Princess Component Analysis:

A Sentiment Analysis on Disney Movie Scripts

Data Mining & Text Mining Final Project Report
By Hailey Skoglund, Gus Lipkin, & Jake Greenberg
CAP 4770 | Professor Jikhan Jeong
December 8, 2021

Princess Component Analysis

A Sentiment Analysis on Disney Movie Scripts

Abstract

In this study, our team analyzed the emotional tone present in the movie scripts and subtitles from the theatrical releases of the official Disney Princess movies produced by Walt Disney World Animation Studios. The goal of this study was to explore the overall sentiment in each of the Disney princess movies and analyze the most frequent words used throughout these beloved Disney classics. In order to accomplish this, we implemented data mining and text mining techniques such as principal component analysis, sentiment analysis, and association rule mining to reveal insights about this movie data. These results showed some common themes and words in each movie, and also highlighted the ways that each film is subtly different, in order to keep audiences engaged. All of the data was placed into various types of plots, including ridgeline, violin, and histograms, in order to better visualize each

1 Introduction

Our topic focuses on analyzing movie scripts from the theatrical releases of the Walt Disney World Animation Studios official Disney princess movie line up. Our motivation to select this topic originated because one of our teammates was being considered for a data analytics internship with The Walt Disney Company. This sparked our interest to explore what data could be analyzed to practice some of the data mining and text mining techniques that we have learned throughout the semester. This topic is important because these techniques can be used to visualize the overall emotion over the course of any movie. In this project, our team focused on modeling the overall emotional sentiment present in each of the Disney princess movies and word choice present in these movie scripts throughout the course of these movies.

2 Data

In this data, we explored the overall sentiment in each of the Disney princess movies over time and word choice present in these movie scripts. We collected this data through researching Disney princess movies subtitle files. Our data originated from a text-based dataset found on opensubtitles.org. The variables that we studied in this data include year released, the runtime, and the songs inside the movie.

To analyze this data, we're also using the Afin, Bing, and NRC lexicons for use with sentiment analysis.

3 Methodology

For the analysis, we primarily used the 'data.table','tidytext' and 'tidyverse' packages, with some other supporting libraries, including the 'srt' library for reading the subtitle files.

Starting with our list of movies, we import our movieData.csv, which contains the year released, the runtime, and the same if it had a live adaptation. This ensures we'll have all of the initial data for each movie.

We assign each .srt file to one of our movies. From here, we re-scale the runtimes from 0-1 so we can track the runtime of each movie. Additionally, we add a 'song' column, for movies that possess music.

For movies that contain music, we make sure that the program is aware that those portions of the movie are in song. This will help us track emotion through the songs comparatively to the rest of the dialogue.

Using our three lexicons, Afin, Bing, and NRC, we generate the sentiments for each word of each script, and save them as longer data.

From here, we generated a plot using ggridges, utilizing the NRC sentiment analysis. This allows us to view the NRC emotions over the course of each movie, as well as general changes in the amount of emotion as the movie progresses

For the Afin lexicon, we run the sentiment analysis by both line and by word for all of the movies.

We then generated a violin plot that contains the variation in positive and negative emotions of words, by the type.

Finally, we gather the sentiment scores for each movie. We do this for each of the lexicons, Afin, Bing, and NRC, and we organize it by emotion.

This allows us to generate a 'radar chart' which shows each movie's relation to the relevant emotions. This visualization was thrown out, as we found better options to visualize our results.

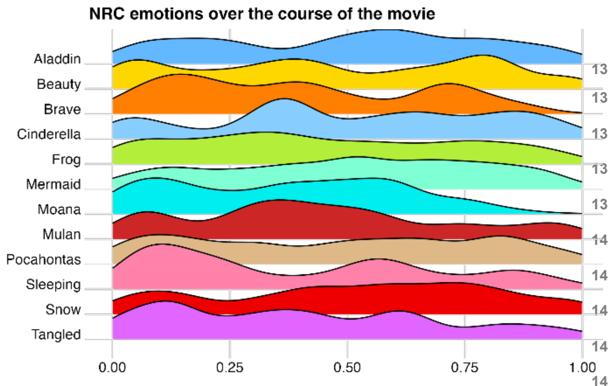
Finally, we decided to create a wordcloud visualization using the ‘worldcloud2’ library, which showcased the most used words, color coded and separated by the movie it appeared in.

4 Results and Analysis

Because this project used many different types of visualizations and techniques, this section has been broken up based on the technique used.

4.1 Ridge Plots (ggplot, ggridges)

The ggridges plot documented the summation of the NRC emotions over the course of the movie’s runtime.



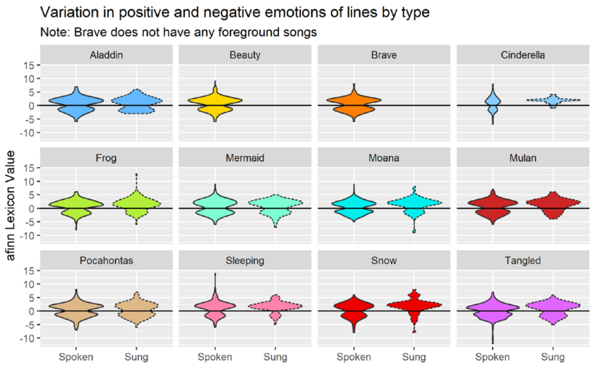
What we found was that the movies had some things in common:

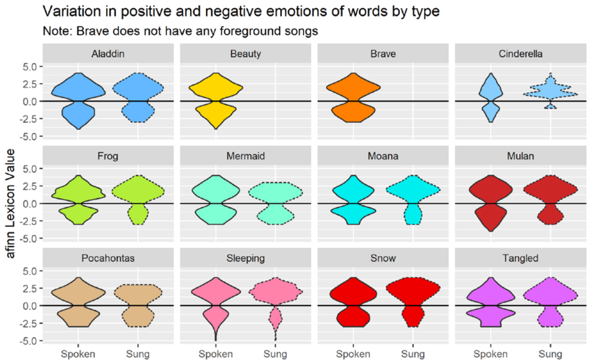
With only a couple of exceptions, the beginnings of each movie (the first 20% or so) all appear to have a local maximum. This implies that the openings of each movie each sport a relatively high peak level of emotion, comparatively to other parts of the movie. We normally then see a downward trend in the summed NRC levels until we approach a turning point or the climax of the film.

Something surprising about the visualization was that no two movies have an extremely similar NRC pattern. By looking at the course of the emotions over the sum of the movie, we can see that each movie has some level of distinctive pacing.

4.2 Violin Plots (ggplot)

The next visualization looked at the positive and negative values by each line of a movie, and by each word of a movie. In order to preserve the fact that some words and lines are spoken in song, the graphs show both spoken and sung words and lines.

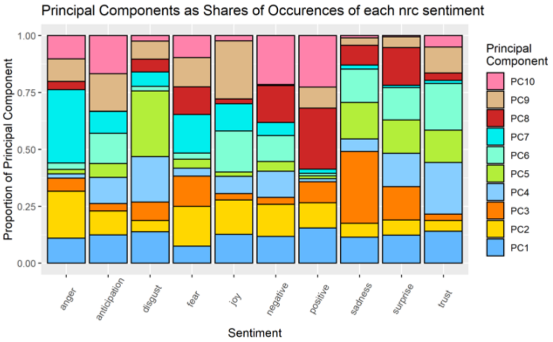




Both plots show the variation in emotion as a result of the words or lines in a given movie. We can see that for the most part, the movies lean towards a mild positivity, with some noticeable exceptions. For example, we can see that Tangled in both is spoken and sung dialogue, sometimes becomes extremely dark and negative, as opposed to a movie like Cinderella, which is primarily positive throughout. Most of the movies have a mix of both positive and negative emotions, as noted by the similar distributions above and below zero, but none of them generally lean too heavily into one or the other.

4.3 Principal Component Analysis (PCA)

The Principal Component Analysis that was conducted for each movie aimed to capture the representation of each NRC emotion within each principal component.

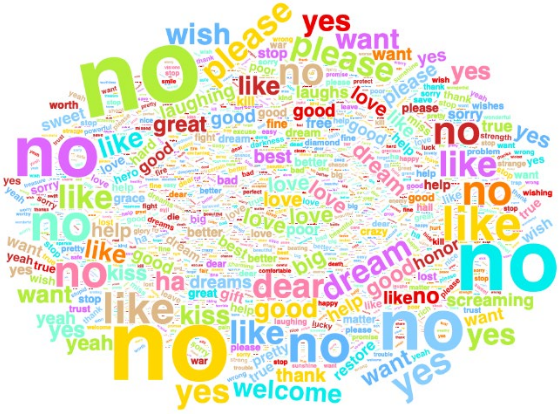


Along the y-axis we have each of the NRC emotions, and the columns above showcase their representation by the PCA they are captured by. For example, we can see that anger is most heavily documented by the 7th principal component, but it barely has any representation within the 4th, 5th, or 6th principal component. Other examples include positivity being captured primarily within the 7th principal component, and sadness being captured primarily within the 3rd principal component.

While these results and the visualization are interesting, the PCA did really manage to capture any of the emotions, or bunch any of them together effectively. This means that the PCA was not very successful.

4.4 Wordcloud

Using the top words of each movie, we generated the following wordcloud:



If anything is immediately noticeable about the wordcloud, it is the prevalence of the word "no". There could be a lot of reasons for this, but we believe the reason is due to the plots of each Disney movie. Each Disney princess is often told off for doing something that they shouldn't have done, or they should not be trying to break the mold. This often ends up becoming a focal point of each movie, so the word "no" could see heavy usage due to this. Other than that, we can see a prevalence of certain words as they apply to certain movies. For example, Cinderella has heavy usage of the words "dream" and "dreams," and this is the primary message of Cinderella.

5 Limitations and Future Study

Some next steps for a future study could involve analyzing the positive or negative sentiments at the end of the movie in relationship with the audience's satisfaction after the movie to determine if there is any correlation between the audience's responses to 'happy endings' or 'sad endings'. With this, we can also determine if Disney movies are more likely to result in 'happy endings' than other movies.

Additionally, the prevalence of the word "no" within the wordcloud means there could be certain recurring topics that are appearing alongside it within each movie. It may be possible to use association rules (arules) or topic modeling (LDA) to find words or topics that "no" is associated with.

References

Frangidis P., Georgiou K., Papadopoulos S. (2020) Sentiment Analysis on Movie Scripts and Reviews. In: Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. *IFIP Advances in Information and Communication Technology*, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1_36

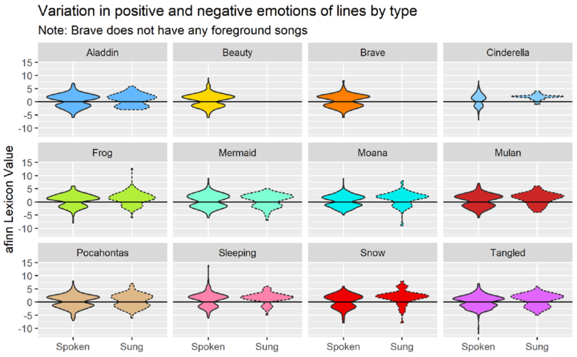
Sureja, N. D., & Sherasiya, F. A. (2017). Using sentimental analysis approach review on classification of movie script. *International Journal of Engineering Development and Research*, 5(2), 616-620.

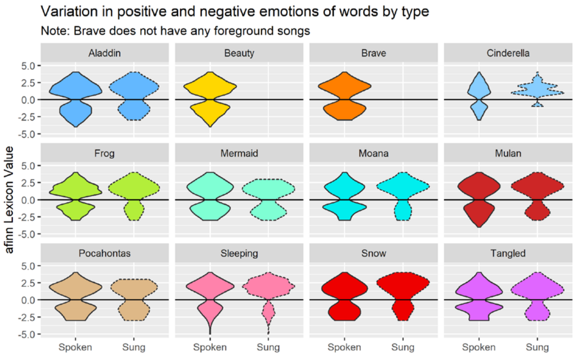
Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv. 50, 2, Article 25 (June2017)*

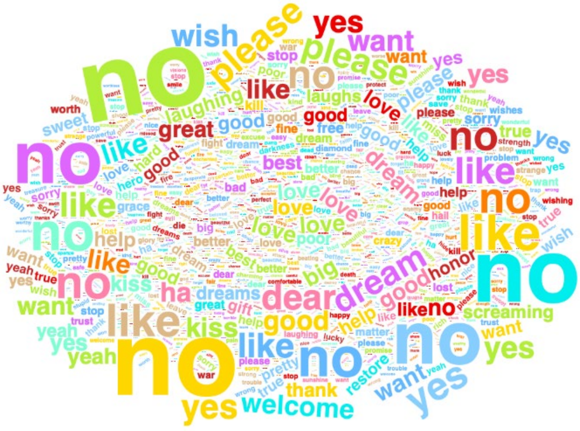
Lipkin, G., Pierstorff, A., Skoglund, H. (2020). Final Report for Intro to Data Science. Florida Polytechnic University. [https://github.com/guslipkin/disney_ds/blob/master/project/report_and_code.pdf](https://githu b.com/guslipkin/disney_ds/blob/master/project/report_and_code.pdf)

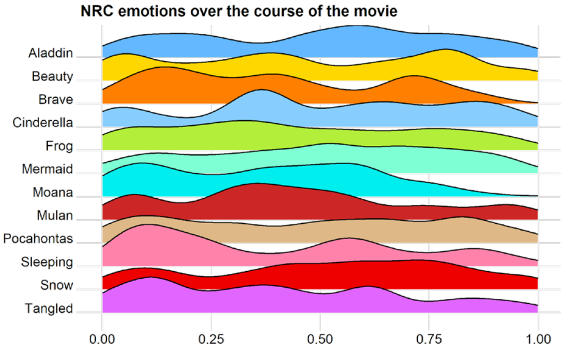
Happily ever after? A sentiment analysis of 7 disney princess films. Recovery Decision Science. (2016, July 20). Retrieved November 12, 2021, from https://recoverydecisionscience.com/happily-ever-after-a-sentiment-analysis-of-7-disney-princess-films/.

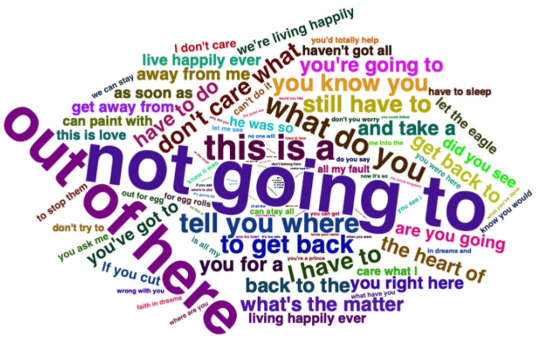
A Appendices

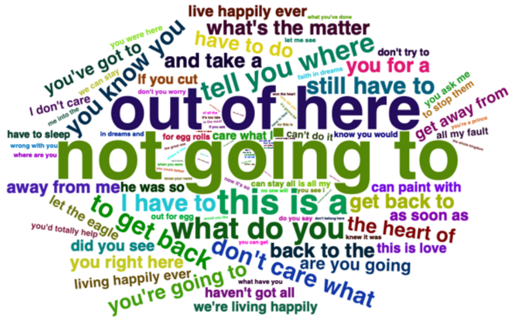


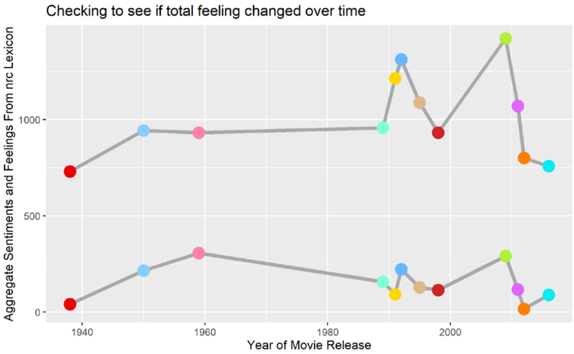


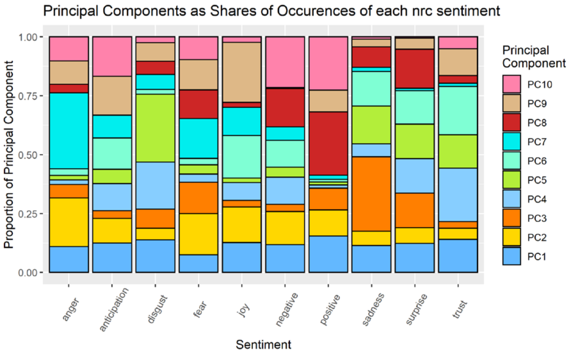


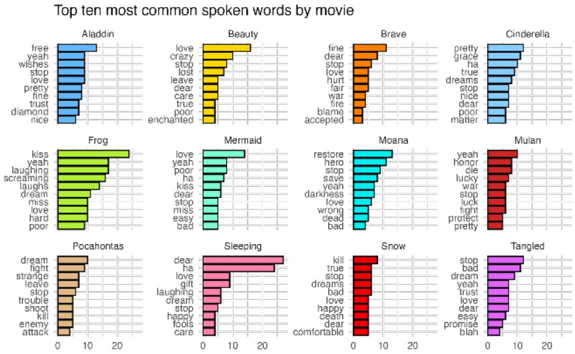


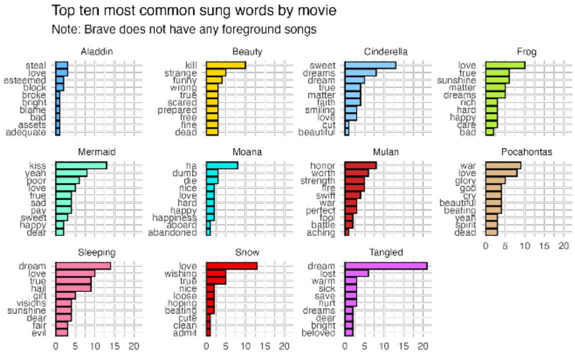





Checking to see if total feeling changed over time
Aggregate Sentiments and Feelings From nrc Lexicon
0
500
1000
1940
1960
1980
2000
Year of Movie Release


Principal Components as Shares of Occurences of each nrc sentiment
Proportion of Principal Component
0.00
0.25
0.50
0.75
1.00
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
Sentiment
Principal Component
PC10
PC9
PC8
PC7
PC6
PC5
PC4
PC3
PC2
PC1


Top ten most common spoken words by movie
Aladdin
free
yeah
wishes
stop
love
pretty
fine
trust
diamond
nice
Beauty
love
crazy
stop
lost
leave
dear
care
true
poor
enchanted
Brave
fine
dear
stop
love
hurt
fair
war
fire
blame
accepted
Cinderella
pretty
grace
ha
true
dreams
stop
nice
dear
poor
matter
Frog
kiss
yeah
laughing
screaming
laughs
dream
miss
love
hard
poor
Mermaid
love
yeah
poor
ha
kiss
dear
stop
miss
easy
bad
Moana
restore
hero
stop
save
yeah
darkness
love
wrong
dead
bad
Mulan
yeah
honor
die
lucky
war
stop
luck
fight
protect
pretty
Pocahontas
dream
fight
strange
leave
stop
trouble
shoot
kill
enemy
attack
Sleeping
dear
ha
love
gift
laughing
dream
stop
happy
fools
care
Snow
kill
true
stop
dreams
bad
love
happy
death
dear
comfortable
Tangled
stop
bad
dream
yeah
trust
love
dear
easy
promise
blah
0
10
20


Top ten most common sung words by movie
Note: Brave does not have any foreground songs
Aladdin
steal
love
esteemed
block
broke
bright
blame
bad
assets
adequate
Beauty
kill
strange
funny
wrong
true
scared
prepared
free
fine
dead
Cinderella
sweet
dreams
dream
true
matter
faith
smiling
love
cut
beautiful
Frog
love
true
sunshine
matter
dreams
rich
hard
happy
care
bad
Mermaid
kiss
yeah
poor
love
true
sad
pay
sweet
happy
dear
Moana
ha
dumb
die
nice
love
hard
happy
happiness
aboard
abandoned
Mulan
honor
worth
strength
fire
swift
war
perfect
fool
battle
aching
Pocahontas
war
love
glory
god
cry
beautiful
beating
yeah
spirit
dead
Sleeping
dream
love
true
hail
gift
visions
sunshine
dear
fair
evil
Snow
love
wishing
true
nice
loose
hoping
beating
cute
clean
admit
Tangled
dream
lost
warm
sick
save
hurt
dreams
dear
bright
beloved
0
5
10
15
20