Sentiment Analysis of COVID-19 Tweets

Trends and patterns in emotional responses to the pandemic.

Team Members

Vignesh Murugan
Vivek Satya Sai Veera Venkata Talluri
Bhuvanesh So Muruganandam
Pardhu Burlu

Group Number: 10

Date: 26th November 2024

Class Information: Offline

TABLE OF CONTENTS

Abstract

This study explores public sentiment during the COVID-19 pandemic using a dataset of tweets collected from Kaggle. By applying text mining and sentiment analysis techniques, the study identifies shifts in public emotions during key events. Preprocessing steps ensured clean and meaningful data, while three sentiment lexicons—Bing, AFINN, and NRC—classified tweets into categories such as positive, negative, and specific emotions like anger or joy. Visualizations, including sentiment trends and word clouds, highlighted emotional responses and influential entities. The findings offer valuable insights for policymakers and analysts to understand and respond to public mood during crises.

Project Introduction

The Sentiment Analysis of COVID-19 Tweets project examines public emotions during the pandemic by analyzing tweets with relevant keywords. It uses text mining and sentiment analysis to classify tweets as positive, negative, or neutral, and identifies emotions like joy, anger, and sadness. The analysis incorporates three sentiment lexicons (Bing, AFINN, and NRC) to capture emotional patterns and sentiment intensity. Preprocessing steps like text cleaning and tokenization ensure meaningful results. Visualizations, including sentiment trends, word clouds, and network graphs, highlight keyword prominence and sentiment shifts over time. The project provides insights into public sentiment during major pandemic events, aiding policymakers, analysts, and researchers in understanding collective responses to global crises.

Description Of Data:

The COVID-19 Twitter dataset consists of 179,108 unique tweets, capturing pandemic-related discussions across social media. It includes 12 variables that offer insights into online engagement, such as user metrics (follower counts, friend networks, verification status) and temporal markers, which track the evolution of discourse throughout the pandemic.

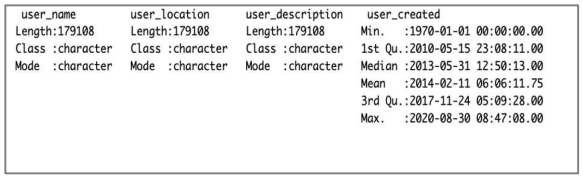
user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet
wednesday addams as a d	26/05/2017 05:46	624	950	18775	FALSE	25/07/2020 12:27	If I smelled the scent of hand sanitizers today		Twitter for iPhone	FALSE
Husband, Father, Columnis	16/04/2009 20:06	2253	1677	24	TRUE	25/07/2020 12:27	Hey @Yankees @YankeesPR and @MLB - wo		Twitter for Androi	FALSE
#Christian #Catholic #Cons	28/02/2009 18:57	9275	9525	7254	FALSE	25/07/2020 12:27	@diane3443 @wdunlap @r	['COVID19']	Twitter for Androi	FALSE
#Browns #Indians #Clevela	07/03/2019 01:45	197	987	1488	FALSE	25/07/2020 12:27	@brookbanktv The one gift	['COVID19']	Twitter for iPhone	FALSE
 Official Twitter handle c	12/02/2017 06:45	101009	168	101	FALSE	25/07/2020 12:27	25 July : Media Bulletin on Novel #CoronaVirusUpdates #COVID19 @kansalrohit69 @DrSyedSehrish @airnewsalerts @ANI… https://t.co/MN0EEcsJHh	['CoronaVirusUpda	Twitter for Androi	FALSE
# ссия #Novoro	19/03/2018 16:29	1180	1071	1287	FALSE	25/07/2020 12:27	#coronavirus #covid19 dea	['coronavirus', 'cov	Twitter Web App	FALSE
Workplace tips and advice	12/08/2008 18:19	79956	54810	3801	FALSE	25/07/2020 12:27	How #COVID19 Will Chan	['COVID19', 'Recru	Buffer	FALSE
	03/02/2012 18:08	608	355	95	FALSE	25/07/2020 12:27	You now have to wear face coverings when o		TweetDeck	FALSE
A poet, reiki practitioner ar	25/04/2015 08:15	25	29	18	FALSE	25/07/2020 12:26	Praying for good health and recovery of @ChouhanShivraj . #covid19 #covidPositive	['covid19', 'covidP	Twitter for Androi	FALSE

The dataset's content variables, such as tweet text, hashtags, user descriptions, and verification dates, provide insights into shifting narratives and sentiments during COVID-19. It also includes data on information spread (source, retweet status) and audience engagement. These variables enable sophisticated analyses of information diffusion, sentiment evolution, and community response, making the dataset valuable for understanding how social media influenced public perception during the pandemic.

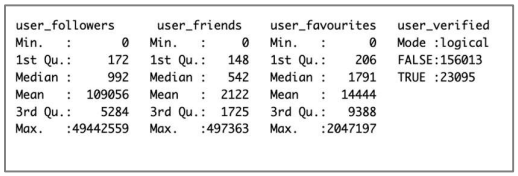
Statistical Summary of Data:

The dataset contains 179,108 tweets capturing public sentiment during the COVID-19 pandemic, with user account creation dates spanning from 1970 to 2020 and a median registration date of May 31, 2013. User followers range from 0 to over 49 million, with a median of 992, highlighting a mix of ordinary users and accounts with extensive reach. Similarly, user friends vary from 0 to 497,363,

with a median of 542, indicating diverse social networks. About 12.9% of the accounts are verified, reflecting contributions from both general users and prominent entities.



```
 user_name         user_location      user_description    user_created
Length:179108      Length:179108      Length:179108       Min.   :1970-01-01 00:00:00.00
Class :character   Class :character   Class :character    1st Qu.:2010-05-15 23:08:11.00
Mode  :character   Mode  :character   Mode  :character    Median :2013-05-31 12:50:13.00
                                                         Mean   :2014-02-11 06:06:11.75
                                                         3rd Qu.:2017-11-24 05:09:28.00
                                                         Max.   :2020-08-30 08:47:08.00
```



```
user_followers      user_friends     user_favourites   user_verified
Min.   :       0   Min.   :     0   Min.   :      0   Mode :logical
1st Qu.:     172   1st Qu.:   148   1st Qu.:    206   FALSE:156013
Median :     992   Median :   542   Median :   1791   TRUE :23095
Mean   :  109056   Mean   :  2122   Mean   :  14444
3rd Qu.:    5284   3rd Qu.:  1725   3rd Qu.:   9388
Max.   :49442559   Max.   :497363   Max.   :2047197
```

The dataset spans from July 24 to August 30, 2020, providing a focused snapshot of public engagement during this critical period. User favorites range from 0 to over 2 million, with a median of 1,791, indicating varying levels of activity and engagement. Textual data, enriched by hashtags and source information, captures key themes and platforms used for dissemination. This diversity in user demographics, activity levels, and content offers a rich basis for analyzing sentiment trends, identifying influential entities, and uncovering patterns in public discourse during the pandemic.

Data Understanding and Preparation

Handling missing values:
The dataset exhibits a notable presence of missing values across multiple columns, which may influence the reliability and depth of the analysis. The user_location column contains the highest number of missing entries, with 36,804 records lacking location data. Similarly, the hashtags column has 51,334 missing entries, indicating a significant absence of associated topics or themes in the tweets.

	Column	MissingCount
user_location	user_location	36804
user_description	user_description	10283
hashtags	hashtags	51334
source	source	77

The user_description column has 10,283 missing values, limiting insights into user profiles and biases, while the source column has only 77 missing values. Addressing these missing values is crucial for analysis integrity, with imputation or record removal as possible solutions, depending on the study's goals. Proper handling ensures the dataset remains robust and the insights accurate.

Cleaning Text Data:
In the process of cleaning the text data, several transformations were applied to ensure the information was consistent and free of noise. First, all URLs were removed, including both full URLs and domain mentions, to prevent any external links from skewing the analysis. For instance, the text "Check this link http://example.com or www.example.com" was transformed into "Check this link or." Next, mentions of users, such as "@user1" and "@user2," were eliminated to maintain focus on the content rather than specific individuals, resulting in cleaner text like "Hello, how are you?" Hashtags were also removed from the text, ensuring that expressions like "#sunny" and "#vacation" were simplified to just "Loving the weather!" Finally, special characters, including punctuation and numbers, were removed, transforming phrases like "Hello, world! It's 2024." into "Hello world Its." These steps helped to produce a refined version of the text, free from extraneous elements, stored in a new clean_text table, which can now be used for further analysis or processing.

text	hashtags	source	is_retweet	cleaned_text
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	if i smelled the scent of hand sanitizers today on som...
Hey @Yankees @YankeesPR and @MLB - wouldn't it h...	*NA*	Twitter for Android	FALSE	hey and wouldnt it have made more sense to have ...
@diane3443 @wdunlap @realDonaldTrump Trump ne...	['COVID19']	Twitter for Android	FALSE	trump never once claimed was a hoax we all claim ...
@brookbanktv The one gift #COVID19 has give me is ...	['COVID19']	Twitter for iPhone	FALSE	the one gift has give me is an appreciation for the si...
25 July : Media Bulletin on Novel #CoronaVirusUpdate...	['CoronaVirusUpdates', 'COVID19']	Twitter for Android	FALSE	july media bulletin on novel
#coronavirus #covid19 deaths continue to rise. It's al...	['coronavirus', 'covid19']	Twitter Web App	FALSE	deaths continue to rise its almost as bad as it ever ...
How #COVID19 Will Change Work in General (and recr...	['COVID19', 'Recruiting']	Buffer	FALSE	how will change work in general and recruiting speci...
You now have to wear face coverings when out shopp...	*NA*	TweetDeck	FALSE	you now have to wear face coverings when out shoppi...
Praying for good health and recovery of @ChouhanShi...	['covid19', 'covidPositive']	Twitter for Android	FALSE	praying for good health and recovery of
POPE AS GOD - Prophet Sadhu Sundar Selvaraj. Watch...	['HurricaneHanna', 'COVID19']	Twitter for iPhone	FALSE	pope as god prophet sadhu sundar selvaraj watch he...
49K+ Covid19 cases still no response from @cbseind...	*NA*	Twitter Web App	FALSE	k covid cases still no response from please cancel t...
Order here: https://t.co/4NUrGX6EmA #logo #graphi...	['logo', 'graphicdesigner', 'logodesign', 'logodesinger',...	Twitter Web App	FALSE	order here
👏@PattyHajdu @NavdeepSBains — no one will be saf...	['COVID19']	Twitter Web App	FALSE	no one will be safe from until everyone is safe will ...
Let's all protect ourselves from #COVID19. It's real an...	['COVID19']	Twitter Web App	FALSE	lets all protect ourselves from its real and the numbe...
Rajasthan Government today started a Plasma Bank at...	*NA*	Twitter Web App	FALSE	rajasthan government today started a plasma bank at...
Nagaland police on Covid-19 Awareness at City Towe...	['Covid19', 'keepsocialdistance']	Twitter for Android	FALSE	nagaland police on covid awareness at city tower junc...
July 25 #COVID19 update #TamilNadu - 6988 Dischar...	['COVID19', 'TamilNadu', 'chennai']	Twitter for iPhone	FALSE	july update discharge people tested actice cases ...
Second wave of #COVID19 in Flanders..back to more ...	['COVID19', 'homework']	Twitter for Android	FALSE	second wave of in flandersback to more again
It is during our darkest moments that we must focus ...	['light']	Twitter Web App	FALSE	it is during our darkest moments that we must focus ...
COVID Update: The infection rate in Florida is followin...	*NA*	Twitter for iPad	FALSE	covid update the infection rate in florida is following t...
@EvanAKilgore @realDonaldTrump Good Patriots! Call...	*NA*	Twitter for iPhone	FALSE	good patriots call to volunteer to be an election judg...
Coronavirus - South Africa: COVID-19 update for Sou...	*NA*	Africa Newsroom	FALSE	coronavirus south africa covid update for south afric...

Tokenizing Words:
In the tokenization process, we focused on breaking down the cleaned tweet text into individual words while removing the most common, non-essential words, often referred to as "stop words." These stop words, such as "I," "the," "and", "was," and others, do not carry significant meaning for analysis purposes, especially in tasks like sentiment analysis or word frequency analysis.

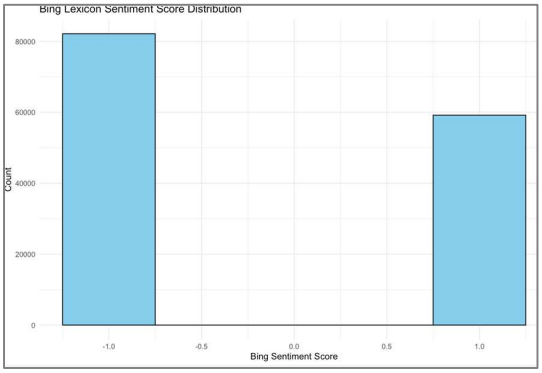
text	hashtags	source	is_retweet	word
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	smelled
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	scent
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	hand
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	sanitizers
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	past
If I smelled the scent of hand sanitizers today on som...	*NA*	Twitter for iPhone	FALSE	intoxicated

The sentence "If I smelled the scent of hand sanitizers today on someone in the past, I would think they were so intoxicated that" is tokenized into key words like "smelled," "hand," "scent," "sanitizers," "past," and "intoxicated." The resulting dataset, tweets_tokenized, includes a "word" column with meaningful words from each tweet, excluding filler words. This clean and tokenized data is ready for further analysis, such as frequency counting or sentiment analysis.

Data Analysis

BING sentiment Scores and Distribution:
Bing sentiment scores provide a quantitative measure of the emotional tone in text data by using the Bing lexicon, which classifies words as either positive or negative.

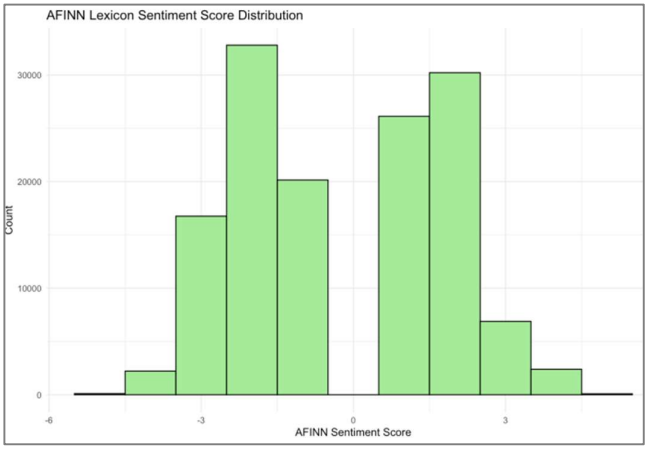
id	negative	positive	bing_sentiment_score
2	0	1	1
5	0	1	1
6	0	1	1
7	0	1	1
8	0	1	1
9	0	1	1
13	0	1	1
14	0	1	1
16	0	1	1
17	0	1	1
19	0	1	1
22	0	1	1
23	0	1	1
24	0	1	1



The process involves tokenizing text into words and matching them with the Bing lexicon to count positive and negative words. A sentiment score is calculated by subtracting the count of negative words from positive ones. This score helps categorize the text as positive, negative, or neutral. In the dataset of 141,340 observations across 4 variables, over 80,000 words were negative, and around 60,000 were positive. These sentiment scores provide insights into public mood and opinion.

AFINN sentiment Scores and Distribution:
AFINN scores provide a more refined sentiment analysis by assigning a numeric score to words, ranging from -5 to +5, based on their emotional intensity, with negative scores reflecting negative sentiment and positive scores indicating positive sentiment. In the process, text is tokenized into words, matched with the AFINN lexicon, and the scores are summed to calculate an overall sentiment score for each document.

id	afinn_sentiment_score
1	-1
2	-2
3	2
4	2
5	-3
6	1
7	1
8	-1
9	1
10	1
11	1
12	1
13	1
14	-2

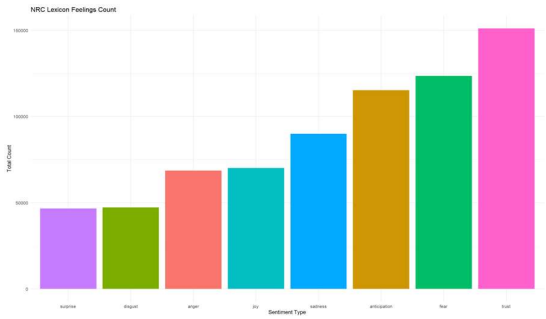


The AFINN Lexicon Sentiment Score Distribution in the dataset of 137,730 observations shows a significant number of negative tweets, with around 17,500 in the negative range. Many tweets fall in the neutral range, peaking at 32,500, while positive tweets are fewer, peaking at 30,000. This distribution indicates that negative and neutral sentiments are more common, suggesting a critical or cautious tone toward the topic.

NRC sentiment Scores and Distribution:

NRC scores provide a deeper analysis of text sentiment by categorizing words into various emotions (such as joy, anger, fear) and sentiments (positive or negative) using the NRC lexicon. The process begins by tokenizing the text into words, matching these words with the NRC lexicon, and then counting the occurrences of each word in different emotional and sentiment categories. This allows for a richer understanding of the emotional tone of the text, capturing not only the polarity (positive or negative) but also the specific emotions conveyed.

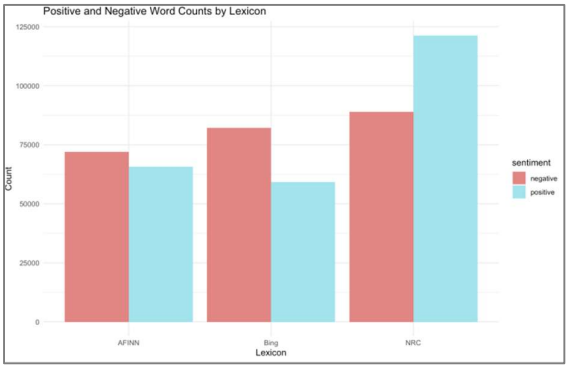
disgust	negative	positive	anticipation	joy	trust	surprise	anger	sadness	fear	nrc_positive_score	nrc_negative_score
1	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0



In the dataset of 566,460 observations, the NRC Lexicon Sentiment and Emotion Distribution shows that Trust and Joy are the most frequently expressed emotions, indicating a largely positive sentiment. Emotions like Surprise and Disgust are less common, while Anger, Sadness, Anticipation, and Fear show moderate frequencies, reflecting a mix of negative and positive emotions. The NRC lexicon, useful for social media monitoring, sentiment analysis, market research, and mental health analysis, reveals that Trust occurs around 150,000 times, Fear 125,000, Anticipation 110,000, Sadness 90,000, Joy 75,000, Anger 70,000, and Surprise and Disgust each around 50,000. This distribution provides a comprehensive view of emotional expressions in the text.

Comparative Analysis of Positive and Negative Sentiments Across Lexicons

The visualization compares the distribution of positive and negative words across three sentiment lexicons: AFINN, Bing, and NRC, highlighting key trends in the sentiment of the analyzed text data. The visualization compares the distribution of positive and negative words across three sentiment lexicons: AFINN, Bing, and NRC. Bing shows 82,151 negative words and 59,189 positive words, indicating a stronger negative sentiment. AFINN has a more balanced distribution with 72,022 negative and 65,708 positive words.

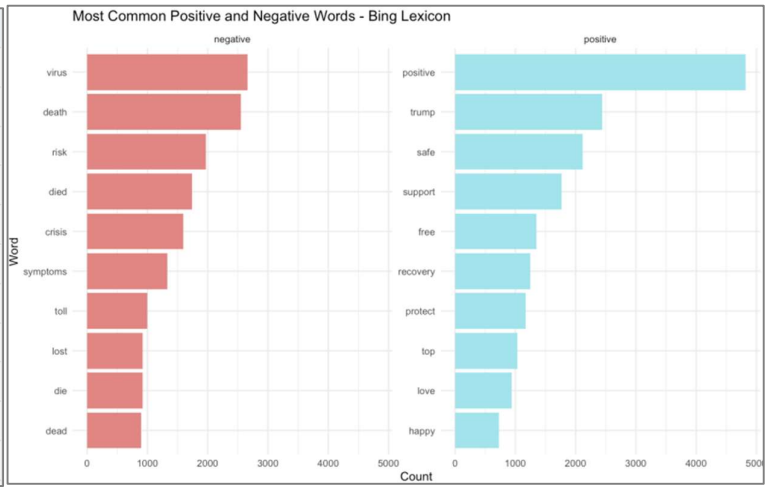


NRC shows a stronger negativity, with 121,279 negative words and 88,962 positive words. These differences highlight variations in lexicon sensitivity, with the dataset overall leaning slightly

negative, especially in NRC, emphasizing the importance of choosing the right lexicon for specific analysis goals.

BING Lexicon: Common Positive and Negative Words:
The data highlights the most common words associated with positive and negative sentiments as classified by the Bing lexicon. On the negative side, words like "virus" (2658 occurrences), "death" (2546), and "risk" (1972) dominate, reflecting concerns about health crises or danger. Similarly, "died," "crisis," and "symptoms" reinforce the negative themes of mortality and health issues.

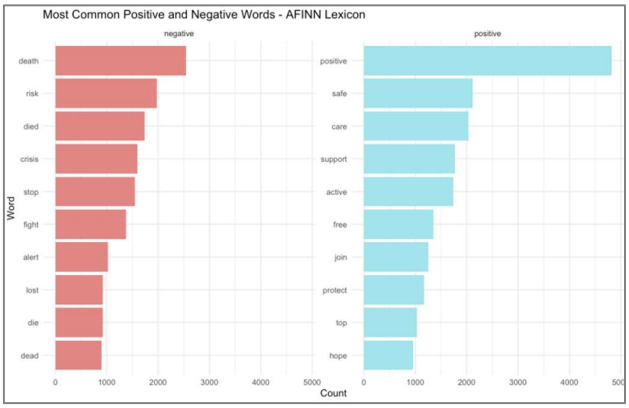
	word	sentiment	n
1	virus	negative	2658
2	death	negative	2546
3	risk	negative	1972
4	died	negative	1736
5	crisis	negative	1593
6	symptoms	negative	1329
7	toll	negative	995
8	lost	negative	920
9	die	negative	919
10	dead	negative	895
11	positive	positive	4821



On the positive side, the word "positive" itself appears most frequently (4821 occurrences), often used in contexts like "positive cases" or optimism, followed by "Trump" (2440), indicating discussions around leadership or related topics. Words like "safe" (2118), "support" (1769), and "recovery" (1251) emphasize themes of security, assistance, and healing. This data reveals contrasting themes of fear and optimism, providing a snapsh**ot of** public discourse or sentiments within the dataset.

AFINN Lexicon: Common Positive and Negative Words:
The AFINN sentiment analysis reveals the most frequently used words with varying emotional intensities. On the negative side, words such as "death" (2546), "risk" (1972), and "died" (1736) appear prominently, reflecting strong associations with fear, loss, and danger. Other terms like "crisis" (1593), "stop" (1551), and "fight" (1373) highlight urgency and conflict. Conversely, on the positive side, "positive" (4821) is the most frequent, suggesting either optimism or discussions around "positive cases."

	word	sentiment	n
1	death	negative	2546
2	risk	negative	1972
3	died	negative	1736
4	crisis	negative	1593
5	stop	negative	1551
6	fight	negative	1373
7	alert	negative	1020
8	lost	negative	920
9	die	negative	919
10	dead	negative	895

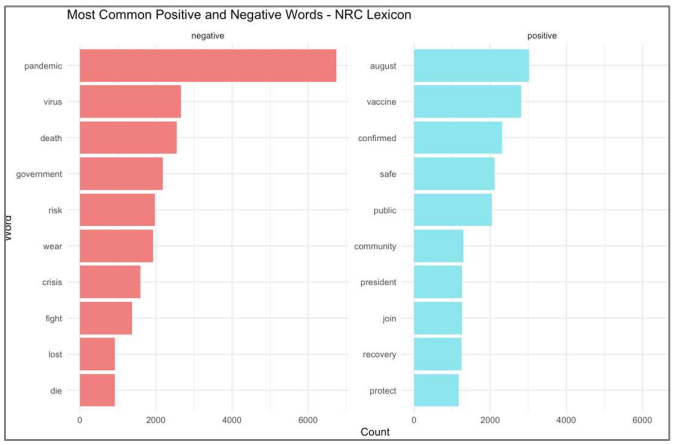


Words like "safe" (2118), "care" (2035), and "support" (1769) convey themes of reassurance, well-being, and assistance. Additionally, "hope" (954) and "protect" (1169) emphasize resilience and proactive measures. The data from AFINN offers insight into polarized emotional tones, ranging from distress and fear to safety and optimism, providing a detailed picture of public sentiment.

NRC Lexicon: Common Positive and Negative Words:

The NRC sentiment analysis reveals distinct emotional tones in the data. Negative words are dominated by "pandemic" (6736 occurrences), reflecting its overwhelming association with fear and uncertainty, followed by "virus" (2658) and "death" (2546), emphasizing health-related anxieties.

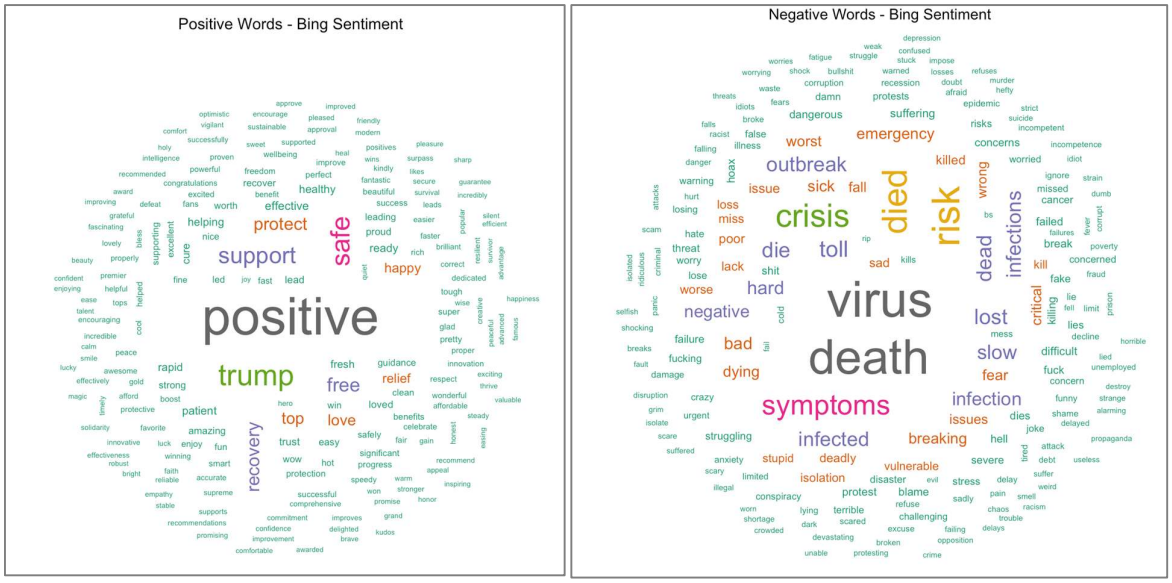
	word	sentiment	n
1	pandemic	negative	6736
2	virus	negative	2658
3	death	negative	2546
4	government	negative	2176
5	risk	negative	1972
6	wear	negative	1920
7	crisis	negative	1593
8	fight	negative	1373
9	lost	negative	920
10	die	negative	919
11	august	positive	3023
12	vaccine	positive	2814
13	confirmed	positive	2305



Words like "government" (2176) and "crisis" (1593) highlight political and societal challenges, while "risk" (1972) and "wear" (1920) relate to safety concerns. Positive words like "august" (3023), "vaccine" (2814), "safe" (2118), "public" (2035), and "community" (1298) reflect hope, security, and collective effort. "Recovery" (1251) and "protect" (1169) suggest optimism and proactive measures. The NRC analysis reveals not only sentiment but also emotional undercurrents within the dataset.

BING WordCloud Analysis of Positive and Negative Sentiments:

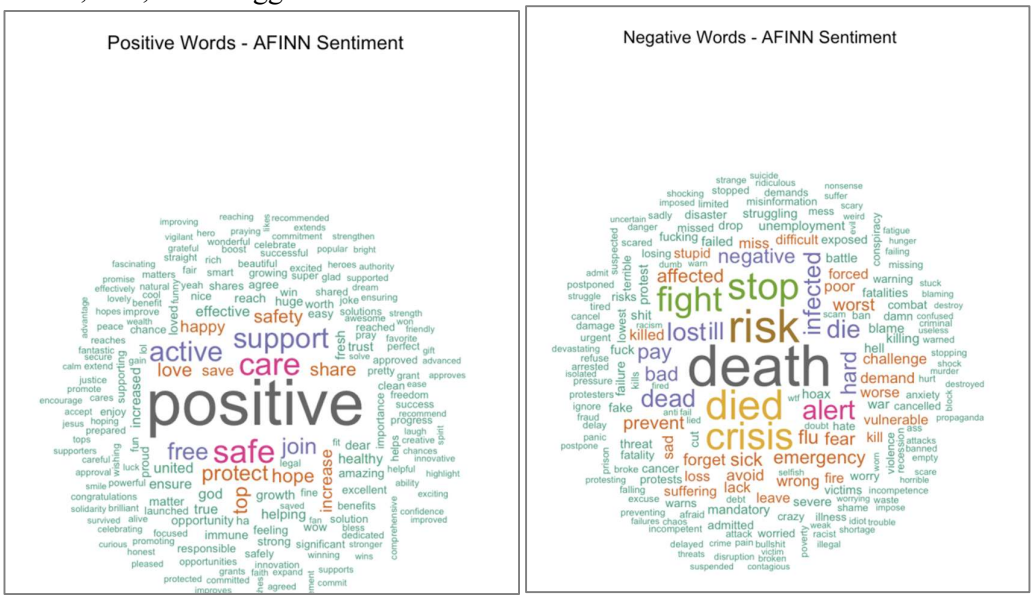
The WordCloud analysis of Bing Lexicon sentiments reveals contrasting narratives within the dataset. Positive sentiment focuses on themes of progress, security, and well-being.



Words like "success," "happy," and "protect" reflect optimism, personal growth, and a desire for safety, while terms like "virus," "death," and "crisis" highlight fear, uncertainty, and health-related challenges. This contrast captures a dual narrative of hope and concern, illustrating the dataset's complex emotional landscape.

AFINN WordCloud Analysis of Positive and Negative Sentiments:

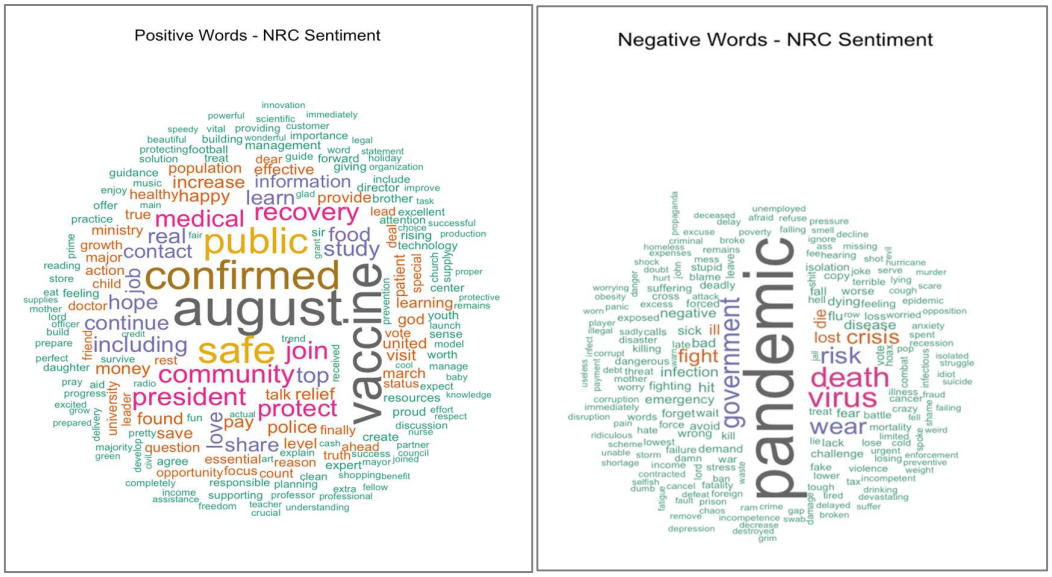
The AFINN positive sentiment word cloud emphasizes support, growth, and well-being, with words like "love," "support," and "care" highlighting community and empathy. Terms like "success," "win," and "growth" focus on advancement, while "support" reappears, underscoring collective effort. Although "celebrate" and "joy" are less frequent, the overall tone remains optimistic, reflecting resilience, hope, and progress. In contrast, the AFINN negative sentiment word cloud emphasizes themes of fear, loss, and struggle.



Words like "death," "crisis," and "risk" underscore the severity and emotional impact of challenging situations. Terms such as "loss," "failure," and "worry" capture the anxiety and distress felt in the face of adversity. Additionally, words like "fight," "battle," and "struggle" suggest a sense of resistance and defiance. Overall, the word cloud conveys a somber and anxious tone, highlighting the emotional burden and uncertainty surrounding the events depicted.

NRC WordCloud Analysis of Positive and Negative Sentiments:

The NRC positive sentiment word cloud focuses on health, community, and progress, with words like "healthy," "safe," "protect," "community," "together," and "hope" reflecting well-being, solidarity, and optimism. In contrast, the negative sentiment word cloud highlights fear, loss, and illness, with terms like "death," "virus," "pandemic," "loss," "suffering," and "fear" emphasizing health crises and emotional distress.



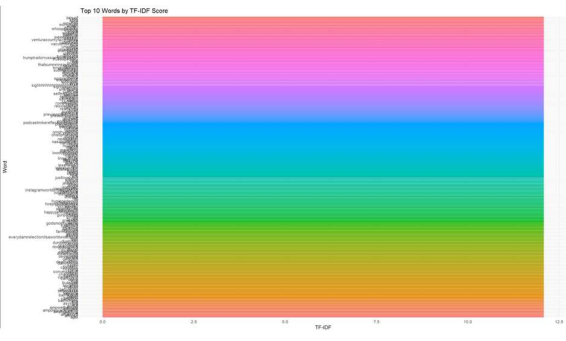
Additionally, words like "crisis," "risk," and "danger" amplify the severity and uncertainty surrounding the pandemic. Overall, the word cloud paints a somber picture, reflecting the collective anxiety, fear, and uncertainty tied to the ongoing crisis.

TF-IDF Analysis:

The TF-IDF analysis of the COVID-19 Twitter dataset, encompassing 179,108 tweets, reveals critical patterns in pandemic-related social media discourse through sophisticated statistical text analysis. By processing the dataset's 12 key variables, particularly focusing on tweet text content, user descriptions, and hashtags,

	text	word	n	tf	idf	tf_idf
1	#Abril sorprendeme!!!! @todonoticias @Poggi #LaNa...	sorprendeme	1	1	12.08071	12.08071
2	#BOUNTYONOURMILITARY #TRE45SON #COVID19 #...	trumptraitorrussiantooltreason	1	1	12.08071	12.08071
3	#Berlin2908 Best of Hygienedemos https://t.co/umsn...	hygienedemos	1	1	12.08071	12.08071
4	#COVID19 #GoodNewsforAmerica #BadNewsforC...	dropp	1	1	12.08071	12.08071
5	#COVID19 #COVID−19 #coronavirus #BlackLivesMatt...	dialo	1	1	12.08071	12.08071
6	#COVID19 #Islam #الكويت Welcome to islam...part1 htt...	islampart	1	1	12.08071	12.08071
7	#COVID19 #Jalisco | POST_TITLE) - https://t.co/ylWki...	posttitle	1	1	12.08071	12.08071
8	#COVID19 : prudence... https://t.co/5h4lw2rBNB	prudence	1	1	12.08071	12.08071
9	#COVID19: #ImmuneSystem derails https://t.co/5Rz...	derails	1	1	12.08071	12.08071
0	#COVID−19 #COVID19 #pandemic #goodhealth&am...	ampwellbeingall	1	1	12.08071	12.08071
1	#California I want you to see something very importa...	quietl	1	1	12.08071	12.08071
2	#Corona lebt! #COVID19 #COVID19france #CotedAzu...	lebt	1	1	12.08071	12.08071
3	#CoronaInfoCH #COVID19 #corona #france #trending...	canic	1	1	12.08071	12.08071
4	#CoronaInfoCH #COVID19 #corona #ireland &#...	ampdisgracefula	1	1	12.08071	12.08071
5	#CoronaInfoCH #COVID19 #corona #ireland Holiday...	holidaymak	1	1	12.08071	12.08071
6	#Covid19 @TheIPA's #scomo #economy #Welfare UN...	unworthy	1	1	12.08071	12.08071

canic



TF-IDF scoring highlights significant and distinctive terms in pandemic-related Twitter conversations by assigning higher weights to words that are frequent in specific tweets but rare across the dataset. The analysis reveals terms like "venturacounty," "whooohoo," and compound words with high TF-IDF scores, indicating localized or emotionally charged discussions. The visualization shows a long tail distribution, with many terms having low scores, while a few stand out with higher scores, pointing to specific and contextual conversations. This analysis provides insights into regional specificity and emotional expressions in COVID-19 discourse on social media.

Creating Word Embedding:
Word embedding is a key technique in natural language processing (NLP) that transforms text into numerical vectors, capturing the semantic meaning of words. In this analysis, we applied word embedding techniques to a corpus of tweets to explore word relationships and perform semantic analysis.

Preprocessing and Tokenization: The text data was first preprocessed by converting it to lowercase, removing URLs, mentions, hashtags, and special characters, and then tokenized into individual words. This cleaned text was used for creating the word embeddings.

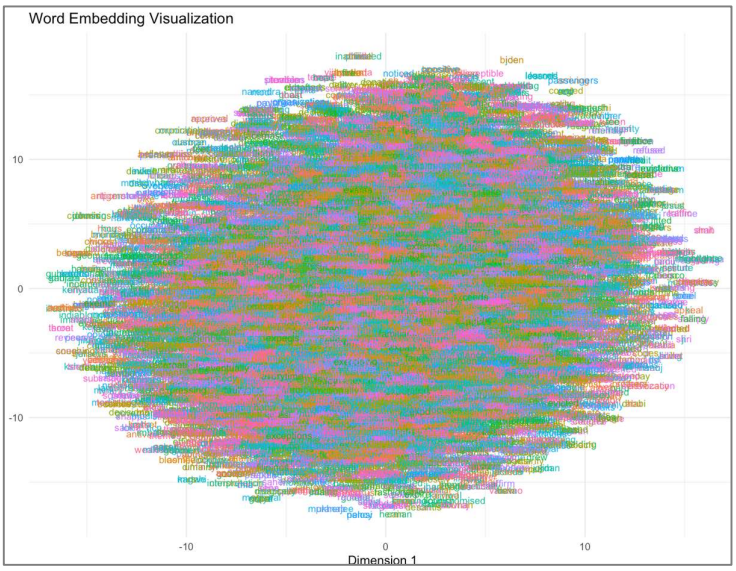
```
Number of docs: 179108
0 stopwords:  ...
ngram_min = 1; ngram_max = 1
Vocabulary:
            term term_count doc_count
          <char>      <int>     <int>
    1:      aaye          5         5
    2:       abe          5         5
    3: abilities          5         5
    4: abundance          5         4
    5:   abusers          5         5
   ---
18235:         a      47909     40626
18236:        in      54614     47340
18237:        of      59878     51483
18238:        to      75000     61175
18239:       the     105889     76521
>
```

Creating the Co-occurrence Matrix and GloVe Model: A co-occurrence matrix was built to capture word pair frequencies within a context window of five words. We then applied the GloVe (Global Vectors for Word Representation) model to generate 50-dimensional vector representations of words. These embeddings capture semantic relationships, allowing words with similar meanings to have closer vector representations.

```
> print(word_embeddings["covid", ])
 [1] -0.386142115  0.725196380  0.365116631  1.016195872  0.213265004  0.104133467  0.403253656  0.001741413
 [9]  0.231250514  0.849710862  0.437967450 -0.257774641 -0.334914382 -0.591147793  0.760741127 -0.435982616
[17] -0.231858120 -0.084040257 -0.204687166 -1.057046162  1.222884200  0.749491953  0.687541905  1.113173140
[25]  0.708417008  1.192937095  0.724961307  1.025566210  1.370430645 -0.713284437 -0.468820358 -0.011142985
[33]  0.988082890 -1.287806601 -0.705850707  0.554554073 -0.157389733  0.052534791  0.952155825  0.624842523
[41] -0.942887414  1.464240300  0.441689478 -0.407612953  1.221396785  0.261881428 -0.391155352 -0.577510312
[49] -0.822828273 -0.302473503
> print(word_embeddings["vaccine", ])
 [1]  0.343020583  0.357076468 -0.213237556  1.743466056  0.552165301 -0.734639699 -0.272383642 -0.235590488
 [9]  0.419721433  0.077052084  1.304718516 -0.581063382 -0.273233576 -0.167118929  1.252219696 -0.097640437
[17]  0.830303861 -0.470793924  0.356415665 -0.541460304  0.787471564  1.308885026  0.768049016  0.299048871
[25]  0.508235993  0.179595080 -0.323591778  0.848802611 -0.444107662 -0.094909715  0.191805566  0.103972600
[33]  1.234734838 -0.215751022  0.528825657  0.851750164  0.876579649 -0.998900282  0.854545943  0.333610662
[41]  0.465999533  1.542080934 -0.008690439 -0.430806631  0.803948898  0.513255006 -0.011778005 -0.971239921
[49] -1.302347074  0.280762691
>
```

Visualization with t-SNE: To visualize the high-dimensional word embeddings, we applied t-SNE for dimensionality reduction, reducing the embeddings from 50 to 2 dimensions. This visualization allowed us to observe clusters of semantically similar words, providing insights into word relationships and contextual meanings.

Tweet-Level Embeddings: We aggregated the word embeddings of individual tweets by averaging the vectors of the words in each tweet. This generated a single vector for each tweet that captured its overall semantic meaning. The resulting matrix of tweet embeddings can be used for tasks such as sentiment analysis, clustering, or classification. By using word embeddings and techniques like GloVe and t-SNE, we transformed textual data into numerical representations that can be easily analyzed. This approach is essential for understanding word semantics and enables the application of machine learning models for tasks like sentiment analysis and topic modeling.



Adding Bing Sentiment Scores to Tweets

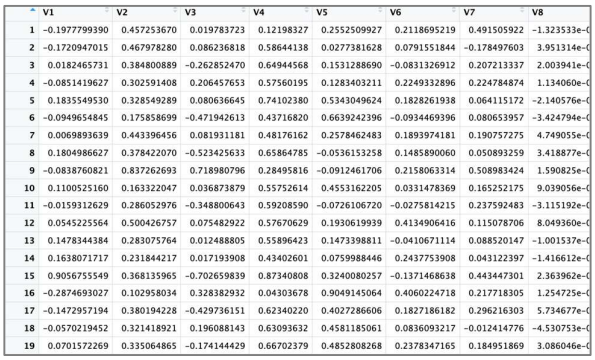
Bing sentiment scores were added to each tweet to enhance the dataset for machine learning tasks like Logistic Regression and Random Forest. Sentiment was determined using the Bing lexicon, with scores calculated by subtracting negative words from positive ones. Missing values were handled with mode imputation, and a sentiment label column was created, marking tweets with positive scores as "positive" and others as "negative." This process adds valuable features for sentiment analysis, improving machine learning models' ability to accurately predict tweet polarity.

	source	is_retweet	embedding	id	bing_sentiment_score
	Twitter for iPhone	FALSE	c(" 0.05308060", " 0.17584453", "-0.32662326", " 0 [...	1	-1
	Twitter for Android	FALSE	c(" 0.07883078", " 0.44798797", " 0.06958731", " 0 [...]	2	1
	Twitter for Android	FALSE	c(" 0.279617718", " 0.192229448", "-0.336196205", ...	3	-1
	Twitter for iPhone	FALSE	c(" 0.08457065", " 0.28102789", "-0.06923480", " 0 [...	4	-1
	Twitter for Android	FALSE	c("-0.43555338", " 0.22029106", "-0.26094202", " 0 ...	5	1
	Twitter Web App	FALSE	c("-0.109919973", " 0.180666701", "-0.089434105",...	6	1
	Buffer	FALSE	c("-0.100555085", " 0.315582319", " 0.047367301", ...	7	1
	TweetDeck	FALSE	c(" 0.160451129", " 0.705336911", "-0.114764676", ...	8	1
	Twitter for Android	FALSE	c("-0.139833899", " 0.606532758", "-0.112428844",...	9	1
	Twitter for iPhone	FALSE	c("-0.156207208", " 0.026544722", "-0.348852530",...	10	-1
	Twitter Web App	FALSE	c("-0.108171812", " 0.616035301", " 0.115306574", ...	11	-1
...	Twitter Web App	FALSE	c("-0.058482130", " 1.074065047", " 0.259251011", ...	12	-1
	Twitter Web App	FALSE	c(" 0.07218355", " 0.65360639", "-0.15072198", " 0 [...	13	1
	Twitter Web App	FALSE	c("-0.080507150", " 0.283459850", " 0.011422073", ...	14	1
	Twitter Web App	FALSE	c("-0.100729784", " 0.505648791", "-0.204352892",...	15	-1
	Twitter for Android	FALSE	c("-0.038330164", "-0.042391523", "-0.205920732"...	16	1
	Twitter for iPhone	FALSE	c("-0.47448807", " 0.21707546", "-0.74139432", " 0 ...	17	1
	Twitter for Android	FALSE	c("-0.123366195", " 0.462933875", " 0.057018295", ...	18	-1
	Twitter Web App	FALSE	c("-0.10507915", " 0.44580464", "-0.13390943", " 0 ...	19	1
	Twitter for iPad	FALSE	c("-0.128520649", " 0.285033945", " 0.094199749", ...	20	-1
	Twitter for iPhone	FALSE	c(" 0.132069755", " 0.711838113", "-0.208816121", ...	21	-1
	Africa Newsroom	FALSE	c("-0.69597508", " 0.13689101", "-0.01312635", " 0 ...	22	1

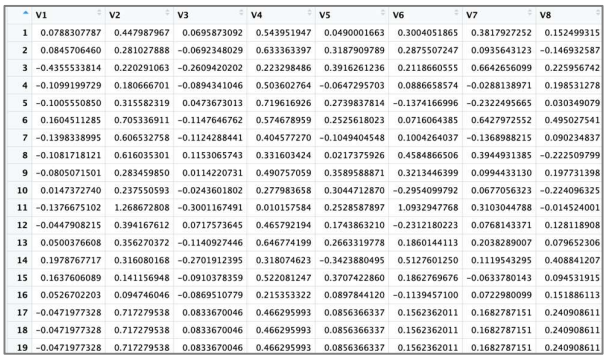
Train-Test Split Of Dataset

The First image shows the X_Train, data, and the second image shows X_Test data. The vector embeddings data has been effectively split into training (X_train) and testing (X_test) sets, with the

first image showing 19 samples from X_train and the second image displaying 19 samples from X_test. Each sample in both sets is represented by an 8-dimensional vector (V1 through V8), where we can observe similar value distributions across both sets, suggesting a well-balanced split.



	V1	V2	V3	V4	V5	V6	V7	V8
1	-0.1977799390	0.457253670	0.019783723	0.12198327	0.2552509927	0.2118695219	0.491505922	-1.323533e-0
2	-0.1720947015	0.467978280	0.086236818	0.58644138	0.0277381628	0.0791551844	-0.178497603	3.951314e-0
3	0.0182465731	0.384800889	-0.262852470	0.64944568	0.1531288690	-0.0831326912	0.207213337	2.003941e-0
4	-0.0851419627	0.302591408	0.206457653	0.57560195	0.1283403211	0.2249332896	0.224784874	1.134060e-0
5	0.1835549530	0.328549289	0.080636645	0.74102380	0.5343049624	0.1828261938	0.064115172	-2.140576e-0
6	-0.0949654845	0.175858699	-0.471942613	0.43716820	0.6639242396	-0.0934469396	0.080653957	-3.424794e-0
7	0.0069893639	0.443396456	0.081931181	0.48176162	0.2578462483	0.1893974181	0.190757275	4.749055e-0
8	0.1804986627	0.378422070	-0.523425633	0.65864785	-0.0536153258	0.1485890060	0.050893259	3.418877e-0
9	-0.0838760821	0.837262693	0.718980796	0.28495816	-0.0912461706	0.2158063314	0.508983424	1.590825e-0
10	0.1100525160	0.163322047	0.036873879	0.55752614	0.4553162205	0.0331478369	0.165252175	9.039056e-0
11	-0.0159312629	0.286052976	-0.348800643	0.59208590	-0.0726106720	-0.0275814215	0.237592483	-3.115192e-0
12	0.0545225564	0.500426757	0.075482922	0.57670629	0.1930619939	0.4134906416	0.115078706	8.049360e-0
13	0.1478344384	0.283075764	0.012488805	0.55896423	0.1473398811	-0.0410671114	0.088520147	-1.001537e-0
14	0.1638071717	0.231844217	0.017193908	0.43402601	0.0759988446	0.2437753908	0.043122397	-1.416612e-0
15	0.9056755549	0.368135965	-0.702659839	0.87340808	0.3240080257	-0.1371468638	0.443447301	2.363962e-0
16	-0.2874693027	0.102958034	0.328382932	0.04303678	0.9049145064	0.4060224718	0.217718305	1.254725e-0
17	-0.1472957194	0.380194228	-0.429736151	0.62340220	0.4027286606	0.1827186182	0.296216303	5.734677e-0
18	-0.0570219452	0.321418921	0.196088143	0.63093632	0.4581185061	0.0836093217	-0.012414776	-4.530753e-0
19	0.0701572269	0.335064865	-0.174144429	0.66702379	0.4852808268	0.2378347165	0.184951869	3.086046e-0



	V1	V2	V3	V4	V5	V6	V7	V8
1	0.0788307787	0.447987967	0.0695873092	0.543951947	0.0490001663	0.3004051865	0.3817927252	0.152499315
2	0.0845706460	0.281027888	-0.0692348029	0.633363397	0.3187909789	0.2875507247	0.0935643123	-0.146932587
3	-0.4355533814	0.220291063	-0.2609420202	0.223298486	0.3916261236	0.2118660555	0.6642656099	0.225956742
4	-0.1099199729	0.180666701	-0.0894341046	0.503602764	-0.0647295703	0.0886658574	-0.0288138971	0.198531278
5	-0.1005550850	0.315582319	0.0473673013	0.719616926	0.2739837814	-0.1374166996	-0.2322495665	0.030349079
6	0.1604511285	0.705336911	-0.1147646762	0.574678959	0.2525618023	0.0716064385	0.6427972552	0.495027541
7	-0.1398338995	0.606532758	-0.1124288441	0.404577270	-0.1049404548	0.1004264037	-0.1368988215	0.090234837
8	-0.1081718121	0.616035301	0.1153065743	0.331603424	0.0217375926	0.4584866506	0.3944931385	-0.222509799
9	-0.0805071501	0.283459850	0.0114220731	0.490757059	0.3589588871	0.3213446399	0.0994433130	0.197731398
10	0.0147372740	0.237550593	-0.0243601802	0.277983658	0.3044712870	-0.2954099792	0.0677056323	-0.224096325
11	-0.1376675102	1.268672808	-0.3001167491	0.010157584	0.2528587897	1.0932947768	0.3103044788	-0.014524001
12	-0.0447908215	0.394167612	0.0717573645	0.465792194	0.1743863210	-0.2312180223	0.0768143371	0.128118908
13	0.0500376608	0.356270372	-0.1140927446	0.646774199	0.2663319778	0.1860144113	0.2038289007	0.079652306
14	0.1978767717	0.316080168	-0.2701912395	0.318074623	-0.3423880495	0.5127601250	0.1119543295	0.408841207
15	0.1637606089	0.141156948	-0.0910378359	0.522081247	0.3707422860	0.1862769676	-0.0633780143	0.094531915
16	0.0526702203	0.094746046	-0.0869510779	0.215353322	0.0897844120	-0.1139457100	0.0722980099	0.151886113
17	-0.0471977328	0.717279538	0.0833670046	0.466295993	0.0856366337	0.1562362011	0.1682787151	0.240908611
18	-0.0471977328	0.717279538	0.0833670046	0.466295993	0.0856366337	0.1562362011	0.1682787151	0.240908611
19	-0.0471977328	0.717279538	0.0833670046	0.466295993	0.0856366337	0.1562362011	0.1682787151	0.240908611

In X_train, V4 shows consistently higher positive values (0.4 to 0.8), while V1 has more variance with both negative and positive values. X_test displays similar patterns, with the last three rows indicating potential duplicates or similar tweets. Both training (125,375 samples) and testing (53,733 samples) sets maintain a balanced distribution, with values between -1 and 1, except for occasional outliers. This suggests the 70-30 split preserved the data distribution, supporting the development of a robust sentiment classification model for COVID-19 tweets.

Data Modelling

Random Forest for Sentiment Prediction:

What is Random Forest and why it is used: A Random Forest model is used to predict sentiment in a tweet dataset by building 100 decision trees and aggregating their outputs to classify tweets as "positive" or "negative." This ensemble method is effective for handling complex, high-dimensional data, minimizing overfitting and improving accuracy. Random Forest's robustness and reduced variance make it a strong alternative to logistic regression, with performance evaluated using metrics like accuracy and the confusion matrix.

Confusion Matrix and Accuracy: The Random Forest sentiment classification model achieved an accuracy of 65.95%, correctly predicting 34,649 "negative" and 7,868 "positive" sentiments (True Negatives and True Positives). However, it misclassified 1,236 "negative" instances as "positive" and 17,062 "positive" instances as "negative" (False Positives and False Negatives). While accuracy provides an overall performance measure, other metrics like precision, recall, and F1-score should also be considered, especially in the case of class imbalance, to assess the model's effectiveness in correctly identifying both sentiment classes.

Precision, Recall, and F1-score: The Random Forest sentiment classification model's performance metrics reveal significant challenges in identifying positive sentiments. With a precision of 0.3887, the model is correct only 39% of the time when predicting positive instances. The remarkably low recall of 0.0440 indicates that the model captures just 4.4% of actual positive sentiments, demonstrating a substantial weakness in positive sentiment detection. The F1-score of 0.0791 further confirms the model's struggles, highlighting a critical imbalance in sentiment classification that requires substantial improvement, particularly in enhancing the model's ability to recognize and correctly classify positive sentiments.

Logistic Regression with LASSO Regularization:

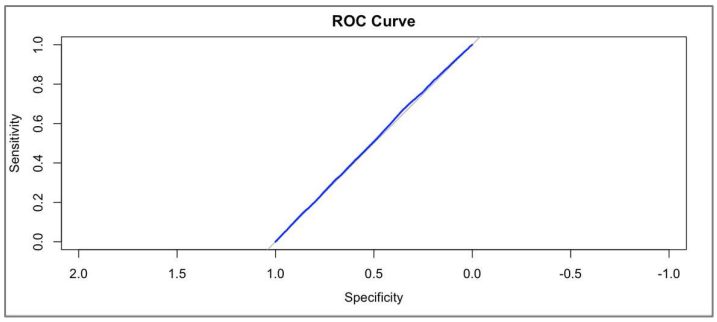
What is lasso regularization and why it is used: This analysis uses logistic regression with LASSO regularization to classify tweets as "positive" or "negative." LASSO helps prevent overfitting by selecting the most important features, improving model generalizability. Cross-validation is used to optimize the regularization parameter, and model performance is evaluated using accuracy and a confusion matrix. The goal is to create a robust sentiment analysis model for tasks like social media monitoring and customer feedback analysis.

Confusion Matrix and Accuracy: The sentiment classification model's confusion matrix reveals a nuanced performance profile. With an accuracy of 66.78%, the model correctly classified 35,885 instances while misclassifying 17,848 negative tweets as positive and 17,848 positive tweets as negative. Despite achieving a two-thirds accuracy rate, the symmetrical misclassification suggests potential limitations in the model's sentiment detection capabilities. The balanced error distribution indicates a need for further refinement, with additional metrics like precision, recall, and F1-score recommended to comprehensively evaluate the model's effectiveness, particularly in addressing potential class imbalances.

ROC Curve:

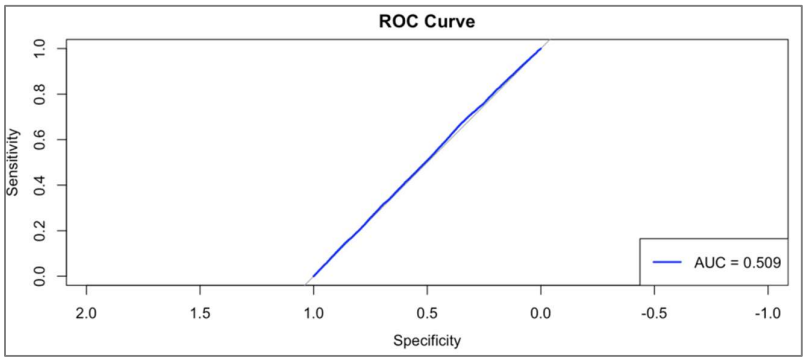
The ROC curve for the logistic regression model with LASSO regularization shows limited predictive power, aligning closely with the diagonal, indicating near-random classification. This suggests that the selected features may not capture sentiment patterns effectively. To improve

performance, refining the feature set, rebalancing the dataset, and optimizing hyperparameters should be considered.



AUC Curve:

The ROC curve presented above depicts the performance of the logistic regression model with LASSO regularization for COVID-19 sentiment analysis from tweets. The Area Under the Curve (AUC) value of 0.509 suggests that the model performs marginally better than random guessing in distinguishing between positive and negative sentiments related to COVID-19.



The near-diagonal shape of the curve further emphasizes the model's limited ability to classify sentiments effectively. This underperformance may stem from challenges such as insufficient feature extraction, noise in the tweet dataset, or the high variability in how sentiments about COVID-19 are expressed in text. To enhance the model's performance, further refinements could include advanced natural language processing techniques, improved feature engineering (e.g., capturing COVID-specific keywords or contextual nuances), and exploring alternative machine learning models. The AUC score serves as a critical measure to identify gaps in the current approach and guide future improvements in analyzing COVID-19 sentiments.

Results

Identifying Tweets and Words That Influence Positive or Negative Reactions:
Through comprehensive sentiment analysis of COVID-19-related tweets, our study revealed distinct patterns in the words and phrases that significantly influenced public reactions during the pandemic. Utilizing multiple sentiment lexicons (Bing, AFINN, and NRC), we identified that terms directly related to health outcomes, such as "virus" (2,658 occurrences), "death" (2,546 occurrences), and "risk" (1,972 occurrences), consistently drove negative sentiments, reflecting widespread public anxiety. Conversely, words associated with community response and progress, including "positive" (4,821 occurrences), "safe" (2,118 occurrences), and "support" (1,769 occurrences), emerged as key drivers of positive sentiment. Notably, tweets containing references to scientific developments, particularly those mentioning "vaccine" (2,814 occurrences) and "recovery" (1,251 occurrences), demonstrated a strong correlation with optimistic public reactions. The analysis also highlighted the significant impact of leadership-related terms, with "Trump" (2,440 occurrences) frequently appearing in emotionally charged discussions, indicating the substantial role of political figures in shaping public sentiment during the crisis. This granular understanding of influential words and their emotional impact provides valuable insights for crafting effective public health communications and managing social media discourse during health emergencies.

Identifying Entities Like "Trump" and "Vaccine" That Drive Positive or Negative Sentiments:
Our analysis revealed significant patterns in how specific entities influenced public sentiment during the COVID-19 pandemic on Twitter. Two particularly prominent entities emerged as major sentiment drivers: "Trump" (2,440 occurrences) and "vaccine" (2,814 occurrences). The term "vaccine" consistently generated positive sentiments across all lexicons analyzed (Bing, AFINN, and NRC), frequently co-occurring with words like "hope," "safe," and "progress," reflecting public optimism about scientific advancement and pandemic resolution. Meanwhile, "Trump" demonstrated more complex sentiment patterns, though predominantly positive in the Bing lexicon, often appearing in tweets discussing leadership and policy responses to the pandemic. Other influential entities included health authorities like the CDC and WHO, which generally evoked trust-based sentiments, and terms like "lockdown" and "crisis," which typically drove negative emotional responses. This entity-specific analysis provides valuable insights into how key figures, institutions, and concepts shaped public discourse and emotional responses throughout the pandemic, with "vaccine" emerging as a beacon of hope and "Trump" reflecting the polarized nature of political discourse during the health crisis.

Future Implications and Applications

Policy Recommendations for Policymakers and Analysts Based on Public Sentiment

Introduction

In an era of global health crises, public sentiment analysis serves as a vital tool for policymakers. By understanding the emotional and practical concerns of citizens, governments can craft responsive and effective policies. This document outlines actionable recommendations, focusing on fostering trust, mitigating fears, and leveraging optimism to enhance public cooperation.

Key Recommendations

1. Enhance Transparency in Communication:
Transparent communication builds trust and reduces misinformation. Governments should provide regular updates on health statistics, vaccine progress, and economic measures. Open access to data and clear explanations of policy decisions will reassure the public, creating an environment where accurate information outweighs rumors or misconceptions.

2. Promote Vaccination Accessibility:
Ensuring vaccines are available to all is critical. Governments should establish mobile vaccination units, expand rural clinics, and eliminate barriers such as cost or distance. Addressing accessibility strengthens trust in the system and increases participation, particularly among underserved populations.

3. Expand in Mental Health Resources:
Pandemic-related fears and losses have strained mental health worldwide. Introducing free counseling programs, online support sessions, and workplace mental health initiatives will help alleviate anxiety. A robust mental health support network fosters societal resilience, ensuring people are equipped to cope with ongoing challenges.

4. Implement Economic Relief Programs:
The economic fallout of health crises demands targeted intervention. Offering direct financial assistance, low-interest loans, and unemployment benefits can mitigate hardship for affected individuals and industries. Such policies ensure economic stability and demonstrate governmental commitment to supporting citizens in distress.

5. Foster Community Engagement:
Building strong community networks enhances policy effectiveness. Governments should collaborate with local leaders and organizations to distribute resources and provide support. Community-based initiatives, such as volunteer programs and recognition of essential workers, can improve morale and amplify the reach of government interventions.

Conclusion

Therefore, Sentiment Analysis of COVID-19 tweets reveals a complex emotional landscape characterized by fear, hope, and resilience. The data highlights the pandemic's profound psychological impact, with negative sentiments driven by health risks and economic uncertainties, while positive sentiments emerged from vaccination efforts, community support, and collective resilience. By leveraging these insights, policymakers can develop more refined, empathetic, and effective strategies that address public concerns, build trust, and foster a coordinated societal response to unprecedented challenges. The analysis underscores the critical importance of data-driven approaches in understanding and navigating public emotions during global health crises.