Monkeypox Analysis in the United States

by

Taeyonn Reynolds

A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Mathematics

Kenneth Jones, Ph.D. Committee Chair	Julian Allagan, Ph.D. Committee Member
Mohammed Talukder, Ph.D. Committee Member	Jeffery Ingram Committee Member

Department of Mathematics, Computer Science and Engineering Technology
Elizabeth City State University
November 2022

ABSTRACT OF THESIS

Monkeypox Analysis in the United States

Throughout the course of time technology has evolved and led to mass communication throughout the world. One of the biggest types of communication in today's society is social media. Furthermore, it can be defined as a virtual community that allows users to create and share information rapidly. One of the biggest social media platforms today is Reddit, which is a network that allows users to communicate in various subreddits that can be divided into cities. By analyzing the data from a given subreddit it will give a clear description on how citizens are reacting to a given situation that is taking place in the community. Due to the recent events that have taken place in today's society this paper is on the monkeypox outbreak. The data corresponding to the viral disease will be scraped from given states to provide an outlook from different communities. To examine the data there will be data analysis, topic modeling, Natural Language Processing(NLP) techniques and word embedding.

Taeyonn Reynolds
Elizabeth City State University
October 2022

ACKNOWLEDGMENTS

I would like to acknowledge and express my gratitude for Dr.Kenneth Jones for assisting me during this academic process. Dr. Kenneth L Jones played a pivotal role in my life by mentoring and encouraging me throughout the years. To Dr. Julian Allagan and Dr. Mohammed Talukder, thank you for teaching me valuable skills and giving me guidance at my time at Elizabeth City State University. You all have impacted me in many ways with a overwhelming attitude to assist me from classwork to projects outside of the classroom. I would also like to thank Jeffery Ingram for serving as a committee member.

Contents

List of Figures

1 Introduction

In the year 2020, the modern world discovered the impact a pandemic can cause throughout all walks of life. Two years later monkeypox has the potential to provide another change in our communities. By using Reddit and various techniques there will be a clear understanding on how the people are expressing themselves over the concern at hand. This was completed by using Natural Language Processing(NLP), a branch of Artificial Intelligence (AI), which is the capability of a program to understand how humans interact with one another. Also a key component to the project was using Reddit Application Programming Interface(API), an API allows two computer programs to share data and execute tasks together, that correlates with Python Reddit API Wrapper(PRAW). A wrapper is used to allow one API to communicate with one another API without changing it.

1.1 Monkeypox Timeline

The Moneykpox outbreak has caught the world by surprise as it comes two years after the COVID-19 pandemic. Monkeypox can be expressed as a rare disease that is caused by infection with the monkeypox virus[2]. It is also in the same family of viruses as smallpox and does not have any relation to chickenpox. Monkeypox was discovered in 1958 when two outbreaks of a pox-like disease occurred in colonies of monkeys kept for research[2]. The origin of the virus was never documented even though it was given the name

monekypox. Furthermore, in 1970 the first human case of monkeypox was recorded in the Demonic Republic of Congo. Cases began to rise a decade later as it started to develop immunity to the smallpox vaccine.

In 2003 there was a monkeypox outbreak in the United States that led to forty seven confirmed cases and few more probable cases. This was a key moment in history as this was the first time humans have contracted monkeypox in a place other than its origin. During the month of April that year, there was a large shipment of roughly 800 small animals of nine different species[13]. Among those animals were the rodents rope squirrels, African giant pouched rats, brush-tailed porcupines, dormice and striped mics. This outbreak occurred due to interactions with pet Pierre dogs that were residing beside animals from Ghana when they were imported into the United States. When the dogs were brought to the United States an animal distributor in Texas began to sell them to various customers not knowing the events that have transpired.

In doing so, the dogs had found new homes and eventually let to the spread of monkeypox to Illinois, Indiana, Kansas, Missouri, Ohio, and Wisconsin. The outbreak was initially discovered when a young child contracted the virus and became ill due and developed a rash soon after. The virus began to spread from person to person and raised concerns on the initial transmission of the virus. Through an investigation and testing it was stated that twelve of the rodents from Ghana had contained monkeypox. To combat the spread of the virus many government agencies utilized their skills to-

gether by extensive laboratory testing; deployment of smallpox vaccine and treatments; development of guidance for patients, healthcare providers, veterinarians, and other animal handlers; tracking potentially infected animals; and investigation into possible human cases[13]. This led to the temporary ban of sale of prairie dogs and an embargo on importation of rodents from Africa.

During September of 2017, a race occurrence happened as this was the first time in nearly forty years of a reported case of monkeypox in Nigeria. The Nigeria Centre for Disease Control(NCDC), a federal government agency utilized to combat vaccine curable diseases, was alerted of a potential case of a monkeypox infection. There was a total of 172 suspected cases and sixty one confirmed cases have been reported through the year. Also, Laboratory-confirmed cases were reported from fourteen states (out of 36 states)/territory: Akwa Ibom, Abia, Bayelsa, Benue, Cross River, Delta, Edo, Ekiti, Enugu, Lagos, Imo, Nasarawa, Federal Capital Territory (FCT) and Rivers[9]. Suspected cases were reported from 23 states/territories including: Abia, Adamawa, Akwa Ibom, Bayelsa, Benue, Cross River, Delta, Edo, Ekiti, Enugu, Federal Capital Territory (FCT), Imo, Kaduna, Kano, Katsina, Kogi, Kwara, Lagos, Ondo, Oyo, Nasarawa, Niger, and Rivers[9]. At the time, nearly seventy five percent of all cases were male with the average age of thirty years old and one death was reported. The NCDC developed a rapid response team over the nation to combat the spread through contact tracing and investigation of the situation through surveillance. Another

attempt to slow the spread for the public health was the creation of a task force named The Surveillance and Outbreak Response Management System (SORMAS). It was utilized to slow the spread by providing real time solutions and communicating with the public in November of 2017. For example, one of the various techniques utilized was isolating units and medical care workers were advised on the proper protocol to minimize the spread.

The following year there were cases of monkeypox reported in the United Kingdom during the month of September. It was the first occurrence of the virus throughout the history of the country. This was made possible due to an individual from Nigeria who was traveling to a naval base in Cornwall which is an historic county in the South Western portion of England. The patient was later transferred to the expert infectious disease unit at the Royal Free Hospital, London where they were able to receive the appropriate care. The Public Health England(PHE), agency tasked with the department of health in England, was able to work with the hospital as they have an specialized disease unit. Through contact tracing there were able to identify potential citizens who may have contracted the virus as a precaution. A week later, another individual became diagnosed and through investigation it was determined there was no connection from the other incident. This event took place in Blackpool which is a tourist attraction in the northwest coast of England. Soon after a medical worker who was tasked to the situation in Blackpool contracted the virus.

In May of 2019, a man was admitted to a isolation ward in Singapore

once lab results stated the man tested positive for monkeypox. Nearly twenty two people who were in close contact of the individual were quarantined to stop the spread. Of the people quarantined it consisted of eighteen members of the workshop and a few personal from the hotel/venue. Medical staff and government officials successfully identified the guest of the workshop as some commuted home to their respective communities. The older male was in the country to participate in a workshop and attended a weeding in Nigeria a week before. A few symptoms that were experienced were fever and chills as well as a rash that began to spread to multiple parts of the body. Furthermore, he was isolated in his hotel room until he was admitted to a hospital days after. He was later able to receive medical attention from the staff due to effective personal protective equipment.

The Centers for Disease Control and Prevention (CDC) and the United States of America confirmed a case of monkeypox from one its citizens that returned from travel in July of 2021. Local health officials communicated with authorities to monitor the spread by gathering information on the other passengers who were also on commercial flights. As the year 2022 approached the middle of May the first case of monkey pox was documented in United Kingdom. In the following weeks there were cases in the surrounding countries and eventually made its way to the western hemisphere through Canada and the United States of America. In fact, The World Health Organization(WHO) stated there were ninety two confirmed cases from various countries such as Spain, Sweeden, Italy, Germany, Australia, Belgium and France at the

time. Shortly after WHO declared the monkepox virus a escalating global monkeypox outbreak a public health emergency of international concern[9].

In the following week, many unions began conversations to administer the vaccination for moneypox. The Department of Health and Human Services (HHS) has been distributing JYNNEOS vaccine from the Strategic National Stockpile since May 2022[6]. In fact, The United States government ordered 2.5 million JYNNEOS vaccines on July 15, 2022. On August 9, 2022, the U.S. Food and Drug Administration issued an Emergency Use Authorization allowing healthcare providers to administer a smaller dose of JYNNEOS into the skin layers of the forearm (like a tuberculosis skin test) as the preferred option to the standard dose usually given in the upper arm[6]. Furthermore, WHO declared monkeypox outbreak a global health emergency which was only used two times over the course of history for the the COVID 19 pandemic and Polio.

There two vaccines that were approved by the Food and Drug Administration to prevent monkeypox was JYNNEAOS and ACAM2000. JYNNEAOS was manufactured to protect people from smallpox and monkeypox. In addition, ACAM2000 was originally developed for smallpox. JYNNEAOS is the primary vaccine for the the United States of America and they are currently working on the vaccine to determine the effectiveness.

There is also an alternate route to combat monkeypox which is an antiviral drug that is called TPOXX (tecovirimat). TPOXX was initially approved by the Federal Drug Administration(FDA) in December of 2018 to properly

treat smallpox for all citizens. Even though it is effective for smallpox, the FDA did not approve it for monkeypox. Recently, TPOXX use for treatment of monkeypox is authorized under FDA regulations, and CDC holds an expanded access Investigational New Drug (EA-IND) protocol. As of today, there were 3,233 patients prescribed or treated with TPOXX and this number underestimates the number of patients who are receiving TPOXX treatment as healthcare providers can start treatment before submitting IND paperwork to CDC[20]. The drug has been proven to be effective in animal trials as it has dropped the death rate when used early in the initial process. This option for treatment is suggested for people who are experiencing/have human immunodeficiency virus (HIV), leukemia, eczema, burns, impetigo and pregnant or breastfeeding people. When considering the use of tecovirimat, clinicians and patients should understand the lack of tecovirimat effectiveness data to date in people with monkeypox, the lack of data indicating which patients might benefit the most from tecovirimat and the concern for the development of resistance to tecovirimat, which could render the drug ineffective for any treated patients[20]. There are more serious symptoms that are high occurring if the use of intravenous therapy rather than prescribed pills.

1.2 Background of Monkeypox

Majority of monkeypox cases arise in Central and West Africa but are transmitted by imported animals and international travel. The virus is

mostly deadly in Africa as one out of every ten people are pronounced dead when they contract moneypox. People with monkeypox get a rash that may be located on or near the genitals (penis, testicles, labia, and vagina) or anus (butthole) and could be on other areas like the hands, feet, chest, face, or mouth[2]. The rash will see many changes while healing that can cause sores and become itchy.

Also, there are also other symptoms that can have an effect on the host such a headache, chills, fever, swollen nymph nodes, exhaustion and respiratory symptoms(e.g., sore throat, nasal congestion, or cough). There is no prediction on what symptoms will occur or how severe they will become. The symptoms will mostly likely appear within three weeks of contracting monkeypox and experiencing flu like symptoms will cause for the rash to appear within one to four days. It can still be transmitted until the rash if fully removed from the body. Also, after two to four weeks of the virus the pateint should be fully recovered.

The monkeypox virus can be spread through close contact with an infected person or animal. In fact, it can spread through contact with a persons rash, scabs or bodily fluids. Having extended contact or physical contact by kissing, cuddling or sexual intercourse will result in the transmission of the virus. As of right now, the data states that most cases are from homosexual or bisexual man but anybody can contract the virus through close contact. It can also be spread through fabric if the infected host bodily fluids were on a blanket or clothing. In addition, a pregnant woman can transmit the virus

to her fetus.

1.3 Monkeypox vs Smallpox

The smallpox virus, also known as variola virus, emerged and began causing illness and deaths in human populations, with smallpox outbreaks occurring from time to time[4]. The last case of smallpox in the USA was 1940 and it was classified as eradicate in 1980 by the WHO. This was a significant event in history as it was the first infectious disease to be classified as such. Scientists believe the only way smallpox can return if its released through bio terrorism. During its tenure it would initially cause a fever, mouth sores and vomiting. Muscle plain, headaches, fatigue, nausea and backaches also were common results of the virus. As it began to move its course through ones body blisters where formed that left scares on the skin. Also, one of the most severe symptoms was the loss of vision.

Smallpox would be transmitted between direct contact or contact with an surface that touched by an contaminated person. At the peak of smallpox the fatality rate was fairly high as every three out of ten people would die if they contracted the virus. In fact, there was a total of nearly three hundred million deaths due to the virus in the 20th century alone. The spread of smallpox was successfully slowed down due to the smallpox vaccine. Monekypox and smallpox are similar due to them being part of the same group of viruses. Smallpox is considered more severe and contagious than monkeypox but the symptoms are closely related. Also, monkeypox causes large nymph

nodes which help determines the diagnosis. The similarities of the two virus allows for the same medication to be used for the initial treatment.

1.4 Purpose of this Study

For this study, the main area of focus will be the United States of America and how monkeypox has impacted communities across the nation. There has been 64,290 confirmed cases and 24,364 of those cases are from the USA as of September 22, 2022. That is nearly forty percent of all confirmed cases throughout the world. In addition to, 684,980 doses of the vaccination has been administered to citizens since May. This project was intended to analyze the number of confirmed cases through age, gender, race/ethnicity and symptoms to provide a clear understanding on how the population is responding to the virus by word embedding and cosine similarity and topic modeling. The main research questions for this project will be:

1. Does race and gender play a significant role in contracting the monekypox infection?

2. What is the current severity of monkeypox and how fast it can be spread?

3. Is there a particular age group that contributes the most cases in the United States of America?

4. How is the country reacting to the current outbreak?

2 Literature Review

To achieve a understanding on the study there has been many examinations of informative relevant works and statistical data. By doing so, new insights were available on the project and led to a clear analysis on previous works that contributed to research and analysis of global statistics that provided research that is applicable to the study.

2.1 Global, National, and State Statistics

The world is constantly monitoring the severity of the monkeypox virus by initially investigating the measurements in regards to contaminations and deaths because of the illness. Global cases have amassed a number of 65,415 to the date September 25, 2022. In terms of symptoms that occur from contraction with the virus death is not a usually occurrence as there are twenty six deaths worldwide and that is less than one percent. The most deaths from the virus come from countries that have historically reported monkeypox over history. As though death is not a common event the body can be impacted in many other ways. The United States alone has 24,846 cases which account for nearly 37.983 percent of cases total. In fact, it was found that most cases are coming from Europe and South America as they contain the top ten countries associated with monkey pox cases if you include the United States. This can be understandable as these are the countries who first saw transmission of the virus. Brazil and Spain both account for nearly

over 14,000 cases and five deaths.

A vast majority of the overall infections are from the most populated states across the country. That can be revealed by eight out of top ten are also in the top ten states with the most confirmed cases to date. The top four states in regards of population are California, Texas, Florida and New York and they account for 13,397 cases overall. There has been 684,980 vaccination doses administered for monkeypox in the United States over the course of four months dating back to June 19, 2022. Also, that number is a collective sum of the first and second dose. Vaccines has been shipped worldwide to countries and made available to citizens in need. To be fully vaccinated a person should take both dose four weeks apart according to the FDA. There are 152,208 people fully vaccinated by the JYNNEAOS vaccine. A large majority of vaccines have been administered to men and that accounts to nearly half a million for the gender.

Communities all over the world has their own protocol for monkeypox but some practices are universal. There is still scientific evidence needed to know if JYNNEOS will fully protect against monkeypox virus infection in this outbreak, so infections may occur even if you are vaccinated[8]. To combat the spread of the virus it is essential to avoid close contact with the rash that develops with the monkeypox virus. If the development of symptoms occur it is best to stay home and contact a local health provider. Also, wearing a mask in situations where there may be a chance to be around high levels of traffic will reduce the chances of contracting the illness. The use of soap and

alcohol based hand sanitizer should be used regularly throughout the day. Health officials stated there should not be any skin to skin contact and that includes but not limited to cuddling, hugging or kissing. In addition, it is very important to not touch any items that have became in contact with a person who has had monkeypox. A individual can become infected if they do not wear the proper equipment when changing the bed sheets, cleaning essentials and clothing.

2.2 Relevant Studies

Most confirmed cases of the virus are from ninety nine countries who have not had a history of a reported monkeypox case. There have been many projects devoted to the findings of monkeypox and how other health issues can contribute to the severity. Those who are healthy may only need a small amount of medical attention to determine the severity . However, because prognosis depends on multiple factors, such as initial health status, concurrent illnesses, previous vaccination history, and commodities, supportive care and pain control may not be enough for some patients (for example, those with weakened immune systems)[6]. A few components that were used to learn more about the situation was the location, age, sexual preference, race/ethnicity and previous health concerns of a person. Some people are considered high risk if they have had anonymous or multiple sexual partners within two weeks, sex workers of any sexual orientation or gender, staff at establishments where sexual activity occurs, people who are living with

HIV/AIDS, people who have been diagnosed with any sexually transmitted infection in the past three months[8]. Having contracted monkeypox and another illness can provide negative effects to ones body and can ultimately lead to death due to the changes to the body.

The research points to homosexual activity between men display the most cases in the United States. An estimated 19 million Americans (8.2 percent) report that they have engaged in same-sex sexual behavior and nearly 25.6 million Americans (11 percent) acknowledge at least some same-sex sexual attraction[5]. A study was conducted to analyze how men were attempting to slow the spread and decrease the risk of harming the public health. Overall, forty eight of respondents reported reducing their number of sex partners, fifty percent reported reducing one-time sexual encounters, and 50 percent reported reducing sex with partners met on dating apps or at sex venues since learning about the monkeypox outbreak[17]. Evidence has shown that urban districts had a much higher vaccine rate, approximately twenty eight percent, than suburban and other areas. According to the CDC, among 4,460 cases with known sexual orientation and gender, 4,159 (93.3 percent) were gay or bisexual men, 242 (5.4 percent) were straight or heterosexual men, and 59 (1.3 percent) were straight, lesbian, or bisexual women[17].

2.3 Statistical Analysis

Through the use of statistical analysis monkeypox has been examined to see the affect it has on the population and how it can impact the world in the future. The data was retrieved from the Center for Disease Control as it provides accurate updates every week from hospitals around the world by verifying entries. The information started being recorded on May 17, 2022 as this was the beginning of the monkeypox outbreak in the United States. The data has proven that the global monkeypox outbreak is currently primarily affecting gay, bisexual, and other men who have sex with men[1]. By using race, ethnicity, and geography this study will provide a clear understanding of the ongoing situation that is effecting the country. There are many different practices that are being constantly debated on the various proper treatment for patients. Also, the effectiveness accurate laboratory tests, government involvement and level of necessary personal protection are also many topics at hand. There were 106 countries with reported monkeypox cases and their location ranged from places from all over the globe.

An attempt to further examine the spread of the current outbreak there was a scientific investigation conducted on a sample size of patients that contracted the virus. Recent sexual history can be defined as any sex and/or close intimate contact in the three weeks preceding symptom onset and among 7,378 people with data on both recent sexual history and gender, 78.9 percent reported man-to-man sexual contact (269). Data was provided throughout multiple jurisdictions across the country as some provided addi-

tional information and to provide an accurate analysis. Among 4,460 cases with known sexual orientation and gender, 4,159 (93.3 percent) were gay or bisexual men, 242 (5.4 percent) were straight or heterosexual men, and 59 (1.3 percent) were straight, lesbian, or bisexual women (269).

There were a wide range of events that came with contraction of monkeypox and some symptoms are highly likely to be experienced together. Among cases for which symptom data were available, nearly all reported a rash that occurred in 97.5 percent of all victims. This is not surprising as the initial diagnosis of monkeypox is the creation of a rash that may spread to multiple places on the body. The next few symptoms that were high occurring all ranged fairly close to one another as they were not life threatening but took a toll on the host by fever (66.4 percent), malaise (63.7 percent) and chills (61.4 percent). Additionally, people can experience headaches, enlarged lymph nodes, muscle aches, itching, rectal pain, rectal bleeding, blood in stools, and tenesmus. Some experienced on vomiting or nausea, abdominal pain, proctitis and conjunctivitis.

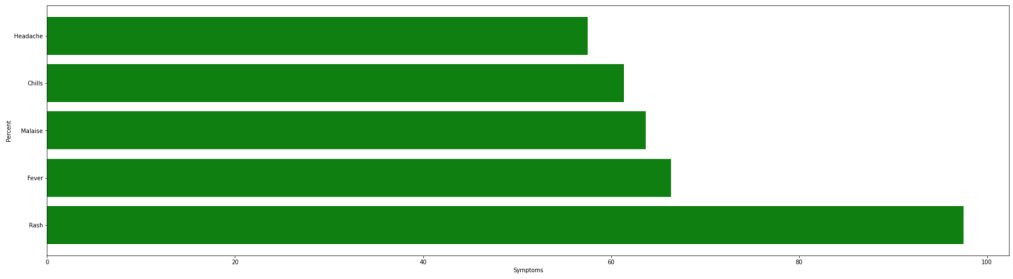

Figure 1: Symptoms of Monkeypox

The age group with the most confirmed cases of monkeypox was the 31 to 40 year old adults. The median age for all ages is 35 years old and the highest age reported is 78 years of age. As shown in Figure 2, this particular age group experienced significantly higher infections by totaling a number of 7,139 which accounts for 41.6 percent. The next highest was ages 21 to 30 (5032 cases, 28.9 percent), followed by ages 41 to 50 (3272 cases, 18.8 percent) and then ages 51 to 60 (1352 cases, 7.7 percent). The age groups with the least amount of cases was children under ten and the next lowest was 71 and older individuals. As the most general gender to appear in the data was man and they were responsible for 16,726 of total infections with 96.2 percent. The next highest gender was woman as they accounted for 365 cases and there was a few data entries that were not connected to a specific gender(139 cases). Transgender woman and men both accounted for 121 and 36 for their classifications.

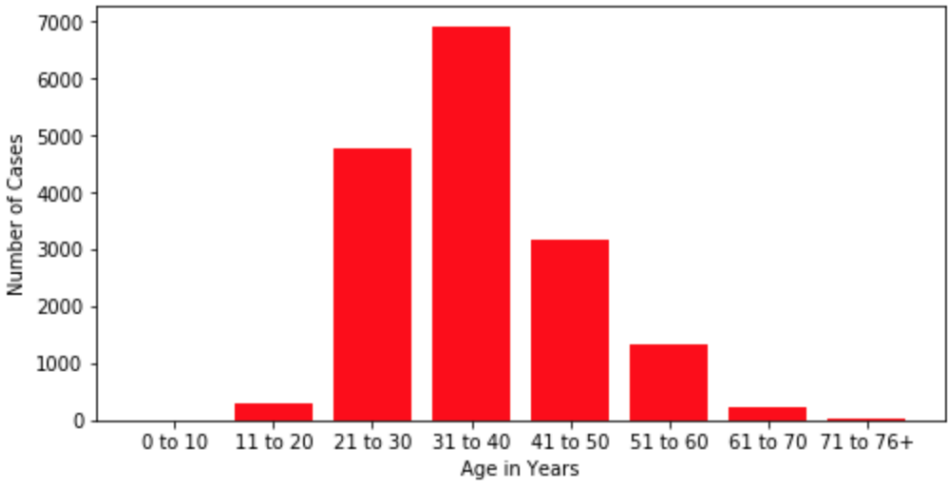

Figure 2: Total monkeypox cases for each age group

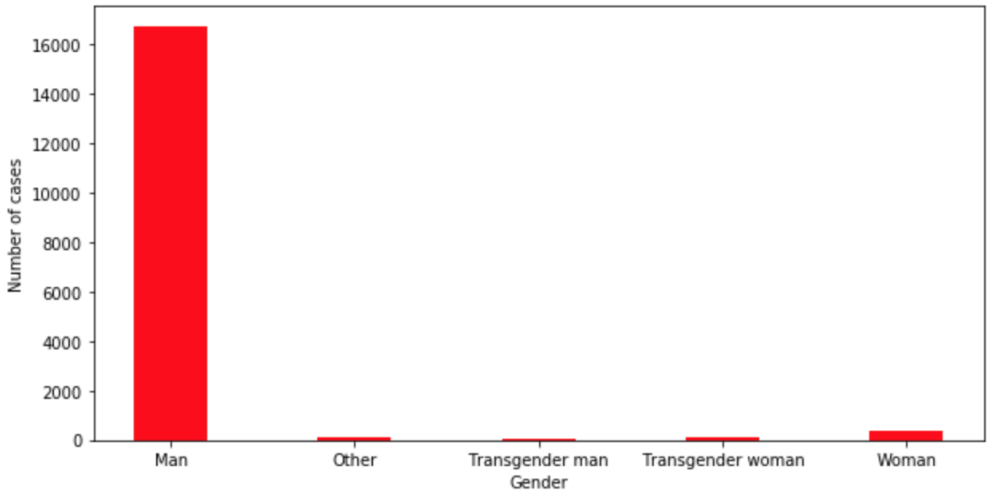

Figure 3: Total monkeypox cases for each gender

Over the past few months one of the main attempts to stop the spread was through a series of vaccines. There were 684,980 vaccine Doses Administered in the 48 U.S. Jurisdictions reporting data as of September 20, 2022.

Of those doses administered 479,391 were of the male sex, 39,558 of the infected were female and 8,869 were classified as unknown. Furthermore, to be considered fully vaccinated a person would have to take two shots at least two weeks apart. There are 155,205 people that are currently considered as fully vaccinated. By analyzing the data the most frequent age group to be proactive in getting vaccinated was the ages 25 through 39 as they lead all others by a large margin. There was a total of 252,422 doses administered and the next largest age group is 50 to 64 years old (110,344) followed by 40 through 49 (96,422). The two lowest groupings were 65 and older(28,382) and 17 and younger(639).

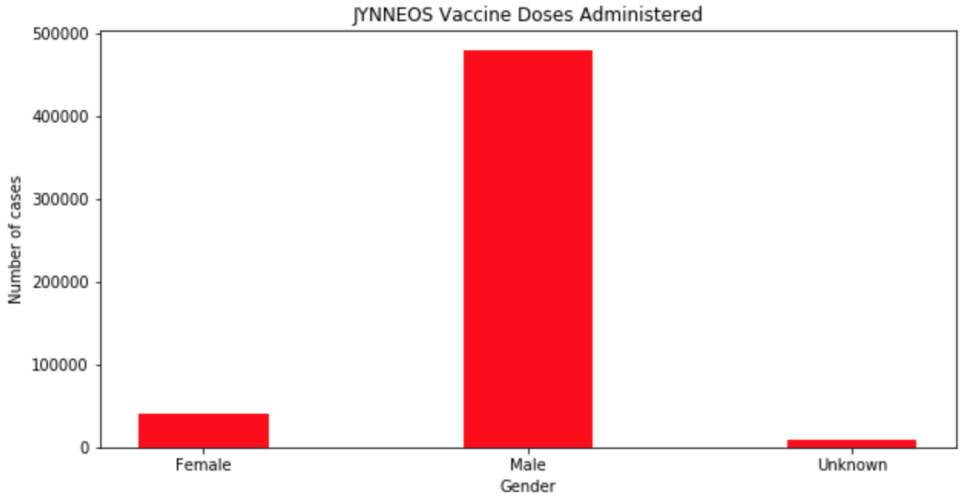

Figure 4: Total JYNNEOS vaccines doses administered by gender

The data set that displayed the given race/ethnicity for vaccinated citizens was mostly composed of White, non-Hispanics. This specific group has counted for 248,140 vaccination doses and 47 percent of all doses recorded.

The next highest grouping were Hispanic and they were the only other race/ethnicity to record over 100,000 doses (109,764). Furthermore, Black/Non-Hispanic was third(59,853), followed by Asian/Non-Hispanic (37,819) with multiple and other race/ethnicity rounding up the remainder doses.

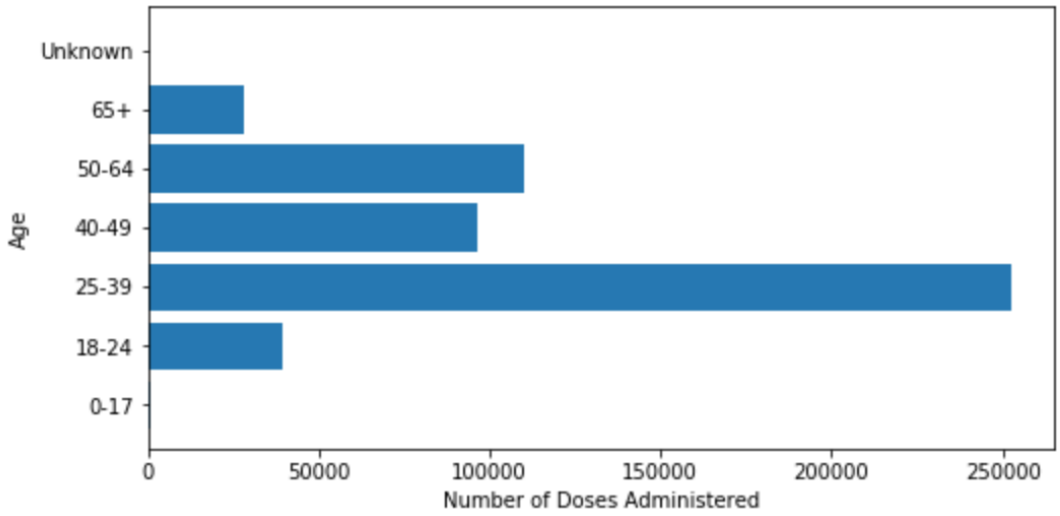

Figure 5: Total JYNNEOS vaccine doses administered by age

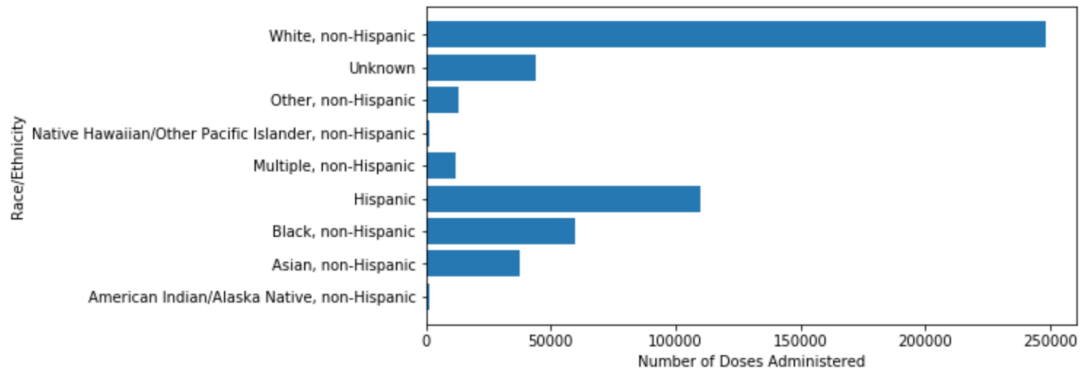

Figure 6: Total JYNNEOS vaccine doses administered by race/ethnicity

3 Methodology

Reddit has the capability to provide a platform for its users to express themselves on various topics through posts. Each given reddit post that corresponded to the topic returned a given title, body and comments. This project focused on filtering the information to obtain all data on monkeypox. For this to be completed Reddit API and PRAW was utilized for it to be a smooth process. Once that was completed artificial intelligence and deep learning, with the assistance of neural networks, was used to provide a clear interpretation of the data. A neural network is a learning system that transfer data through different layers and falls under the umbrella of machine learning. "Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy"[21].

3.1 Data Collection

The data used in the research was collected from the subreddits of the states California, Texas, Florida, New York, Georgia, Illinois, Pennsylvania, New Jersey, Maryland, Washington and North Carolina. Furthermore, to accurately see how citizens responded to the outbreak post were collected

from the time range January 1, 2022 to September 1, 2022. For a post to be included in the data it had to contain monkeypox, monkey pox or MPXV. By building a URL that was connected with Pushshift API it allowed the program to filter out posts that did not fit the required specifications. Each given post has a individual ID that corresponds to the information associated with that entry. Once the extraction of the data is complete it was transferred to a file.

3.2 Data Preprocessing

A major step following data collection was data Preprocessing. This was a key component as upon completion the data will be ready for further analysis. Data preprocessing can be expressed as raw data that is transformed to a specific format. To begin the process it was essential to remove the given URLs of the given post as they did not have any bearing on the results that were intended to receive. After that, the next phase was to lower case all of the data. It was a straightforward task and was successful due to the program operation that was already installed. Lower-casing was a needed step as it did not allow words to be grouped separately if the first letter was capitalized.

Next, text tokenization was used to separate every sentence into individual words. This was done by using Natural Language Toolkit (NLTK) which is a interface to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization,

stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries[11]. To get the data to be more clear stopwords were removed from the information. Stopwords are frequent used words that will not change the topic of the content if it is removed. A few examples of stopwords are the words 'are', 'is' and 'that'. It is also very helpful to eliminate any punctuation from the text. Punctuation removal is a common machine learning technique for preprocessing and working with user generated data provides a large amount of punctuation's. By tokenzing the data it provides a easy process for removing stopwords and punctuation.

A pivotal decision was utilizing lemmatization rather that stemming to analyze the meaning behind a word. Stemming is the process that chops off the beginning or ending of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes[16]. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma[16]. For example, by applying a stemming operation on the words studies and studying the two words will produce studi and study. The original words have the same meaning but after the process was completed they return two different results. Lemmatization provides more accurate results and using the NLTK WordNet Lemmatizer the data set was properly returned with a part of speech parameter.

3.3 Word Embedding

Word embedding is a technique that analyzes a text by utilizing the words as vectors to be implemented into a machine learning model. It is very popular among natural language processing applications as it can be implemented into many projects to provide great findings. A major feature of word embedding is allowing a word to predict the words near it. Furthermore, word embedding was performed using the application Word2Vec.

The word2Vec program was revealed in 2013 and provides natural language processing as it uses deep learning techniques to accumulate information within a corpus. Natural language processing is the teaching of a computer to comprehend written or verbal information as a human would. It can be broken down into morphological and lexical, analysis, syntactic analysis, semantic analysis, discourse integration and pragmatic analysis. As two people communicate a word can have multiple meanings and understanding the context of the conversation is necessary for proper comprehension of the intended message. Humans have developed this skill over time by the understanding of semantic relationship of words. Through the use of neural networks Word2vec take each word and produces vectors that show relations to one another through cosine similarity. It is constructed of the algorithms continuous bag-of-words(CBOW) and skipgram as it analyzes a corpus, a particular body of sentences, in their respective ways.

Moreover, CBOW can be expressed as having the text to be processed and then produces a word that is related to the data. The skipgram model

is the opposite as it receives a given word and tries to predict the context. By using window size, given number of words and workers, the application will provide accurate findings for each individual corpus.

Both models have their advantages and disadvantages based on their constructive architecture. It is noted that the CBOW model is expected to have a faster processing time and provides better results on words who are more common. On the other hand, a skipgram model works best with a small amount of text and with words that are not usual. The reason it works best with words that are uncommon is it does not calculate the average of word embeddings. By not doing so it can separate the meanings and provide an accurate analysis. A CBOW model is faster as it produces less computations than a skip gram model. By taking the desired number of words on both sides of the intended word are then used as information to produce the relevant data. Additionally, Skipgram will take much longer due to the necessary functions that are required to correctly understand the meaning of a word in its given context.

To gather a better understanding the figure below will show the process of a CBOW model. The words will be inputs using the one hot encoded method and are displayed as V. Also, N represents the number of neurons and V represents vectors that occur in the hidden layer. Wvn is a weight matrix that charts what has been inserted to the hidden layer and W'nv is a weight matrix that returns the hidden layer results to the final output layer. In addition to, a CBOW approach may be beneficial as it does not require

alot of storage but it does not have the ability to separate different meanings of the same word in an effective way.

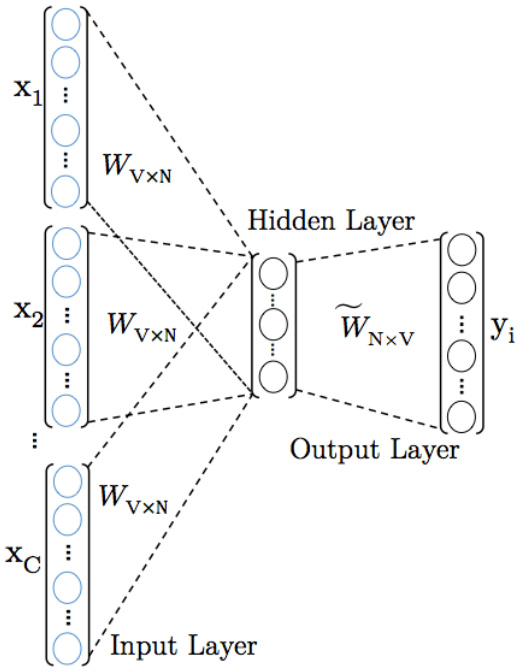

Figure 7: Architecture of a CBOW model

As for the skipgram model, the structure can be seen as reversed but that is not necessarily what has transpired. The given word is put into the model by using individual embedding layers through a one-hot encoded vectors that returns word embedding for each pair of terms. By using the dot product of the embeddings that value is returned and then calculated to be compared to other values which is achieved by backprorogation. This model is beneficial as it can differnate between different meanings of the same word.

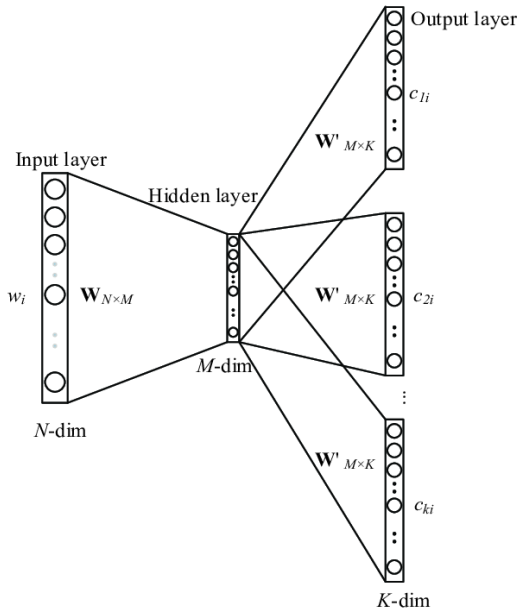

Figure 8: Architecture of a skip gram model

A commonly used measurement examination in data mining to compare two vectors is cosine similarity. Measuring the cosine similarity is achieved through the equation $cos(x, y) = x * y/||x|| * ||y||$. The equation obtains its desired results by using a dot product of the vectors. As the calculations return a cosine value relative to one it means the vectors are closely matched and the angle is small. When the cosine value approaches zero it means the vectors do not share a close relation and produces a larger angle. After the calculations are complete on the vectors they are then used to examine the similarities and differences of the corpus.

Another method that was used in word embedding was Global Vectors for Word Representation [14]. GloVe utilizes a corpus for its input and returns a co-occurrence matrix from the vectors that displays resulting representations who show interesting linear substructures of the word vector space[14]. This algorithm combines a local context-based method with matrix factorization ideas such as Latent Semantic Analysis. It uses a Log-Bilinear Regression Model that is equipped with a straightforward weighted least squares method. By computing the probabilities it will provide a clearer view on the given body of sentences.

An example of how the co-occurrence probability is constructed can be displayed in Figure 9. As one might expect, ice co-occurs more frequently with solid than it does with gas, whereas steam co-occurs more frequently with gas than it does with solid[14]. Both words co-occur with their shared property water frequently, and both co-occur with the unrelated word fashion infrequently[14]. Only in the ratio of probabilities does noise from non-discriminative words like water and fashion cancel out, so that large values (much greater than 1) correlate well with properties specific to ice, and small values (much less than 1) correlate well with properties specific of steam. In this way, the ratio of probabilities encodes some crude form of meaning associated with the abstract concept of thermodynamic phase[14].

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figure 9: Example of a GloVe model, Source : Stanford NLP

3.4 Topic Modeling

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents[12]. In the business world this is a common practice as it allows the understanding how human interactions are being documented in real time. For an example, many companies will release a new product and would like to know how their customers are responding in real time. It is difficult to process hundred of thousands of data and topic modeling plays a vital role in achieving the results. Topic modeling has many different methods and can range from Latent Dirichlet Allocation (LDA), Non Negative Matrix Factorization (NMF), Parallel Latent Dirichlet Allocation (PLDA) and Pachinko Allocation Model (PAM).

For this project, Latent Dirichlet Allocation(LDA), a customary algorithm for topic modeling, was chosen to be the basis for this project. In a simple form it treats documents to be a mixture of topics and each topic to be seen as a mixture of words which allows documents to "overlap" each other

in terms of content, rather than being separated into discrete groups[15]. Its end goal is to find representative words for a topic. For a LDA model to work it has to be given a corpus who is prepossessed by cleaning the data with the removal of stop words, lemmatization of the data and extraction of special characters. Through the use of genism and the matplotlib applications topic modeling was successful.

Topic modeling can be visualized by a many ways such as structural topic model(STM), cloud representation and plotQuotes. One of the ways the data was displayed was using word clouds that will display the words who are used the most from smallest to largest and how often each appears(monkeylearn). For the data it was best to utilize ten topics with twenty words that will show in each word cloud. After that, a word count was constructed over the topics to display the importance of the topic words.

4 Results

4.1 Data Collection and Preprocessing Results

The data was collected over various subreddits and returned a large sum of data that included the title, body and comments for every post. For this research, the top eleven states regarding monkeypox cases were chosen. Another reason for choosing those states were due to the fact they were also ten out of the top thirteen states in terms of population. The states are California, Texas, Florida, New York, Georgia, Illinois, Pennsylvania, New

Jersey, Maryland, Washington and North Carolina.

Once the data was inserted into a text file it was then ready to be processed. The results of that operation returned a smaller amount of information as words that were not useful were removed. In the figures below they will first show a text file before and after it was processed on a given body of text. From Figure 11, the text can be seen as broken up into individual words that were not flagged from the original sentence. The use of special characters and suffixes were removed throughout the text. For example, the terms contraindicated and 16% was changed to contraindicate and 16. There was a total of 22,622 words after the data collection and preprocessing steps were completed.

```
It's also contraindicated for anyone who's immunocompromised
(cancer, transplant recipients, etc.), has heart disease
(including high cholesterol, high blood pressure, etc) or
is pregnant. Oh, and everyone over 65 (another 16% of the population).
```

Figure 10: Example of a reddit post before preprocessing

```
also, contraindicate, anyone, immunocompromised, cancer,
transplant, recipient, etc, heart, disease, include, high,
cholesterol, high, blood, pressure, etc, pregnant, oh,
everyone, 65, another, 16, population
```

Figure 11: Example of a reddit post after preprocessing

4.2 Word Embedding Results

To examine the word embedding results the GloVe, Skip-Gram and CBOW models were compared to one another as they obtain different key words. Before the analysis could be achieved, each model was evaluated at different parameters to best fit this given corpus. The Skip-Gram and CBOW models utilized a minimum word count of four, window size of five and vector dimension of 400. Also, the GloVe model had a window size of five, thirty epochs, a learning rate of 0.05 and eight threads. By using cosine similarity the words vaccine and rash was used and the top five results.

By using vaccine for the GloVe model the words smallpox, get, monkeypox, man and spread was produced. These words are closely related as smallpox and monkeypox both share a similar Jynneos vaccine. Also, the CBOW and Skip-Gram models each generated monkeypox or smallpox respectfully. The second and fourth most related word was get and man. This correlates with the topic as most vaccine doses were administered predominantly to men around the country. For the word rash, the Skip-Gram model had the most relevant word embedding outcomes as they all correlate in their own way. The model produced the words fluid, scab, direct, body and respiratory. Fluids and scabs are common reactions that transpires as contracting monkeypox most occurring symptom. Furthermore, the words direct and body pairs well together as if their is any direct contact with a person or a surface the rash can be transmitted. The CBOW model results were not far off from the others but did not have much similarity as the

others.

Vaccine		
GloVe	CBOW	Skip-Gram
smallpox	get	smallpox
get	n't	covid
monkeypox	like	need
man	people	good
spread	monkeypox	well

Table 1: Table for GloVe, CBOW and Skip-Gram for Vaccine

Rash		
GloVe	CBOW	Skip-Gram
article	touch	fluid
make	someone	scab
touch	come	direct
thing	disease	body
feel	outbreak	respiratory

Table 2: Table for GloVe, CBOW and Skip-Gram for Rash

4.3 Topic Modeling Results

To produce the Topic Modeling results, a model was created to obtain ten topics from twenty key words for each given topic using LDA. There was a total of twenty passes completed over the corpus. Also, the size of the words in each given topic are proportional to their weight of that given topic. Words who are significantly used more will appear bigger than others

over the word cloud. The top word for each given LDA word cloud in order was gay, contact, monkeypox, see, monekypox, think, people, emergency, remove and million. Word clouds one and nine both produced the biggest differences in the first topic to secondary topics. The first topic correlates to one of the main reasons of the spread of the virus withe the terms male, men and sex being clearly visible. From the data it can support the cause as males account for nearly all cases. Topic 2 is focused on the prevention of the spread. This showcases the steps to stop the outbreak and the words contact, spread and touch are present. It also displays the word lockdown which was a major step to prevent the coronavirus from running its course. Next, topic three is more associated with the awareness of moneypox and the cases that have been calculated. Fear, talk and post are all three words that were represented. For topic four the word see stands out and mask, wear and interactions suggests they can see the given magnitude of the given situation. Similarly to topic four, topic 5 suggests a more serious approach in terms of protective measures by highlighting health, science and vaccine. This is an accurate assessment as the word mandate is present. To pair well with that is CDC and take is associated with this topic adding more elements to message. Topic six takes a more slower approach and is the least visible topic in comparison to the others. From the words that are present the common theme is the protection from the spread. The next cloud express a concern of the previous diseases that have been experienced in the past. Monkeypox, smallpox and covid all appear and they are closely related in some aspects.

Also, the word disease can be seen in this topic. Topic eight contains the words government, emergency, state and federal. From those terms it clearly shows the people are trying to analyze the government intervention into the matter at hand. The words declare and fund are suggesting they should play a role in slowing the transmission of the viral disease. The next topic is more focused on the informative measures in regards to internet platforms. Post, comment and website are words that are closely related to uploading information to the web to communicate with others. In fact, the platform youtube is present and it is capable of mass communication around the world. Topic ten is concerned with the community and includes the words home and live.

After that was completed, a word count distribution graph was constructed from each given topic data. Words are ordered in importance from left to right and the frequency of a word does not correlate to its importance. In fact, in only four of the graphs have the most used word to have the most weigh in its given topic. The graph in Figure 13 display the top ten words and monkeypox was the highest word in four out of the ten topics. In addition to, the term people was the second highest occurring in topics one, six and seven. Most of the words in the data set consist of various expressions that are related to stop the spread of monkeypox such as spread, get and vaccine. Another area that was highly present in the graphs was the overall well being of the people in society. Most terms correlate to the protection from monkeypox and the necessary steps to properly protect one another.

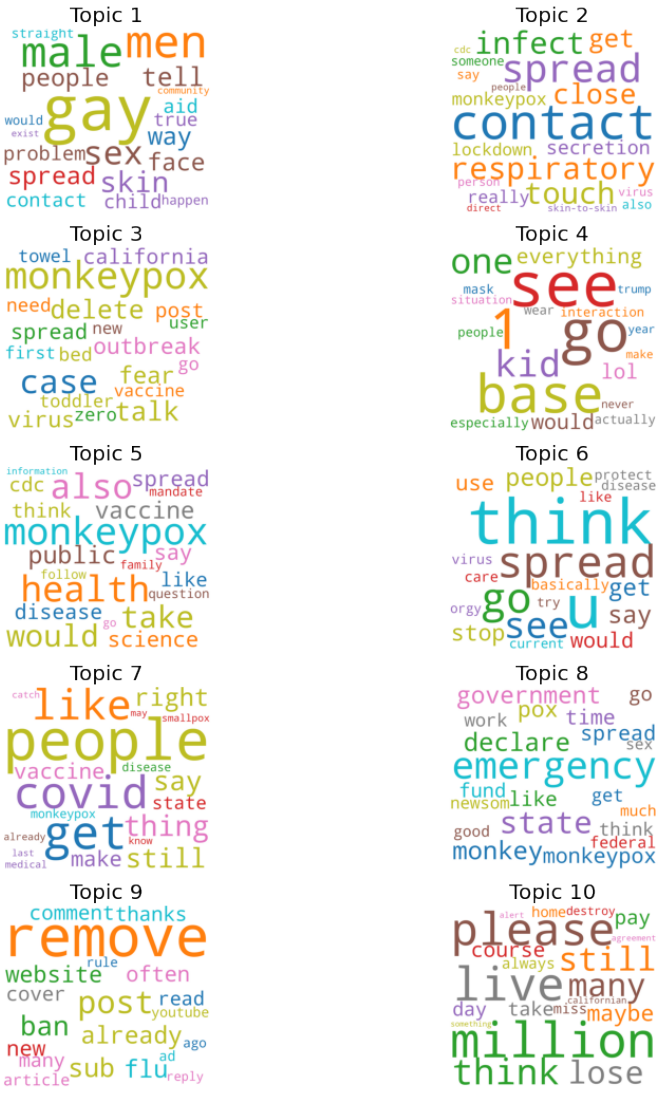

Figure 12: LDA word clouds for Topic Modeling

Figure 13: Word count distributions for Topic Modeling

5 Discussion

At the current point in time, the data suggest monkeypox does not raise a major concern in regards to the public health in the United States of America. There is no need for a shutdown of the country or mask mandate at the current time. Cases have not taken a huge increase and deaths are not being reported. Some symptoms have been common but the most frequent one does not take a drastic toll to the body. Males are most likely to obtain the virus than any other gender and more test are being done to understand. Also, there are more vaccinated individuals then people contracting the virus and that can be due to the various methods that are available. The word embedding results shows the community understands the role the vaccine can play and their desired to obtain it through the three models. They also understand the vaccine relation to smallpox. For the term rash, the severity of becoming in contact and how the transmission is prevalent to stop the spread. It also showcases the timeline of the rash before it runs its course. The topic modeling LDA word clouds express the many positions that have taken place in regards to monkeypox. Data from the subreddits range from the spread of the disease to government involvement on a national scale. In all, all of the task illustrates the people are aware of the situation and very proactive to make sure it does not grow at an exponential rate.

6 Conclusion and Recommendations

6.1 Summary

With the emergence of monkeypox in today's age it sparked interest throughout the globe. Monkeypox is a disease that is related to smallpox and has been around for a long time dating back to 1958. Throughout the years there has been multiple outbreaks that have spread the virus across the world. On random occasions there have been occurrences of outbreaks from animals and international travel. Cases this year have moved to different countries where monkeypox has never been before. Additionally, there were many organizations created to provide surveillance and protection from the disease. People with weaken immune systems have a higher rate of severe symptoms and individuals with multiple sex partners are at a higher risk. The primary result of contracting the virus will most likely cause a rash that can be contagious through surfaces. Monkeypox does have the ability to effect the respiratory system when it comes in contact with a host through various ways. A large sum of vaccinations have been given to the people of the country even though the vaccinations are not proven yet. Also, the age group twenty five to thirty nine have received the most vaccine doses administered. By using Reddit and its API, it was capable to obtain the given title, body and comments for monkeypox in the given subreddits. There was 22,924 words in the corpus after the data collection and preprocessing elements were completed. The preprocessing steps used NLTK and lemmatization

to provide a smooth process for word embedding. In doing so, the corpus utilized cosine similarity to compare the given models on the words rash and vaccine. Topic modeling was also done using a LDA model of word clouds with a word count distribution graph for each topic.

6.2 Primary Findings of this Study

According to the data on the United States it was discovered that the predominant race/ethnicity's to obtain the infection were White, Hispanic/Latino and Black/African American. Also, males is by far the highest gender to contract the virus at a alarming rate compared to others. The overall status of the country is not highly impacted as cases have only amassed around seven percent of the total population and there are not many deaths associated at the current time. Age groups between 21 to 30 and 31 to 40 are seen to be most likely to acquire monkeypox. Word embedding and topic modeling showcased the country is aware of monekypox and demonstrated their knowledge on how they intended to slow the spread of its leading symptom. The analysis also demonstrated how the leading protective measure has connections with smallpox and the urgency to use it. The four research questions that were stated at the beinging have been answered.

6.3 Ways to Improve Future Studies

To further examine monkeypox and its impact it would be useful to

incorporate data on a global scale from all countries from different parts of the world who does not share the same practices. By doing the investigation on a bigger arrangement of information, complete individual information for other domains that has been influenced by the monkeypox disease it will show a more accurate result on its total impact. A major component to improve the study is to gather verified vaccine administration information on other countries from credible databases. With the additions of financial, symptoms, age, gender, vaccine doses and race/ethnicity data from different cultures may provide a different view on the findings that were discovered. With the expansion to different locations it may be necessary to explore different social media outlets as they provide another avenue to analyze multiple audiences. A few examples of a platform to utilize would be twitter, youtube, facebook and etc as they all have millions of daily users that constantly contribute information ranging over various topics with different point of views. All of these suggestions would possible allow for a more clear understanding on severity of the monkeypox disease and the symptoms that occur with it as well as the analysis of the communities around the world.

References

[1] "2022 Outbreak Cases and Data." Centers for Disease Control and Prevention, 30 Sept. 2022, `https://www.cdc.gov/poxvirus/monkeypox/response/2022/index.html`.

[2] "About Monkeypox." Centers for Disease Control and Prevention, 22 July 2022, `https://www.cdc.gov/poxvirus/monkeypox/about/index.html`.

[3] "Demographics of Patients Receiving TPOXX for Treatment of Monkeypox." Centers for Disease Control and Prevention, 9 Nov. 2022, https://www.cdc.gov/poxvirus/monkeypox/response/2022/demographics-TPOXX.html.

[4] "History of Smallpox." Centers for Disease Control and Prevention, 20 Feb. 2021, https://www.cdc.gov/smallpox/history/history.html.

[5] "How Many People Are Lesbian, Gay, Bisexual, and Transgender?" Williams Institute, 3 Feb. 2021, https://williamsinstitute.law.ucla.edu/publications/how-many-people-lgbt/.

[6] "Jynneos." U.S. Food and Drug Administration,, https://www.fda.gov/vaccines-blood-biologics/jynneos.

[7] Ligon, B Lee. "Monkeypox: A Review of the History and Emergence in the Western Hemisphere." National Center for Biotechnology Information, U.S. National Library of Medicine, Oct. 2004, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7129998/.

[8] "Monkeypox Prevention." Virginia Department of Health, 2 Oct. 2022, https://www.vdh.virginia.gov/monkeypox/prevention/.

[9] "Monkeypox – Nigeria." World Health Organization, https://www.who.int/emergencies/disease-outbreak-news/item/21-december-2017-monkeypox-nigeria-en.

[10] "Monkeypox: Background Information." GOV.UK, https://www.gov.uk/guidance/monkeypox.

[11] NLTK, 25 Mar. 2022, https://www.nltk.org/.

[12] Pascual, Federico. "Topic Modeling: An Introduction." MonkeyLearn Blog, 26 Sept. 2019, https://monkeylearn.com/blog/introduction-to-topic-modeling/.

[13] "Past U.S. Cases and Outbreaks." Centers for Disease Control and Prevention, 6 June 2022, https://www.cdc.gov/poxvirus/monkeypox/outbreak/us-outbreaks.html.

[14] Pennington, Jeffrey, et al. "GloVe: Global Vectors for Word Representation." Glove: Global Vectors for Word Representation, 2014, https://nlp.stanford.edu/projects/glove/.

[15] Robinson, Julia Silge and David. "6 Topic Modeling: Text Mining with R." 6 Topic Modeling — Text Mining with R, https://www.tidytextmining.com/topicmodeling.html.

[16] "Stemming and Lemmatization", Cambridge University Press, 7 Apr. 2009, https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[17] "Strategies Adopted by Gay, Bisexual, and Other Men Who Have Sex with Men to Prevent Monkeypox Virus Transmission - United States, August 2022." Centers for Disease Control and Prevention, 1 Sept. 2022, https://www.cdc.gov/mmwr/volumes/71/wr/mm7135e1.htm?s_cid=mm7135e1_w.

[18] "Technical Report 2: Multi-National Monkeypox Outbreak, United States, 2022." Centers for Disease Control and Prevention, 1 Sept. 2022, https://www.cdc.gov/poxvirus/monkeypox/cases-data/technical-report/report-2.html.

[19] Tosh, Pritish K. "What Is Monkeypox, How Does It Spread and How Can It Be Prevented?" Mayo Clinic, Mayo Foundation for Medical Education and Research, 19 Aug. 2022, https://www.mayoclinic.org/diseases-conditions/infectious-diseases/expert-answers/monkeypox-faq/faq-20533608.

[20] "TPOXX (Tecovirimat)." U.S. Department of Health amp; Human Services, https://aspr.hhs.gov/monkeypox/TPOXXOperationalGuidance/Pages/TPOXX-tecovirimat.aspx.

[21] "What Is Deep Learning?" IBM, 1 May 2020, https://www.ibm.com/cloud/learn/deep-learning.