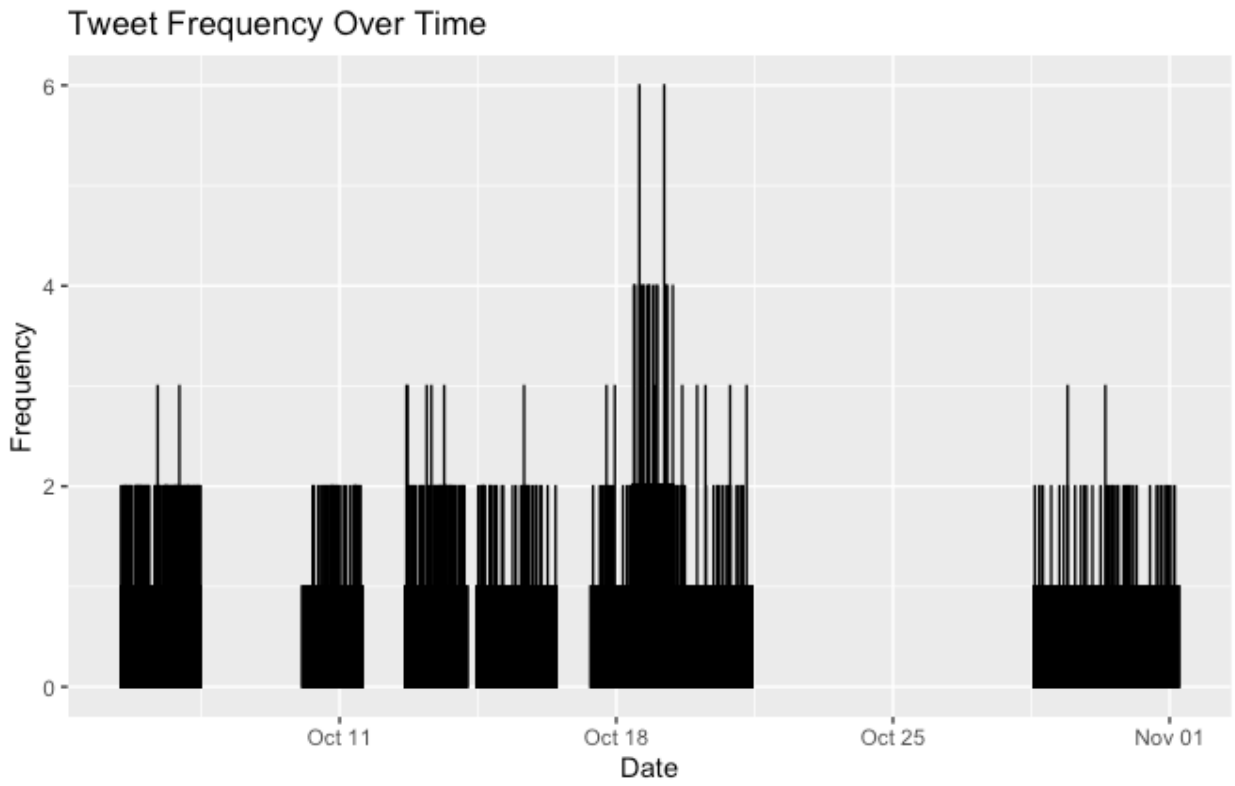
Introduction

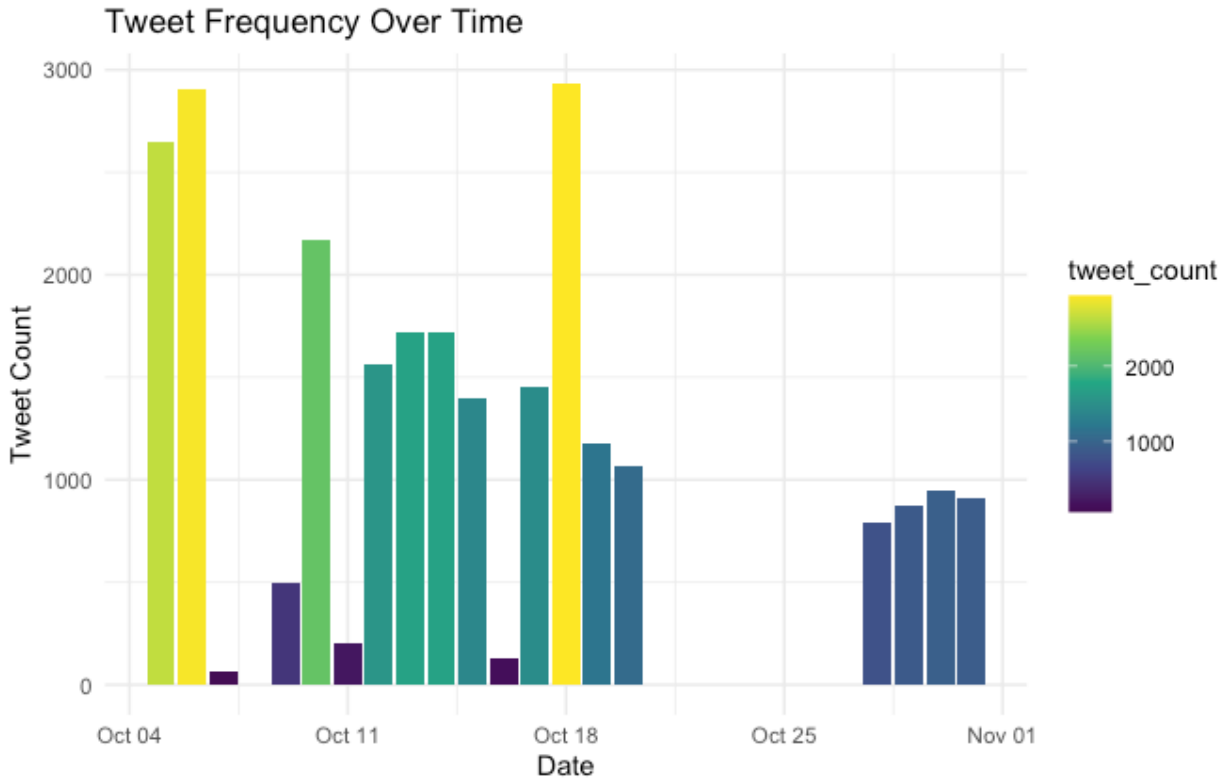
The data has been taken from (CONTRACTOR, D., 2022. *Kaggle.* [Online] Available at: https://www.kaggle.com/datasets/deepcontractor/squid-game-netflix-twitter-data) . This original data has more than 80000 tweets and after cleaning the data, it has only 225187 obs. Of 13 variables which is shown as name: cleaned_tweets. The data cleaning process is shows as bellow:

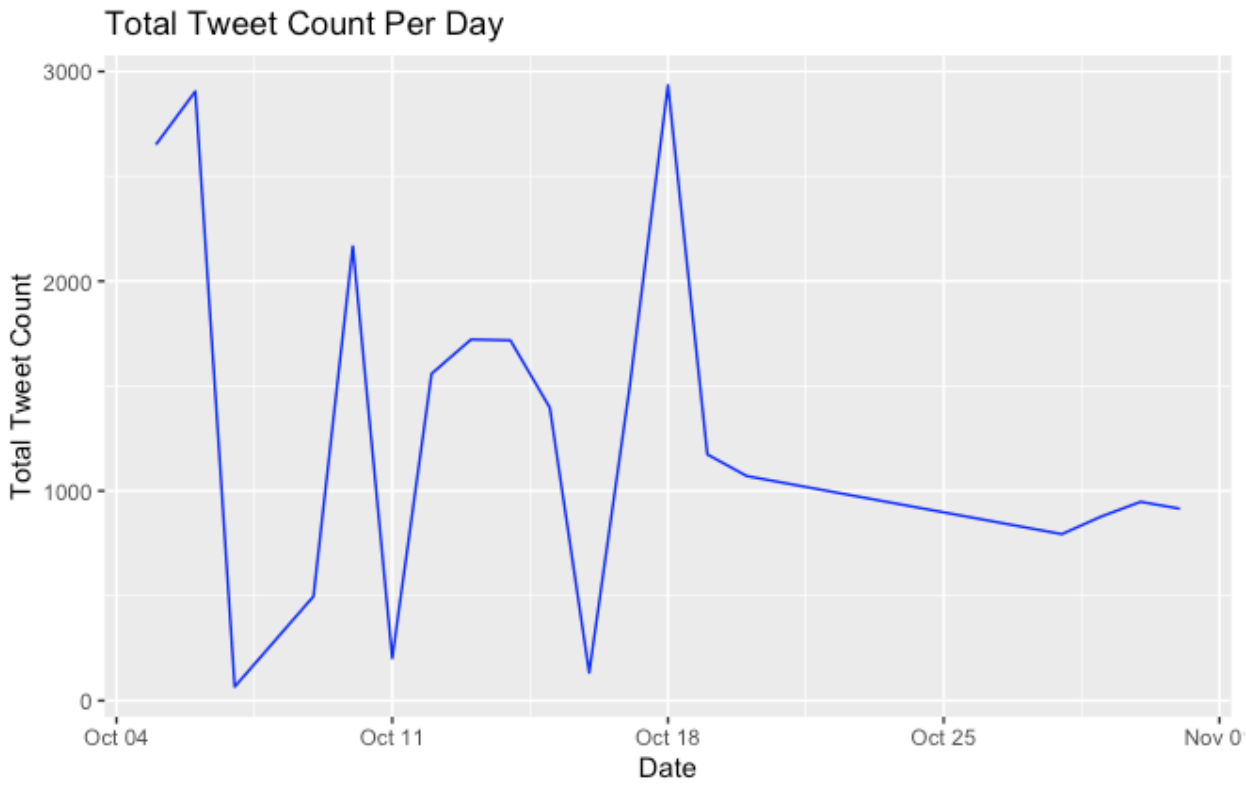
```
>  View(tweets_squidgame)
> library(dplyr)
library(ggplot2)
> tweets_squidgame$text[tweets_squidgame$text == ""] <- NA
> tweets_squidgame$source[tweets_squidgame$source == ""] <- NA
> tweets_squidgame$text[tweets_squidgame$text == ""] <- NA
>
> tweets_squidgame$user_location[tweets_squidgame$user_location == ""] <- NA
> tweets_squidgame$user_description[tweets_squidgame$user_description == ""] <- NA
> cleaned_tweets<-na.omit(tweets_squidgame)
> cleaned_tweets <- na.omit(tweets_squidgame)
```

Now on, as a cleaned_tweets data, this assessment will review the dataset and visualise the frequency of tweets, identify top tweeting locations and present the first 10 and identify the most retweeted tweets (top 20).

visualise the frequency of tweets



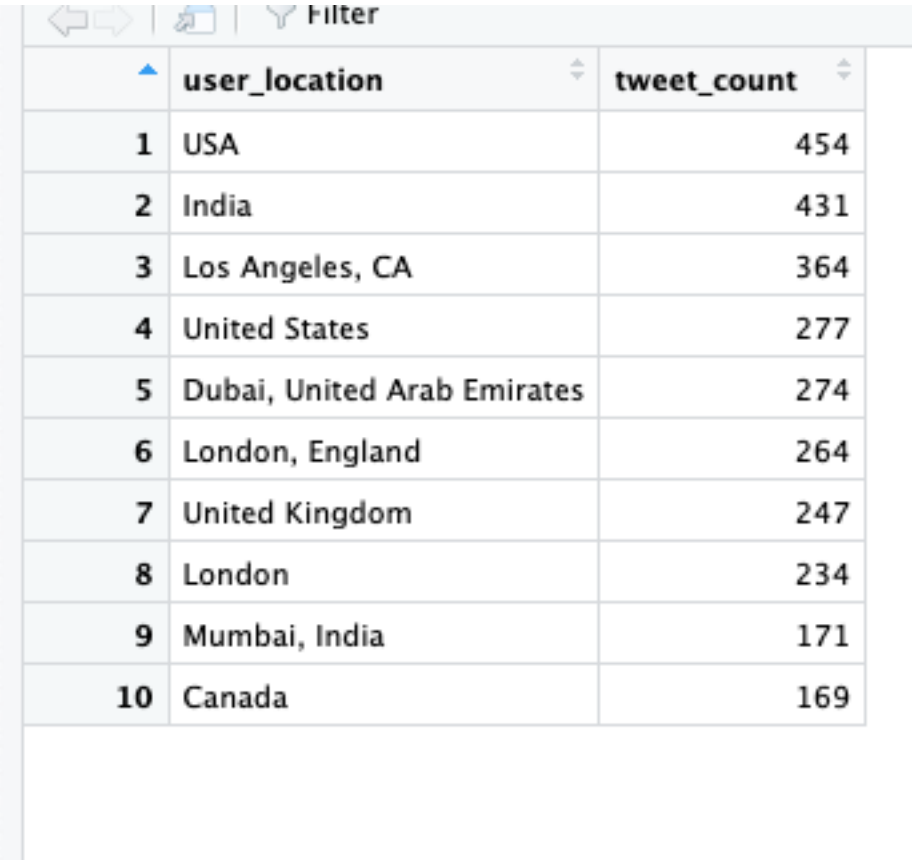




These 3 different graphs illustrate the count frequency over time and per day of tweet whereas, data spans October 5 to October 31 with low as 66 and high as 2934 which happened October 7 and October 18 respectively. The 2nd graph also represents by the color whereas more tweets count presented by yellow and less tweets counts represented by dark blue. Overall, this numbers provide an overview of the tweet activity over the given time period.

Top tweeting locations and present the first 10

The following table represent the top tweeting location:



	user_location	tweet_count
1	USA	454
2	India	431
3	Los Angeles, CA	364
4	United States	277
5	Dubai, United Arab Emirates	274
6	London, England	264
7	United Kingdom	247
8	London	234
9	Mumbai, India	171
10	Canada	169

The above table represent the USA is the most popular location. whereas, 453 tweets posted from the USA and Indian, Los Angeles, CA, United States and more continues in the same format as location and tweet count. Overall, this table provides insight into where is the top location of sources in the data and which information can be useful for where are the most active user are from and their contribution.

The most retweeted tweets (top 20)

The following table shows the most retweeted tweets (top 20) :

	user_name	text	is_retweet
1	YoMo.Mdp	//Blood on 1st slide I'm joining the squidgame thing, ...	False
2	Peyman 🅚🅐🅘	$THG Going to explode to 4B Marketcap very soon. T...	False
3	Kevin Franco	I discovered English audio after 5 episodes of Squid G...	False
4	soph 🍁	— squid game (2021): deaths and foreshadowing 🦑 ...	False
5	Yemzxy	The saga continues Get data cheap @honourworldng...	False
6	Geriatrixx	Dunno what everyone's going crazy about #SquidGam...	False
7	Monkey Magic	#SquidGame In the tug of war game, everyone was ch...	False
8	Anne LaVampyre	I'm really thrilled for Korea. If you would have told me...	False
9	𝑀𝑠.𝒴𝑜𝑜𝓃	let's str3am your fav idol songs with apps premium b...	False
10	Reinhard Omukuba	Cyber security "experts" be like: "I'm a cyber security ...	False
11	Charlie	I Didn't mean to push her off - Squid Game Roblox ht...	False
12	J ✨ new art 🎨	"Ladies and gentlemen... Her" Here's a chibi art of Ka...	False
13	DHL Express Malaysia	Can you win this game?😎 Challenge accepted! #Squi...	False
14	SILCREEK REALTORS	Thank you for giving us the opportunity to serve you ...	False
15	holdmyjilabi	My parents finished #SquidGame in one sitting witho...	False
16	King David Tobi	Different Shows, Same Energy 🌈 #AliceInBorderland ...	False
17	JustSomeGoth	SQUIDGAME CHIBI YCH OPEN! Unshaded - £20 Shade...	False
18	iDiski Times	So Kaizer Chiefs and Orlando Pirates fans, who would...	False
19	Fãs de Keanu Reeves	This is '#SquidGame' would look if it starred #KeanuR...	False
20	제로 ◡̈	#SquidGame actress #JungHoyeon is less than 100k a...	False

The above table username shows the display name of twitter account and text shows the content, mention and hashtags (#Squadgames) and retweeted represent the wheatear the tweet has been retweeted by other (True or False). The above table helps to understand the which tweets in the dataset are more retweeted frequently and which can give insights into popular topics or engaging content.

Conclusion

In conclusion, by employing potent visualization libraries in R, it converted raw data into captivating visual representations. In an accessible manner, our visualization permitted the communication of complex data patter, with tweet frequency being presented through bar charts and trend over time being depicted through line plots. This not only improved our understanding but also provided a means to effectively share insights with other. By identifying top tweeting locations and the most retweeted content, it helps to understand deeper with user preferences and engagement trends which will help to contextualize business data and develop ML and AI solution using R to improve business decision.

Most common positive and negative words

From the Squad game data analysis, by using most common positive and negative words will provide textual data with customer review, social media post which will be helpful to illustrates the sentiment analysis. The following table provides the most common use of positive words.

Positive words

	word <chr>	n <int>
1	like	1326
2	good	751
3	great	418
4	best	389
5	love	388
6	happy	332
7	right	305
8	well	289
9	better	253
10	fun	225

1–10 of 10 rows

The above table represent the like is the most common word with 11326 times, good appeared 751 times and so on.

Negative words

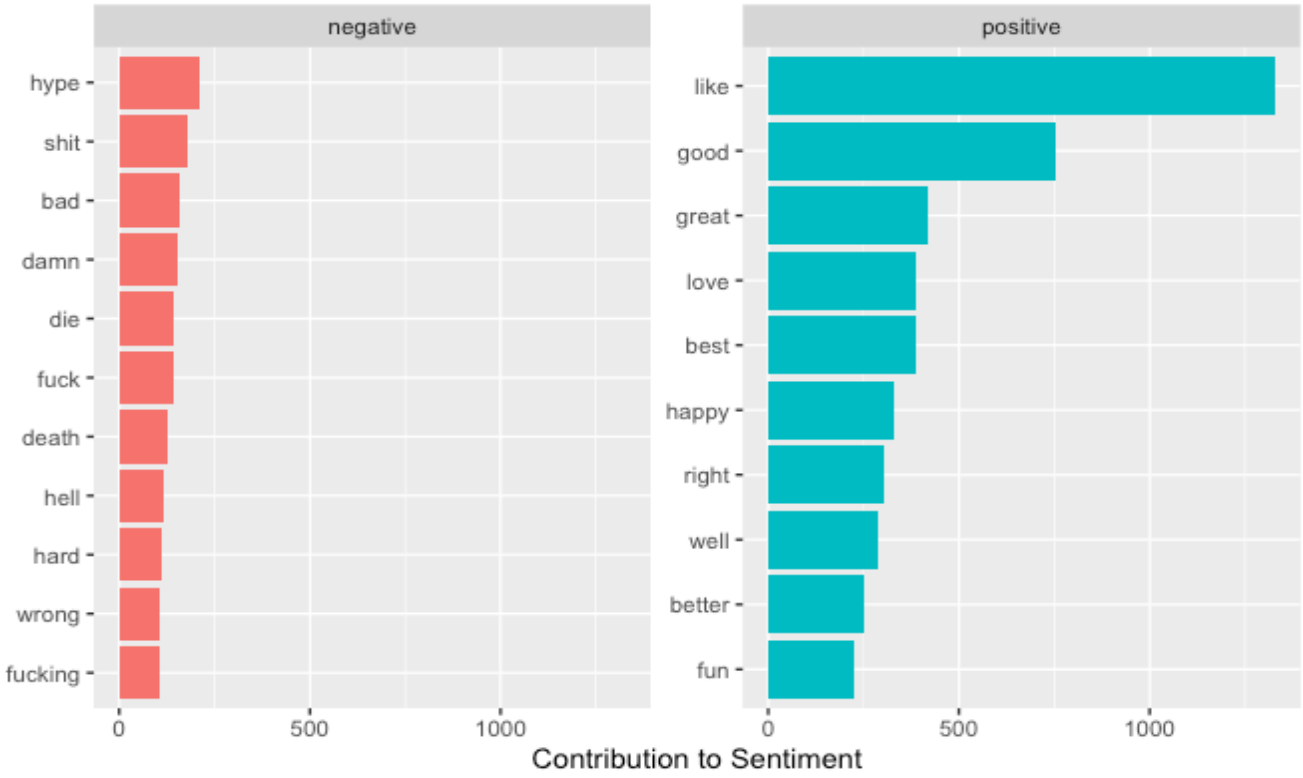
Description: df [10 × 2]

	word <chr>	n <int>
1	hype	211
2	shit	177
3	bad	158
4	damn	152
5	die	142
6	fuck	140
7	death	127
8	hell	115
9	hard	110
10	fucking	107

1–10 of 10 rows

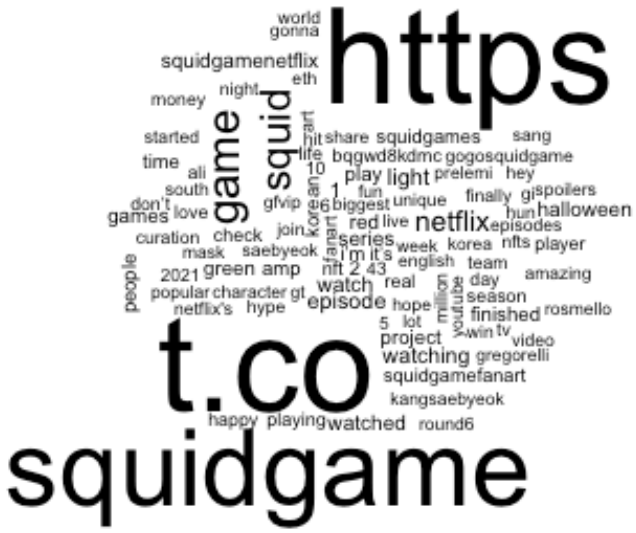
The above table shows the indicative of negative sentiments. Whereas, hype is the height time appeared word with 211 times. Shit has been used 177 times and so on.

Contribution to sentiment



The above graph illustrates the contribution to sentiment which shows the text shaping whether the overall tone is positive or negative. Overall, positive words are more using to express. This chart shows the like, good, great and so on has been contributing positively to the sentiment. On the other hand, negative words such as hype, shit, bad and so on carrying a negative connotation.

Word Clouds



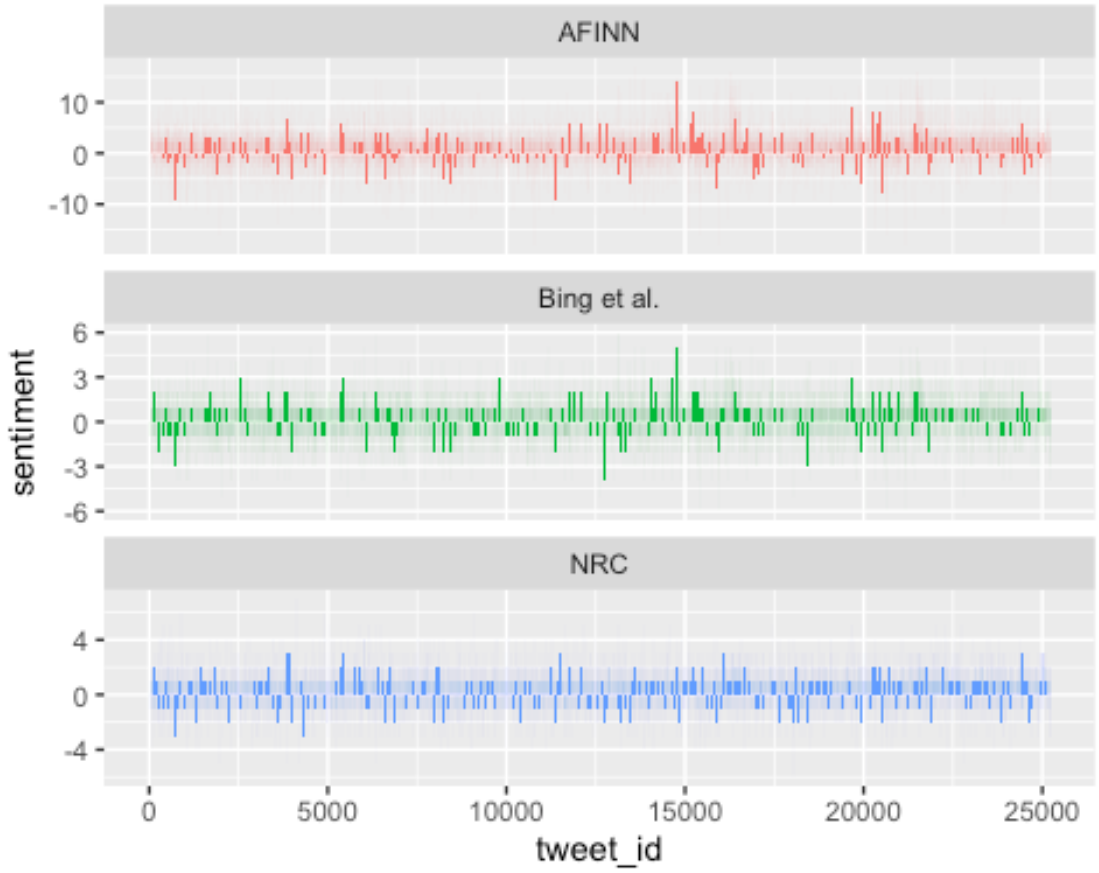
The above picture represents the word clouds of squid game tweets. It shows the visual representation of text data where the size of each word is proportional to its frequency. Https and squid game are the more frequent word with larger displayed. Game, squid, Netflix, episode, finished, green, 2021 are noticeable which means they are using frequently.

Comparison Cloud Function

The above cloud picture represents the comparison function between positive and negative words. Negative words shown on the top and positive words shown in the bottom with light font. This helps for spotting change in sentiment. As there are negative and positive topics, it can better understand the vocabulary associated with each topic and make informed content or communication decision.

Three Sentiment Lexicons Analysis

The sentiment lexicons analysis mainly uses for evaluation the opinion and expression in text. The major lexicons are AFINN, Bing and NRC. The below chart shows the all three sentiment lexicons and illustrates the sentiment changes across the tweet's words.



The above chart shows the three sentiment lexicons analysis output of squads_game tweets dataset, using three distinct sentiment lexicons: AFINN, Bing and NRC. Which are represented color by red, green and blue respectively. X-axis represent the sentiment and y-axis represent tweet _id index. The AFINN lexicon have presented the largest absolute values on positive values. The Bing has lower values whereas, NRC shifted vast on higher comparing to other two.

The following table shows more clearly:

Lexicon	Sentiment	Count
AFINN	Positive	9329
AFINN	Negative	4810
AFINN	Neutral	635
NRC	Positive	9132

Lexicon	Sentiment	Count
NRC	Negative	4496
NRC	Neutral	2036
Bing	Positive	8006
Bing	Negative	4586
Bing	Neutral	1684

The above table indicates the specific sentiment analysis of three lexicons. Whereas, sentiment is categorized on positive, negative and neutral. The count column represents the times of each sentiment categories. In AFINN lexicon, 9329 is the positive identified and similarly, NRC and Bing have 9132 and 8006 respectively. Overall, this helps to illustrates the how each lexicon interprets the sentiment mentioned in the text data.

Comparing three Lexicons

After conducting with chart and table of three sentiment analysis: AFINN, Bing an NRC lexicons output, a clear differentiation in sentiment analysis and scoring emerged. It shows that AFINN and Bing presented relatively similar tendencies while NRC incorporation of shows in more depth which is shown above chart and table as well. However, AFINN is quite easy to use can be more useful for basic sentiment analysis. Similarly, Bing can be useful for analysis of clear positive and negative sentiment expression. NRC lexicons shows the emotion with different sentiment and it represent the in-depth analysis.

On the basis of above provided details according to our tweet data set accuracy, coverage context, NRC could be the better choice for analysis to make right decision and to understand in-depth and it has emotional dimension of the text that would easy to understand with better presentation. As we have tweets dataset, tweets have expression of emotional and subject content and emojis and informal language which can help to distinguish between positive and negative expression of sentiment.

Conclusion

In conclusion, all three lexicons are very helpful in the journey to decode sentiment analysis of squads_game tweets dataset. Whereas, AFINN and Bing presented multifaceted nature of sentiment analysis and NRC shows in-depth emotional dimension. So, selecting these three can be depend on nature of data set and purpose of data. And for this dataset, NRC can be more preferable in terms on Emotion and expression tweets.