this paper such as possible differences in what men tweet as opposed to females. We were unable to guarantee that these names selected would solely produce female users, but we took note in the selection process to seek names that have a high likelihood of being a woman's name such as Mary, Abigail, Barbara, etc. which are common names among women.

As we generated these tweets, we applied the Transformers sentiment analysis tool to help us show the net sentiment of users of a given name. This was an overall effective tool to use for the project, but would give us some errors. One such issue was that it would determine that a tweet possesses positive sentiment when the intent of the tweet was likely sarcastic and should have produced a negative sentiment determination for an accurate answer.

For each of the tweets that we checked with our program, we would check to see if the given tweet contained any of the keywords that we are looking for in relation to a given topic. Such examples we used was {"trump", "biden"} and {"mask", "vaccine"} to find tweets relating to the views of the users relating to the either of the two presidents and the measures taken with in regards to the COVID-19 pandemic (respectively).

In the generation of the tweets from our likely female groups, we would run into the issues of retweets which would often skew our data by producing the same tweet again and again and not giving us a proper insight to what the thoughts of the individual are. To eliminate the noise stemming from retweets, we chose to see if the tweet was a retweet and if so to not include it in our sample.

Experimental Results:

As stated previously, for the internet meme to hold true, we would have to find that people named “Karen” would disproportionately have tweets with negative sentiment, compared to other common female names. What we ended up finding happened to be, for the most part, the opposite.

The overall tweet statistics give an overhead view of the sentiment of the tweets from the female users. The filtered topics usually gave about a few hundred tweets out of 50,000 per name. Due to twitter rate limits, and overall time restraints, we had to limit the overall tweets down to 100 per username and 10,000 per name. The data collected, we believe, still provides enough tweets to analyze our hypothesis.

The way that the *Transformers* package works is that it gives a label to the tweet of “positive” or “negative”. It also provides a percentage based on how confident it is in the rating of the tweet. The overall tweets analyzed had an average confidence score of 95.514%. Being above 95.5%, overall is solid. Especially with 250,000 tweets analyzed, having such a high percentage gave us added trust in the reliability of our data.

The graphs that will be presented below, not just for the overall tweets, but also for the individual filtered topics, will show the percentage of tweets that are positive for each name. The black bar on each graph is used as a visual aid to help point out which bar is for Karen.

(Figure 1)

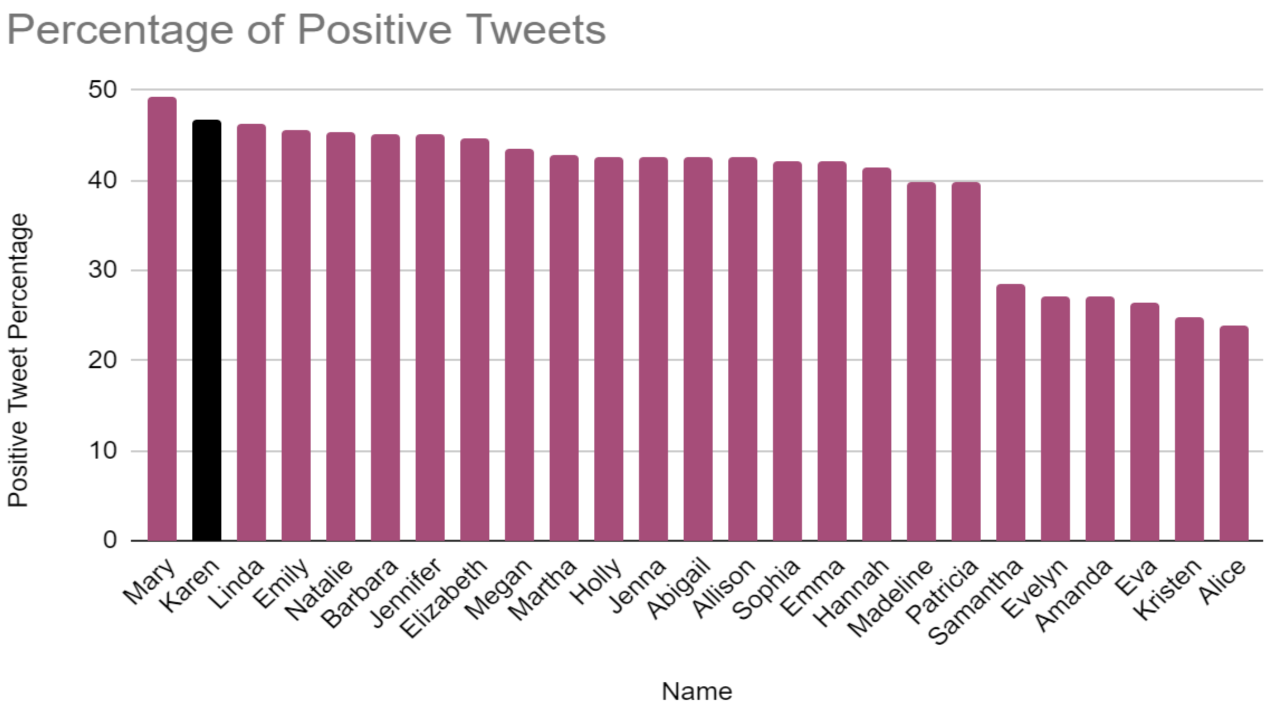

Figure 1 shows the statistics for the overall tweets. As you can see, the name "Karen" ranks 2nd out of the 25 listen names. 46.655% of tweets from users named "Karen" were labeled positive. The most positive name on average for overall tweets was "Mary" which came in at 49.220%. The name that produced the most negative tweets on average was "Alice" at 23.961%. That means that users named "Karen" were almost twice as positive as users named "Alice".

(Figure 2)

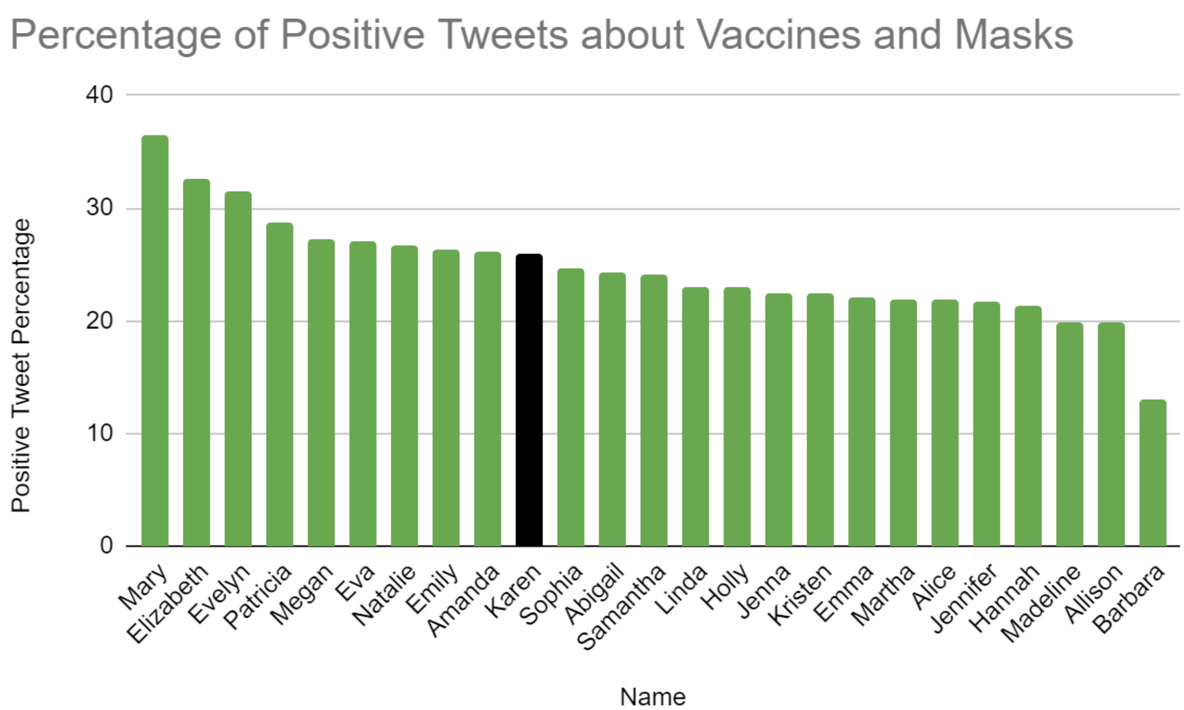

Figure 2 shows the statistics for tweets about vaccines and masks. The tweets were filtered using the keywords "vaccine" and "mask". For each username, the 500 most recent tweets were filtered through to find tweets containing those keywords. This means that for each name, 50,000 tweets were harvested and filtered through. After applying those filters, on average, there were 411 tweets for each name that related to the topic. The category overall ranked 2 out of 3 as the average positive tweet percentage was 24.57%.

As seen in Figure 2, the name "Karen" ranks 10th out of 25 in terms of positive tweet percentage. 26.309% of tweets from users named "Karen" about vaccines and masks were labeled positive. In this category, "Mary" ranked as the most positive name at a positive tweet percentage of 36.452%. The most negative tweets on average in this category came from users named "Barbara" at a rate of 13.043%. Interesting to note, "Mary" was the most positive name on average both overall, and when it came to vaccines and masks.

(Figure 3)

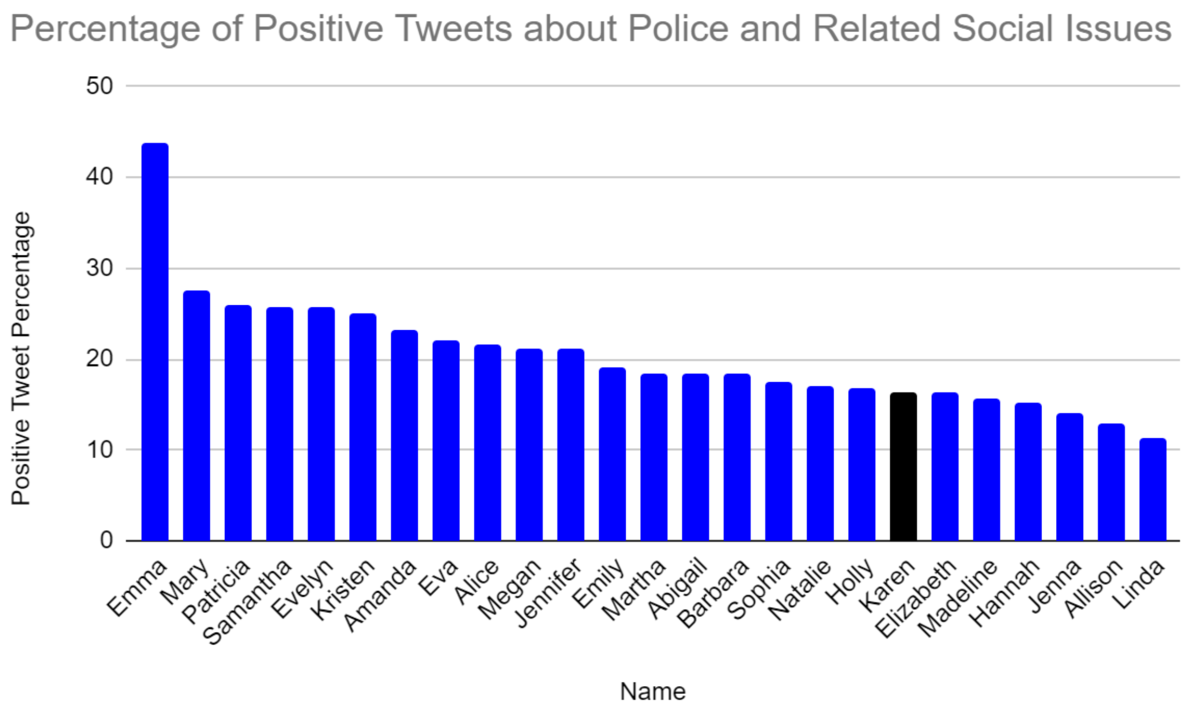

Figure 3 shows the statistics for tweets about the Police and Related Social Issues. There were many words that were used to filter through tweets. Keywords used were: “police”, “cops”, “cop “, “Lives Matter”, “BLM”, “ACAB” and “ALM”. For this topic, the average tweets gathered per name were 267. It’s worth noting that there is a space after “cop “ as this helped us eliminate some irrelevant tweets that contained words with the substring of “cop” within them. This topic was the toughest out of the three to find tweets for, hence the numerous keywords. We tried to use words that represented both sides of the debate around police and the lives matter movements.

The statistics in Figure 3 show that the name “Karen” ranks 19th out of the 25 names tested. 16.355% of tweets from users named “Karen” about Police and Related Social Issues were positive. This topic produced the most negative tweets for users named “Karen”. The most positive name on average for this topic was “Emma” who was high above all other users at 43.873%. The most negative name on average for this category was “Linda” at 11.268%. The keywords for this topic overall produced the most negative sentiment for tweets as the average positive tweet percentage was 20.42%.

(Figure 4)

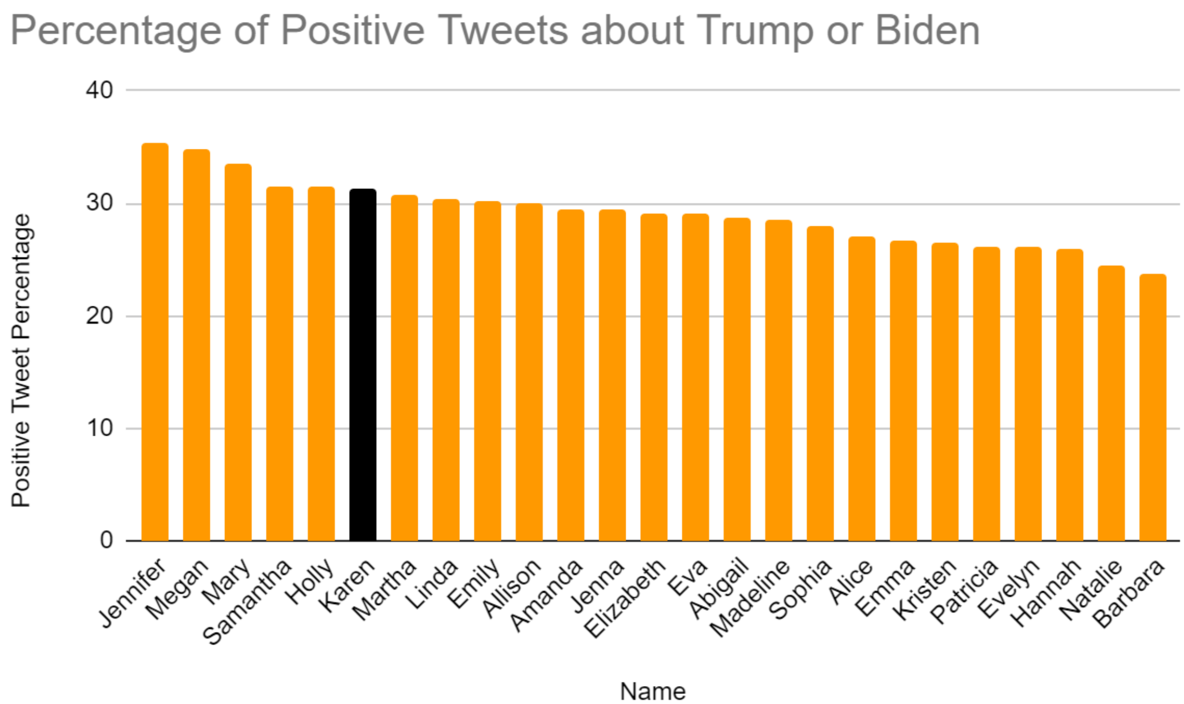

Figure 4 shows the statistics for tweets about Presidents Trump and Biden. The only keywords that were used for this category were "Trump" and "Biden". The keywords appeared frequently when people were tweeting at the twitter accounts of the President, or their family members. The average tweets gathered per name was 721, which was the most out of three topics that we filtered for. The sentiment for this category, overall, was the most positive, with a 29.136 positive tweet percentage.

As shown in Figure 4, the name "Karen" ranked 6th out of the 25 names tested. The positive tweet percentage for "Karen" users was 31.325% for tweets including either of the last two US Presidents' names. The name that was the most positive for this category was "Jennifer" at 35.29%. For comparison, the most negative name for the President category was "Barbara" at 23.81%. This almost 12% difference from least to most positive names was the closest gap out of all the categories we tested.

Analysis:

(Figure 1) The goal of the study is to show that twitter users that contain the name Karen would have tweets that suggest results in an overall negative sentiment. Ironically, out of the 24 other name's unfiltered tweets collected, Karen ranked 2nd in positive sentiment. This would be the first indication that a raw collection of all Karen's tweets does not have a correlation with the Karen Archetype compared to other names such as Alice in the experiment. However this data does not specify a common reason or attribute to why the users tweets are a negative sentiment. Such that the statements in the tweets can be irrelevant to what we define as a Karen archetype to be as well as the majority of the tweets collected from the users are retweets that contain **generalized information**.

@SecondLady45: "Suicide is preventable and #MoreThanEverBefore we must do all that we can to help vulnerable Americans, especially… https://t.co/Mf0eJ82JPr"

[{'label': 'NEGATIVE', 'score': 0.9768712520599365}]

A retweet timeline from a single user including duplications can take a majority of the limited tweets that the Twitter API can collect where a majority of them are irrelevant to our hypothesis. This is the reason why we had formed three relevant filter topics that have the most history in regards to the "Karen" topic.

(Figure 2) Our definition for a Karen Archetype in regards to this filter is that they are more likely to be anti-maskers with statements that result in a negative sentiment. Since this filter is more of a global topic it seems that tweets are collected from twitter users outside of the U.S. which goes against the experiment's domain. Unfortunately since many twitter users do not put their location publicly onto their account, adding a US filter location would significantly limit the amount of tweets collected. So, we then run into the issue of twitter users who tweet about mask policies and vaccine distributions from other countries.

@E_Nwabuoku
Location : Port Harcourt, Nigeria

@EvelynLabour
Location : Deptford, London

@samthesparrow
Location : Nottingham

@samanthabarry
Location : **Global**

@NightswithAlice
Location : **Everywhere**

@ViTheCleaner
Location : **www.vi-the-cleaner.co.uk**

The usernames above are just a couple of users that have publicly stated their location in their profile. Since twitter profile locations are not verified, users are free to put whatever text they want, which could lead to **internet trolls** forming errors in our data or inconsistencies with users outside of the United States.

@KristieMewie: " **"**Wearing a mask under your nose is like wearing a c****m over your b******k.**"** ~ @Nightwatch_Nick quote of the day 🤣🤣🤣"

[{'label': 'NEGATIVE', 'score': 0.9831653237342834}]

**Censored original tweet*

The sentiment analysis result from the tweet above alone is very hard to determine whether they are in support of masks, or antimask due to the sarcastic quotations.

(Figure 3) With this filter, we had to be more specific about the definition of the Karen Archetype. So we had made the assumption that they would have conservative values, such as going against the reform of the police, against Black Lives Matter (BLM), or in support of All Lives Matter (ALM). Another assumption was that twitter users that contained ALM or #ALM would indicate conservative values. Compared to the other filters, Karen had a higher negative sentiment average at 16.3% positive rating overall. Although Emma had a 43.8% positive rating and had 39 more tweets collected than Karen, it is not far off to disprove that a majority of the Karen tweets related to police were overall against police reform or BLM. But after manually reading the tweets that contained both negative and positive sentiments from Karen; There are indications that keywords and language used could not help the transformer package differentiate our definition of a Karen Archetype with its definition of a negative sentiment. For example,

@KarenSpilka: "The **police** reform bill signed into law on New Years Eve was an important start; it is the sprint portion of the marathon we must run, listening to the voices of people of color as we go. #mapoli 19/24"

[{'label': 'POSITIVE', 'score': 0.9840203523635864}]

Shows a positive sentiment by a Karen user that goes against our hypothesis. Other tweets we had viewed that contained All Lives Matter were actually liberals making sarcastic statements about conservatives.

@HollyMaeHenry: "if you're still using the argument "**all lives matter**" you look mentally incompetent, just ~ f y i ~"

[{'label': 'NEGATIVE', 'score': 0.9992997646331787}]

another indication that the filter and transformer package had flaws in the experiment. In conclusion, it was difficult for the analysis tools to prove our definition of the Karen Archetype hypothesis. As a lot of the tweets collected were from Liberals making sarcastic statements to Liberals talking about police brutality in the U.S. which the transformer analysis tool would result in a negative sentiment. Conservatives with negative sentiments would mix with the data resulting in a potential source of errors.

(Figure 4) Since the only filter keywords we had used were Biden and Trump, it had collected the most tweets compared to other filters. Based on the tweets read, they ranged from the time of the start of the election to the storming of the capitol until the inauguration of president-elect Biden. Similar to the Police filter, we had assumed that the Karen Archetype would support conserviatve policies in addition to statements about a rigged election that should result in negative sentiments. The name Karen had a high positive percentage rating at 31.3% , 6 out of the 25 names. This filter would disprove the correlation between the name Karen and the hypothesis. Issues come to surface since the transformers analysis tool does not take into account the political views of the user. After reviewing the tweets regardless of the twitter users political views it would rate it either positive or negative. This would produce an error in our research, because now it is harder to differentiate the correlation for a Karen's political views with the actions of the archetype based on the transformers package and the filter. One way would be to assume that negative tweets that contained Trump would be Liberals and negative tweets that contained Biden would be conservatives,

@OnAirHannah: "**Trump** wins a state: Victory! I am King! **Trump** loses a state: RECOUNT! FRAUD! Can't even with this dude."

[{'label': 'NEGATIVE', 'score': 0.9990556240081787}]

The example above is a special case where the statement that contains Trump is sarcastic, but the negative sentiment still indicates that the user is likely a democrat.

@lifeaseva: "Let me get this straight, **Trumpies LOVE** to bring up "looters and rioters" in regards to BLM, who we all know were clearly opportunists with no stance just trying to get free shit in chaos... but are silent when their own kind forms a TERRORIST group to kill a US GOVERNOR... i…"

[{'label': 'POSITIVE', 'score': 0.9254190921783447}]

Another special sentiment analysis result shows the use of sarcastic keywords that also contain Trump from a democrat but results in a positive score, most likely due to the keyword 'LOVE'. These errors also indicate the lack of variety and perspective in the twitter platform. After reading many positive and negative sentiment tweets from all users in this filter, it seems that a majority of tweets are from democratic supporters. This phenomenon can also be an indication from the timeline when Trump was officially banned from twitter, and had influenced a majority of conservatives and Trump supporters to transition into the recently deployed social platform, Parler, where they believe they can express their freedom of speech.

Evaluation of Potential Sources of Error:

According to Pew Research, the top 10% of twitter users tweet about 80% of all tweets. Therefore, the sample size that we have isn't as diverse and we mostly have the same users tweeting about the same sentiment for each topic. In addition, Twitter's demographics consist of mostly liberal younger women below age 30 which do not match with the demographics of the Karen archetype in that they are likely over the age of 40 and is conservative.

When applying the Transformers sentiment analysis tool, we would run into issues that are regular occurrences with any working sentiment analysis tool. One of the most common errors that we ran into from the Transformers sentiment analysis tool was the inability to detect sarcasm which would produce a false positive when the intent of the message was actually meant to be negative. This would also be the case the other way around where a tweet may return as negative, but intended to be positive (for example a friend saying "I hate you haha" when someone announces online that they were just accepted into a prestigious university). Furthermore, there was a gray area with certain topics that had it lean towards one way instead of the other. For example, there was a tweet that wrote, "Abolish the police" and was categorized as positive sentiment by the Transformers Tool even though the tweet clearly had negative sentiment toward the police. This tweet also shows no detection of sarcasm.

For future data mining experiments studying the Karen Archetype, we would look into more accurate sentiment analysis tools that may have come along. Ideally we would be able to apply more accurate sentiment analysis tools to take advantage of that may have been used before in similar studies into the analysis of the Karen archetype.

The occurrence of spam/meme accounts on twitter would generate a lot of dirty data for the project to deal with. A spam account may have been made just for a singular joke towards another post on twitter and be picked up by our program unsuspectingly grabbing it and as a result skewing our experimental results. The occurrence of meme accounts were especially difficult for us to deal with in this project as the Karen meme remains popular with many users deciding to pretend to be the infamous Karen from the meme. This would produce negative results from a subset of our dataset that should not be included if before being added they were known to be part of a meme page.

For future data mining experiments studying the Karen Archetype, we would want to look at gathering a large dataset of regular user twitter accounts and spam/meme accounts to apply supervised machine learning techniques to help us build another filter for us to work with. This would allow us to further reduce the excess noise of twitter data that is just dirtying on our results. This would likely not remove 100% of our issues relating to spam/meme pages affecting our results, but it would allow us to generate a more accurate sense of the relationship of a name to one's frequency of producing negative content.

Location services with a tweet were a major source of error that we would run into as the GPS coordinates would be infrequently turned on (as a result of not being needed and invasion of privacy concerns). This would have then forced us to rely more heavily on the voluntary input of data for the location of the origin of a tweet resulting in searching for given keywords within the location which relies heavily on the trust that a user produces accurate information describing the actual location of the origin of the tweet.

For future data mining experiments studying the Karen Archetype, we might look into seeing if there is an API we can take advantage of in another social media platform such as Instagram. An image driven social media platform such as instagram would be ideal as geo-tagging features are included in all digital pictures that we take, so it would be possible to trace where the picture was taken or where the post was sent from. With this guarantee for the location, we would also be able to take advantage of the message associated with the image in the post and determine the sentiment from there. This would give us a much more reliable form of testing in relation to location than just by the trust of the twitter user where most people with no technological background could easily turn on or off the location settings associated with their tweet. As a result, we were unable to pinpoint certain locations to find more Karen archetypes.

The common joke in the Karen meme is the phrase “I would like to speak to the manager”. We did try to see if we can find tweets in regards to complaints about a restaurant’s/store’s services, but upon our lack of tweets relating to this trait of the Karen archetype we decided to discontinue this part of our search as we would otherwise have to draw conclusions from an extremely limited dataset. We believe the lack of tweets relating to complaints about a restaurant’s/store’s services is from the social context behind the complaint itself would mostly remain offline and not show up in a tweet.

For future data mining experiments studying the Karen Archetype, we would have to find some other way around this. We are unable to control social behavior to make the Karen archetype post “I would like to speak to the manager” every time she has an incident which was a major contributing factor to how the archetype was manifested. It was more common in recorded videos to see a Karen asking to speak to the manager than written posts and tweets. We would like to take more creative approaches in trying to find people who discuss witnessing such a moment online, but as of the moment we are currently limited by the technology of our time.

Some users in either an effort to protect their identities or from confusion in the setup of their accounts would duplicate their username in the field for their actual name. This would as a result make us miss a lot of potential candidates as they may possess one of the names we were testing for, but from using a creative username such as *@bubblegum2718* as the first name, we would be unable to find and record their tweets for use.

In addition, some users will go by their nickname instead of their given name. For example, in our experiment, the name Elizabeth could go by nicknames like Lizzy, Liz, Beth, and more. Therefore, we were not able to grab those users who went by those names to guarantee that their full name is Elizabeth and were unable to use their tweets. Our sample size would be smaller as a result of that.

For future data mining experiments studying the Karen Archetype, we might want to look into what the accounts that follow her have tagged her in. From there we can take the (admittedly computationally expensive approach) of searching through the tagged posts and seeing if there are occurrences of deviations of the word “birthday” (deviations may include “bday”, “b-day”, “birtday” (misspellings), etc.) and see what name most frequently occurs. This would likely produce for us a name to work with, but is mostly restricted by the computational power presently available for the project team and the budget needed in order to gain access to Twitter’s premium API to grab the information in a timely manner.

Another source of error is the occurrence of retweets which would oftentimes lead to the same tweet being formed multiple times across multiple different users. This would lead to issues in determining the true mindset of the users as retweets would be reflective of the mindset of the originator of the tweet and not necessarily give the greatest insight to the mindset of the individuals we are testing for. To combat this, we filtered out the retweets from our dataset to try to get the most genuine input from the individuals as opposed to analyzing the same individual’s mindset multiple times over.

However, some people make an attempt to give the illusion of originality in their tweets by copying and pasting a popular tweet and try to pawn it off on their own work instead of showing the original writer by retweeting. For future data mining experiments studying the Karen Archetype, we would like to build a small database for us to store the tweets that we have found thus far in our search and try to see if a given post can be found within our database as already found. If it surpasses a certain threshold (let’s say 95% or above of the candidate post is the same as a previously mined tweet), then we would reject the candidate post under the suspicion that it is a plagiarized form of another tweet. This method would be prone to some innocent occurrences of type one errors where we reject the new data coming in when we shouldn’t, but until sufficient plagiarism technology is available for use through Python’s Twitter API, we don’t appear to have many alternatives to address duplication.

Conclusion:

According to our results, the name Karen does not have a direct correlation to the Karen archetype of having negative sentiment toward the vaccine and masks, the police, and relating to Presidents Trump and Biden. Interestingly, people with the name Karen tended to have more positive sentiment and/or equal sentiment in comparison to the average of the 25 other female names that we experimented with. Although we were able to grab significantly positive or negative sentiment for our project, our program fell short of catching all of the sources of error. In retrospect, we applied our approach and worked together on developing an analysis to an odd, yet interesting question.

Resources:

- https://www.theguardian.com/lifeandstyle/2020/may/13/karen-meme-what-does-it-mean
- https://www.theatlantic.com/international/archive/2020/08/karen-meme-coronavirus/615355/
- https://abcnews.go.com/2020/top-20-whitest-blackest-names/story?id=2470131
- https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/
- Twitter's Python API