|  |
| --- |
|  |
| Final Report 7230ICT |
| Big Data Analytics |

|  |
| --- |
| Rey Allen Permale  10-4-2024 |

Table of Contents

**Setting1**

1. Artist Information2

**Data Selection and Exploration2**

1. Collection of Data2
2. Top 5 Actors3
3. Unique Actors4
4. Spotify Data5

**Text Pre-processing9**

1. Term-document matrices9
2. Semantic Network10

**Social Network Analysis12**

1. Centrality Analysis12
2. Community Analysis17

**Machine Learning Models23**

1. Sentiment Analysis23
2. Decision Tree24
3. Topic Modelling26

**Visualisation28**

1. Dashboard28

**Analysis Review29**

1. Alternative methods29

**Setting**

1. Artist Information – Taylor Swift

Taylor Swift start as a country music singer who is in the music industry for 18 years now. She began her career in 2006 with her first album called “Taylor Swift”. Since then she evolved into one of the most prominent figures in music, transitioning from country to pop. Currently, she has been experimenting with indie and alternative genres.

Throughout her career, Taylor Swift has released 11 studio albums ang 4 re-recorded albums (Forbes, 2024). The total number of tracks that she published a total of 274 tracks after the addition of her latest album (Newsweek, 2024).

*References*

Toni FItzgerald. (2024). *Taylor Swift albums*. Forbes. <https://www.forbes.com/sites/entertainment/article/taylor-swift-albums/>

Sophie Lloyd. (2024). *How Many Albums Does Taylor Swift Have? Timeline of Song Releases. Newsweek*. [https://www.newsweek.com/taylor-swift-how-many-albums-songs-tortured-poets-anthology-1892216 - :~:text=On%20Friday%20morning%2C%20Swift%20announced,number%20of%20tracks%20to%20274.](https://www.newsweek.com/taylor-swift-how-many-albums-songs-tortured-poets-anthology-1892216#:~:text=On%20Friday%20morning%2C%20Swift%20announced,number%20of%20tracks%20to%20274.)

**Data Selection and Exploration**

1. Collection of Data

Reddit was chosen as the platform for data collection as the sentiments of the redditors are more related to the thread that was chosen compared to the comment sections in Youtube in the search strategy that was applied. The search strategy only chose the top threads related to the most recent 4 albums released by Taylor Swift including the re-released versions. It should also be noted that showbiz related feud and political thread was ignored to only focus on the improvement of the musical aspect of the artist. They key search words that gave the desired results and the chosen links are:

Search Words:

* Taylor Swift Latest Album
* Taylor Swift Midnights
* Taylor Swift Version
* Taylor Swift Album Megathread

Links:

* <https://www.reddit.com/r/TaylorSwift/comments/y9ivjo/midnights_megathread/>
* <https://www.reddit.com/r/TaylorSwift/comments/17he0ir/1989_taylors_version_megathread/>
* <https://www.reddit.com/r/TaylorSwift/comments/1aj5rqq/the_tortured_poets_department_album_announcement/>
* <https://www.reddit.com/r/TaylorSwift/comments/139b5nl/speak_now_taylors_version_announcement_megathread/>

The number of data points collected are in total of 3366.

*Figure 1. Data Points*

1. Top 5 Influential Actors

A close up of a number

Description automatically generatedThe dataset was removed of any rows that contains a null attribute and was then created into a network graph resulting into these 5 influential actors. The result also shows that people are more interactive to the comments of aran130711, PassionateAsSin and Lyd\_Euh. Furthermore, there is a tendency that [deleted] is a false result as the activities of multiple deleted accounts in reddit can fall into the category of the [deleted] account.

*Figure 2. Top 5 influential actors*

A close-up of a network

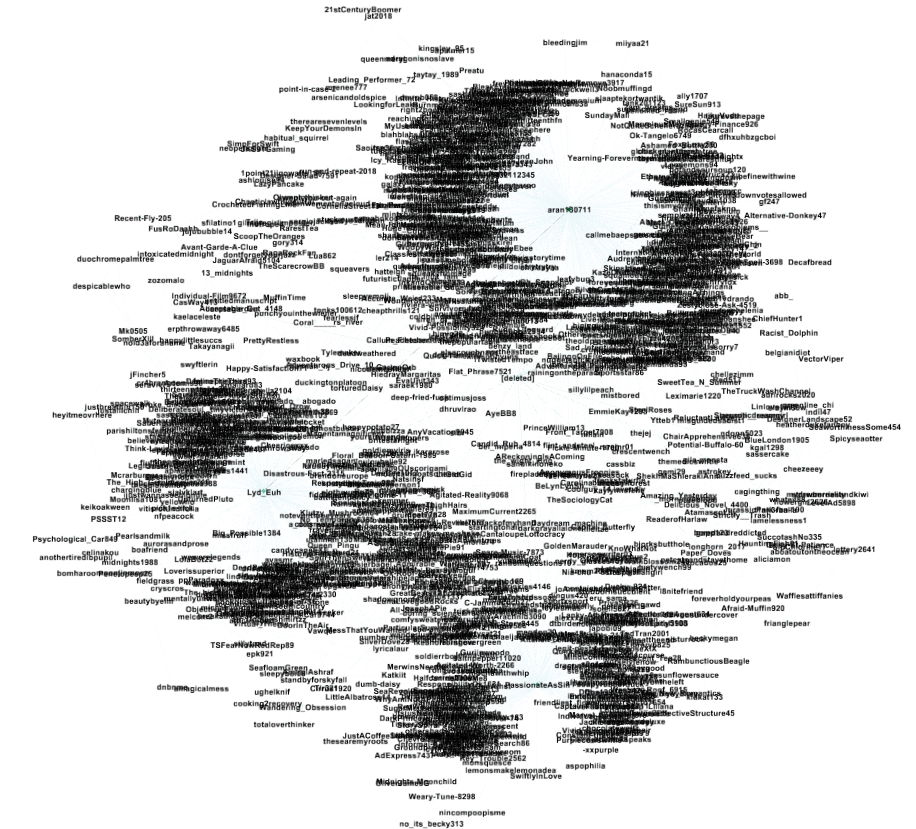
Description automatically generated

*Figure 3. Actor Network Graph visualization*

The reason there are only 3 prominent figures in the visualization is because the gap between the pagerank of the top 3 results compared to the top 4 and top 5.

1. Unique Actors in the dataset.

There are around 1481 unique actors in the actor graph and was obtained by using the length function of R. This will return the number of unique actors which was converted into row in the actor network graph.



*Figure 4. Unique Actors Network Graph visualization*

1. Spotify Data

Active Years

The data collection in spotify for Taylor Swift showed that spotify has the oldest album released in 2014 where as in the searched in part 1 there are albums dating older than 2014. This is probably due to an issue back in 2014 which is not related to the case study.

A screenshot of a music album

Description automatically generated

*Figure 5. Taylor Swift albums in spotify*

Albums and Songs Published

The data collection showed a result of 20 albums released in Spotify and 359 songs published by Taylor Swift. These figures were obtained by searching for albums related to Taylor Swift and looping through each album to create a data set of all\_tracks.

A computer screen shot of a program

Description automatically generated

*Figure 6a. album and tracks code*



*Figure 6n. album and tracks count*

The underlying reason for this result is that there are albums which have different versions such as the folklore having a deluxe edition.



*Figure 7. folklore album with different versions.*

Collaborators

Collaborators in the Spotify data were obtained by unnesting the artist column of each album. The resulting artist was then filtered to remove the “Taylor Swift”artist\_name to only display the list of artists that Taylor Swift did a collaboration with. Figure 8a shows the code to unnest and filter the collaborators and figure 8b shows the list of collaborators.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated*Figure 8a. extracting collaborators*

*Figure 8b. List of Collaborators*

*Figure 8c. Collaborators network with track\_id A green oval with white text

Description automatically generated*

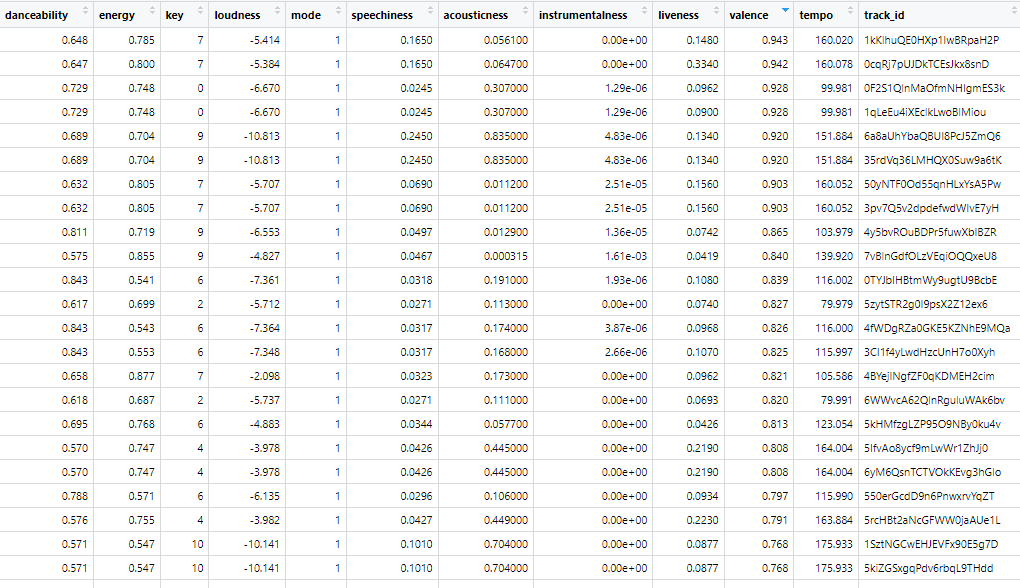
Prevalent Features

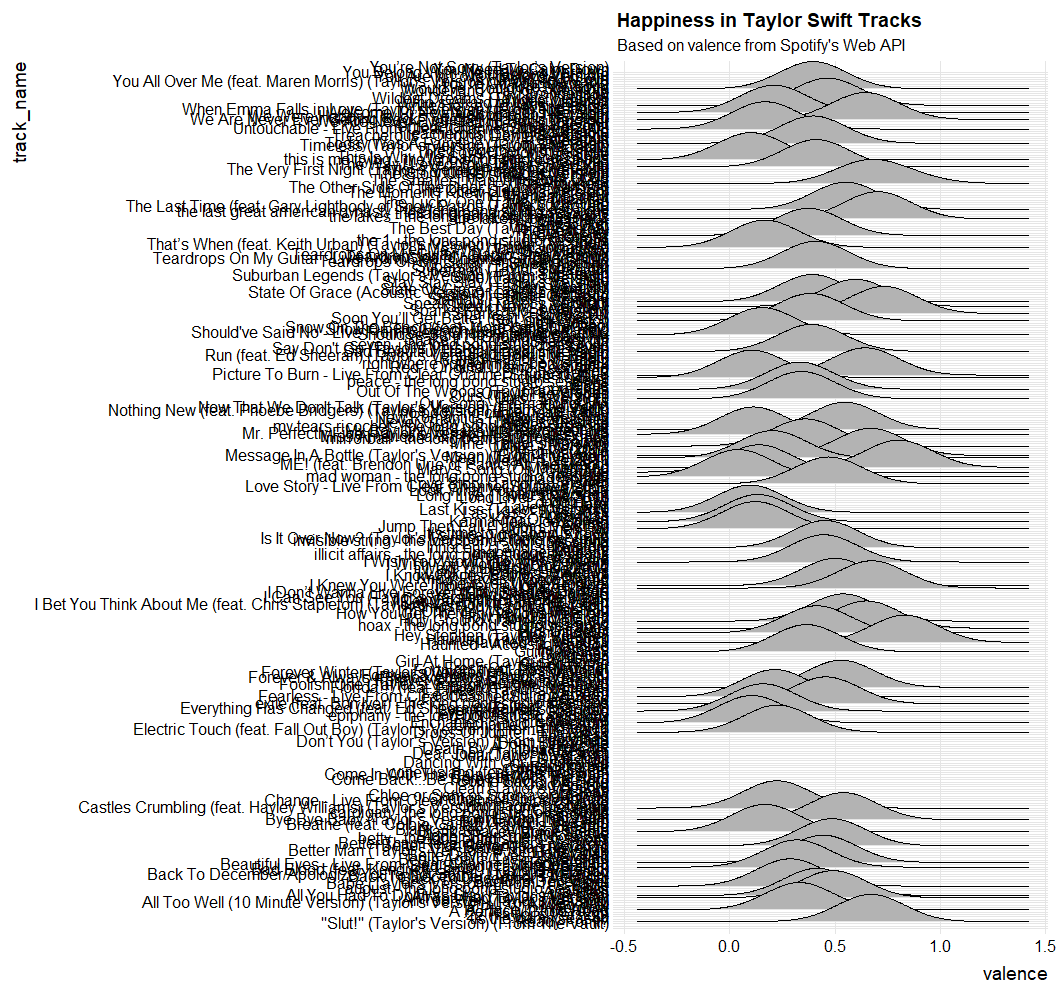
The noticeable features in Taylor Swift’s tracks are the valence and danceability. These attributes can be compared as they behave in a proportional manner where if the valence is high (which means the song is positive), the more it is danceable.

A computer code with black text

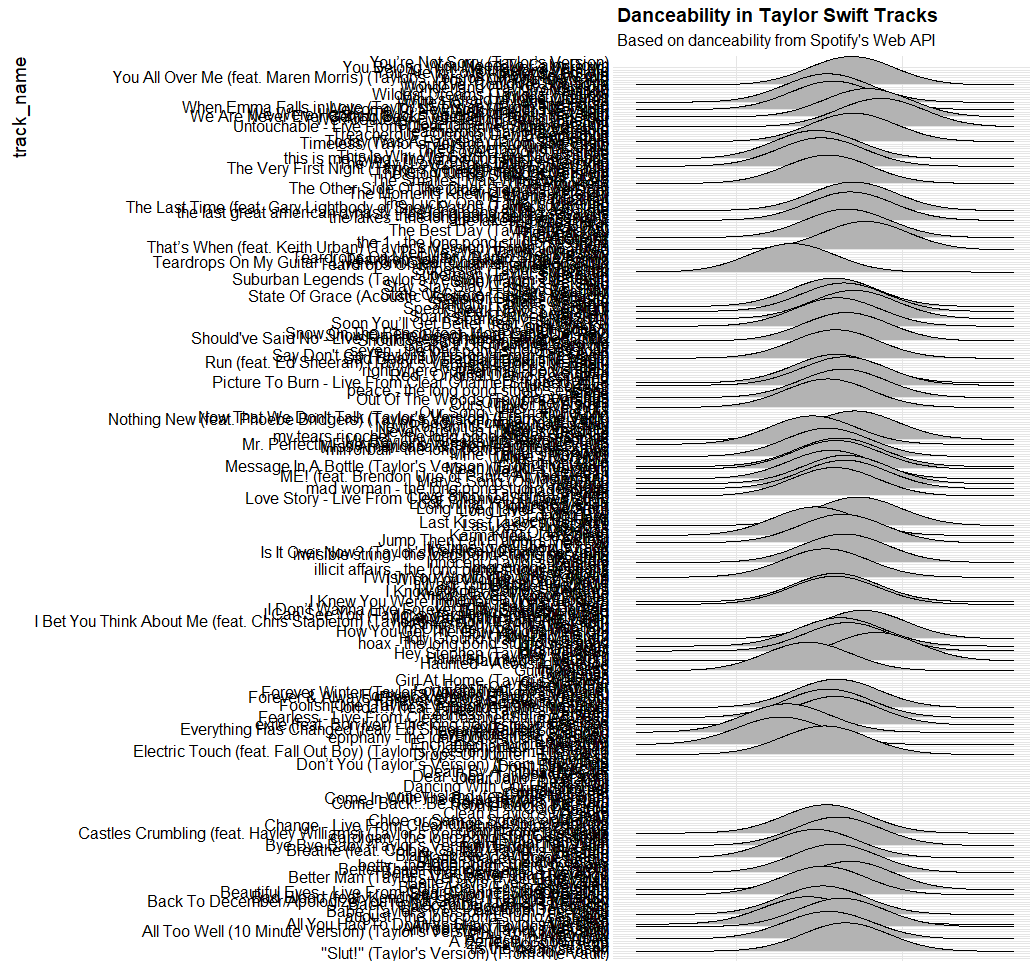
Description automatically generatedFrom the results in these attributes, it can also be inferred that the lesser valued tracks in terms of valence and danceability are those songs that relates to heartbreak or nostalgia. Taylor Swift’s evolution to pop music can also be observed as some of her songs has a high danceability and valence.

*Figure 9a. audio features code*



*Figure 9b. audio features result*

*Figure 9c. valence per track graph*

*Figure 9d. danceability per track graph*

**Text Pre-processing**

1. Term-Document Matrix

The top 10 meaningful terms were achieved by removing unnecessary terms, converting each word to stem words, and removing stop words. The results are as follows:

Album – The highest reoccurring word; High probability because the threads search was about taylor swift’s most recent album.

Song – One of the main items expected to find in a data collection for a singer. The main target is the sentiment of the users for the songs of Taylor Swift

Taylor – The first name of the artist.

Love – These term could either people use the term “love” to describe their sentiment to the tracks or is about how Taylor Swift’s music is mostly about love songs.

Feel – This is expected in a love song. This could be part of the statement where people describe how they are feeling towards the album

Sound – By product of the song, or is used by the users to describe the song sounding like something familiar to them.

Listen – The verb users do in order to feel something from the track.

Track – This is another word for song or recording.

Version – Two of Taylor Swift’s latest album are re-released version; hence versions of the album was discussed.

Time – This term could be related to how to users feel after Taylor’s re-release of the album which happened years before.

A computer code with text

Description automatically generated

*Figure 10a. Top terms used*

A graph of a number of blue bars

Description automatically generated

*Figure 10b. Visualization for top 10 frequent words*

1. Semantic Networks

A close-up of a computer code

Description automatically generatedBy obtaining the bigrams or the consecutive pair of words often appearing are obtained by using the functions in R. It was then sorted according to page rank which returned a result of the top 10 important words from the thread related to Taylor Swift. Those word are as follows:

*Figure 11a. Top 10 most important words according to page rank*

It could be observed that the top term “album” has more then double the occurrence compared to the rest of the terms in the top 10 result. This could be due to the main topic of the subreddit to be the most recent album’s of Taylor Swift. There are 2 results that hold the same meaning such as taylor and taylors where taylor could be when the users are talking about the artist “Taylor Swift” while taylors could be the stem word for the terms taylor’s version, as there are two albums related to the thread that are taylor’s version release. The rest of the terms are music entertainment related and genre related such as pop, tv, songs and favorite. The last term is the profanity which is “fucking” which are also A network of dots and lines

Description automatically generatedoften used by the users as an expression.

*Figure 11b. Top 10 most important words visualization*

**Social Network Analysis**

1. Centrality Analysis

The centrality analysis were performed to different datasets belonging to the main artists and two related artists.

1. Taylor Swift

Centrality analysis was performed to the actor network graph previously created to show the scores and determine if it is relevant to the artist. The related artist are obtained by the related artist function of the spotifyr.

Degree Centrality – The degree centrality measures the number of direct connections a user has. In this case, it would show the number of replies or comment the user has given out or received. The first sorting shows the aran130711’s thread received the most replies whereas the [deleted] user comment the most. This may also be a false data, as the score [deleted] user received where an accumulation of the multiple deleted accounts that commented in the thread. The total sorted shows the total (in and out) activity the user has done or received.

A computer screen shot of a code

Description automatically generated

*Figure 12a. Sorted degree centrality results*

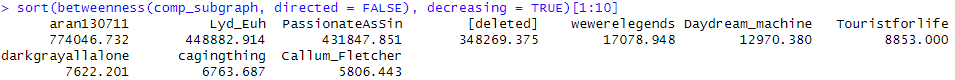
Closeness Centrality – This measure how each of the users are close to the rest of the discussion. This could mean that user may have reached out to more user via the reply they posted. The values shown in the results shows that most users have the same closeness score which means most users can be reached by other users on an equal level and vice versa.

A computer screen shot of a code

Description automatically generated

*Figure 12b. Sorted closeness centrality results*

Betweenness Centrality – This would indicate that these users posts are the most often nodes the appear on the shortest paths between other nodes. The result is also a bit reasonable as the degree centrality of aran130711 was the highest, hence this user’s reply has the most interaction in the thread. Most of the results in the betweenness could be seen in the previous centrality analysis done.



*Figure 12c. Sorted betweenness centrality results*

*A network of lines and dots

Description automatically generated*Relevance of Results – The relevance of the centrality analysis for Taylor Swift is through the revelation of how the users contribute to the topic of her albums and activities. It can also be shown that some users which in this case are fans of Taylor Swift, are relevant to increase her popularity by engaging with fellow users especially in the digital and information age we are in. The users with high centrality tend to drive the narrative that revolves around Taylor Swift and build different discussions which may in turn affect the sentiments other fans have for Taylo Swift.

*Figure 12d. Actor Graph Network Taylor Swift*

1. Olivia Rodrigo Centrality Scores

The component size of Olivia Rodrigo’s search threads are up to 659. This would mean that there would be a gap in the centrality scores when comparing Olivia Rodrigo’s to Taylor Swift’s.

A computer code on a white background

Description automatically generated

*Figure 12e. Component Size Olivia Rodrigo*

Degree Centrality – The fairly noticeable as only one person in this subreddit received a reply of more than a 100, the rest only falls up to a range of 50. This would mean that the subreddit was mostly affected by a single significant person which is “NominalPerson”.

A computer screen with text

Description automatically generated

*Figure 12f. Sorted degree centrality results Olivia Rodrigo*

Closeness Centrality – The result is still somewhat similar compared to Taylor Swift’s result. This is because how the thread system is structured where it is easy for people to connect to another hence the closeness centrality of all users are fairly even.

A computer screen shot of a code

Description automatically generated

*Figure 12g. Sorted closeness centrality results Olivia Rodrigo*

A close-up of a computer screen

Description automatically generated Betweenness Centrality – With a huge difference in community size, it can be observed in the betweenness centrality results. This is because since there are fewer edges, there is a direct proportionality to the difference in number of routes.

*Figure 12h. Sorted betweenness centrality results Olivia Rodrigo*

A close up of a firework

Description automatically generated

the difference in number of routes.

*Figure 12i. actor network Olivia Rodrigo*

1. Ariana Grande Centrality Scores

A computer code on a white background

Description automatically generatedHaving a component size of only 160, it can be seen the drastic difference between the centrality scores of the actor graph in Taylor Swift’s centrality analysis compared to Ariana Grande’s analysis.

*Figure 12j. Component Size Ariana Grande*

A computer screen shot of a program

Description automatically generatedDegree Centrality – The difference in degree centrality compared to Taylor Swift is rather noticeable. Although, it can also be observed that the distribution among the activities per user is more even compared to Taylor Swift’s where only 4-5 users received more than 10 replies where as here, it can be seen that all of the top 10 users received replies of more than 15.

*Figure 12k. Sorted degree centrality results Ariana Grande*

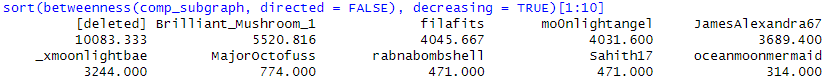
Closeness Centrality – The result is similar with Taylor Swift’s results. This could mean that the structure of the thread system of Reddit also plays a role on how the users are close to other users.

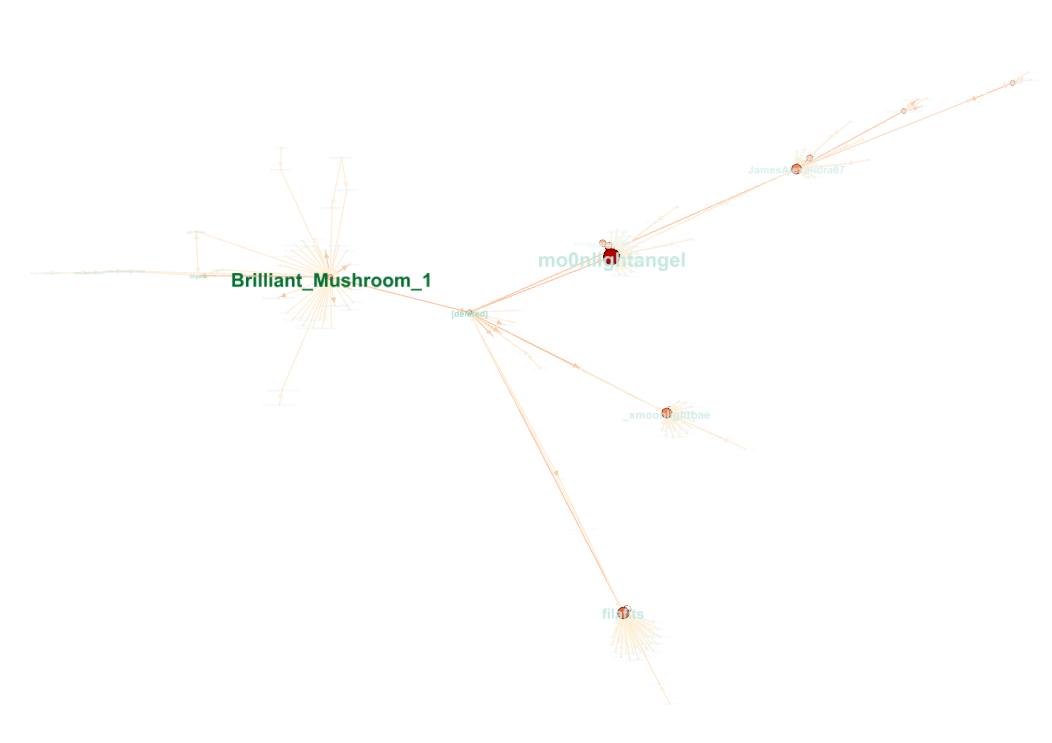
A computer screen with text

Description automatically generated

*Figure 12l. Sorted closeness centrality results Ariana Grande*

Betweenness Centrality – With a huge difference in community size, it can be observed in the betweenness centrality results. This is because since there are fewer edges, there is a direct proportionality to



*Figure 12m. Sorted degree centrality results Ariana Grande*

*Figure 12n. actor network Ariana Grande*

1. Girvan-Newman and Louvain methods
2. Taylor Swift

A number grid with numbers

Description automatically generated with medium confidenceLouvain Algorithm – The modularity-based algorithm that helps us identifies communities which have dense links within the clusters. The relevance would be knowing which communities in the data sets are more active in this case Is the community number 7. Furthermore, we could also identify the user groups, in the community and see how their conversation develop to further promote Taylor Swift’s popularity.

*Figure 13a. Louvain algorithm result.*

*A network of words and lines

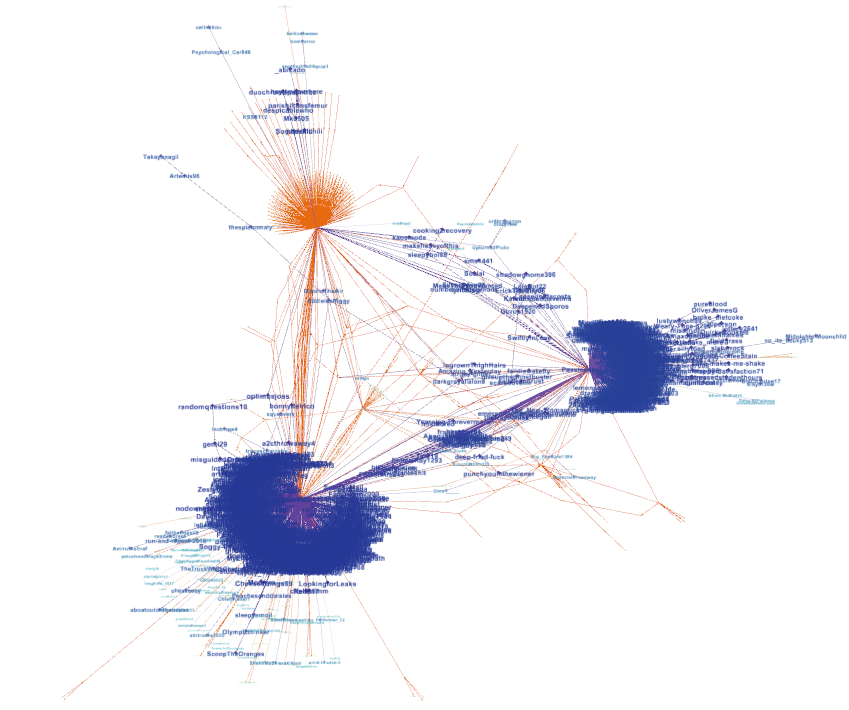
Description automatically generated with medium confidence*

*Figure 13b. Louvain modularity class visualization*

A grid of numbers on a white background

Description automatically generatedGirvan-Newman – This algorithm has the same goal but has a different approach where it progressively removes edges with the highest betweness centrality, to split the network to smaller subgroups hence identifying the communities. This algorithm has a relatively similar result with the Louvain algorithm only having a difference on a minimal amount of users in a community.

*Figure 13c. girvan-newman algorithm result.*

**

*Figure 13d. girvan-newman algorithm visualization.*

Relevance – Overall these algorithms are important in order to have a sub-topic exploration and also observe user engagement patterns. Both of these aspects will be a vital information to find a way to further promote Taylor Swift’s positive semantic response by the public,

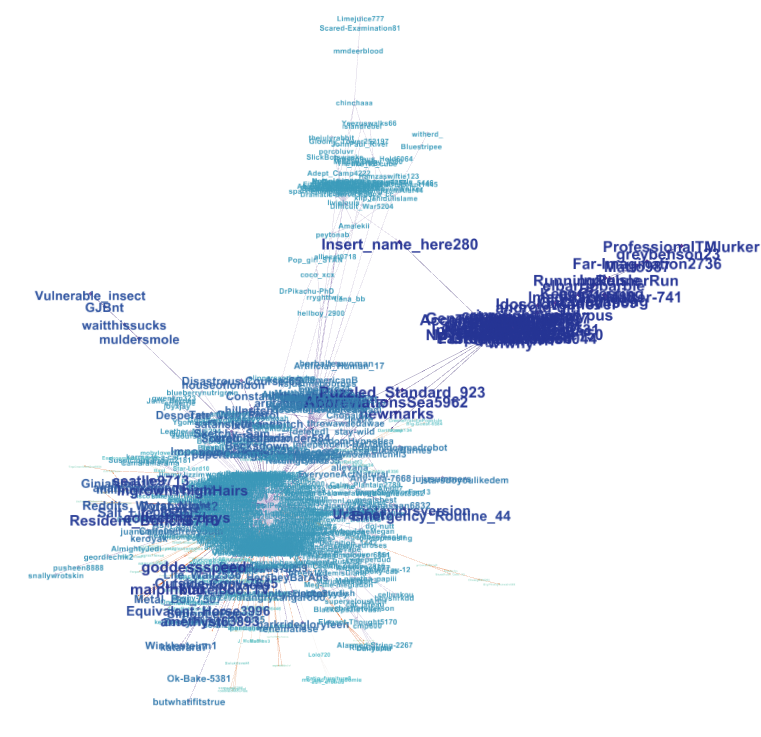
1. Olivia Rodrigo

In general, Olivia Rodrigo’s threads demonstrated a different structure in community analysis, as the distribution of activity only falls mostly in one community comparing to Taylor Swift’s 3 major communities.

A white background with numbers

Description automatically generated with medium confidenceLouvain Algorithm – The community size in Olivia Rodrigo compared to Taylor swift is far lesser and there is only one community that reached more than a hundred of interactions among users.

*Figure 13e. louvain algorithm results olivia rodrigo.*

 *Figure 13f. louvain algorithm visualization olivia rodrigo.*

A number grid with numbers

Description automatically generated with medium confidenceGirvan-Newman – This algorithm shared a similar result with it’s Louvain counter part. Comparing to Taylor Swift’s results, given that there are fewer users in the threads, it is also expected that these algorithm will show a far lesser results.

*Figure 13g. girvan-newman algorithm results olivia rodrigo.*

A blue ball with orange lines and lines

Description automatically generated *Figure 13h. girvan-newman algorithm visualization olivia rodrigo.*

1. Ariana Grande

In general, the huge difference of users active in Ariana’s threads compared to Taylor Swift resulted in a far lesser community generated. It can also be observed that in the communities are fairly distributed among each users.

Louvain Algorithm A number of sizes and numbers

Description automatically generated with medium confidence

A diagram of a light source

Description automatically generated with medium confidence *Figure 13i. louvain algorithm results ariana grande.*

*Figure 13k. louvain algorithm visualization ariana grande.*

A close up of numbers

Description automatically generated Girvan-Newman

*Figure 13k. girvan-newman algorithm results ariana grande.*

*Figure 13l. girvan-newman algorithm visualization ariana grande.*A diagram of a network

Description automatically generated with medium confidence

**Machine Learning Models**

1. Sentiment Analysis

On a high-level analysis of the data collected from reddit. It can be observed that mostly are just categorized on a neutral level. This was achieved by categorizing the phrases into sentiment labels which are Positive, Neutral, and Negative using the function below:

A black and white text

Description automatically generated

*Figure 14a. sentiment labels*

A screenshot of a computer

Description automatically generated

*Figure 14b. sample data set*

A graph with different colored bars

Description automatically generatedUsing the ggplot function in R, a bar graph was produce for easier comparison among the sentiments of the redditors. According to the obtained data, the redditors were mostly anticipating or happy about the newly released albums. The rest of the emotions in the visualization portrays the genre of the album Taylor Swift released.

*Figure 14c. visualization of sentiment analysis*

1. Decision Tree

Using a machine learning algorithm used for both classification and regression task, we are going to predict the likelihood of being a Taylor Swift song against 2 playlist from spotify “Daily Top 50” and “Daily Mix”. This would be interesting as Daily Top 50 are the top 50 popular songs of the world, whereas the Daily Mix is a personalized one.

A screenshot of a computer

Description automatically generatedThe process is done by creating a binary attribute in both dataset (isTaylor). Value 0 indicates that it is not a Taylor Swift song and value 1 indicates that it is. Both of these data will then be combined by music of Taylor to be utilized when we are training the data.

A screenshot of a computer

Description automatically generated

*Figure 14d. Daily Mix mixed with Taylor Swift Figure 14e. Top 50 mixed with Taylor Swift*

A screenshot of a computer

Description automatically generatedBoth datasets are then separated into a 80:20 split, where 80 percent will be used for training the machine whereas the remaining 20 percent will be the testing. On the testing phase of the dataset, the results held a high accuracy of classification. 93.12% for the Top 50 Global mixed with Taylor Swift and 91.78% for Daily Mix with Taylor Swift Songs.

*Figure 14f. Top 50 Testing Data Result*

A screenshot of a computer code

Description automatically generated

*Figure 14g. Daily Mix Testing Data Result*

1. Topic Modelling

Topic modelling is a technique in natural language processing (NLP) which aims to discover the underlying themes or topics in a collection of documents. This will help identify patterns and group them according to similarity. This would be helpful in the music industry as underlying trends and topics can be discovered through this algorithm. Furthermore, feedbacks from fans can also be obtained allowing the artist’s management team to adjust accordingly.

The algorithm was done by using the data collected from reddit and was cleaned and stripped off of any noisy data. The collected data was reduced down to its stop words and is then converted into a document term matrix. Using the Latent Dirichlet Allocation (LDA) with a set number of clusters in this case is 6. This will then return the top topics from data collected data which can then be used to collect information from the redditors.

A group of colorful bars

Description automatically generated with medium confidenceThis algorithm was ran through 3 datasets which are Taylor Swift, Ariana Grande and Olivia Rodrigo. Below are the results of these popular artists and what their fans talk in their threads.

*Figure 15a. Taylor Swift Topic Modelling*

A group of colorful bars

Description automatically generated with medium confidenceA group of colorful bars

Description automatically generated*Figure 15b. Ariana Grande Topic Modelling*

*Figure 15c. Olivia Rodrigo Topic Modelling*

**DATA VISUALIZATION**

1. Data Visualization for Reddit and Spotify
2. Reddit Data

A screenshot of a computer

Description automatically generatedThe visualization using PowerBI will be showcasing mainly of the redditors statistics across the for threads that was data collected. These subreddit thread represents the album released by Taylor Swift in the past 3 years. This visualization we can see how one fan or in business case one social manager can effectively increase an artists popularity

*Figure 16a. Reddit overview for Taylor Swift Data*

The most interesting findings in this chart is how one user garnered more than half of the interactions in reddit under the thread of Taylor Swift. It can be safely assumed that this user is working for Taylor Swift as a social media manager where the users post gained 16M number of comments which is massive for a reddit’s space.

A screenshot of a computer

Description automatically generated

*Figure 16b. Social media manager impact*

1. Spotify Data

The key visual for the Spotify compares Taylor Swift towards other artist. It can be seen how Taylor Swift is dominating in terms of popularity and follower count. One of the things that can be noticed here is about Ariana Grande. Despite having low presence in Reddit, she has around 100M of followers, but her popularity is only at 90th percentile. With this visual we can safely conclude the fan to fan interaction drives popularity of the artist.   
  
A screenshot of a graph

Description automatically generated

**ANALYSIS REVIEW**

1. Different Models and Visualizations

Conducting social media analysis using various methods and algorithms provides unique insights, and there is no single approach that fits all purposes. For tracking current trends in social media or the music industry, NLP algorithms are highly effective. If the goal is to classify artists, decision trees and their counterpart, Bayesian algorithms, can be useful. For analyzing non-linear relationships within data, advanced algorithms like Bayesian Networks and Neural Networks offer robust options.

Additionally, there are diverse visualization techniques that can effectively present data. For instance, heatmaps are valuable for displaying an artist's popularity across different geographical locations, while network graphs can illustrate relationships between users and topics. When it comes to topic analysis, word clouds are useful for quickly conveying the essence of a topic and its associated terms.