Rady | UC San Diego School of Management

ListenFirst Capstone Report

Presented By

Alexander Ilyin (Team Captain)
Ming Ki Toby Cheng (Team Manager)
Ziyuan Yan
Jian Bin Liew

Advised By

Natasha Balac
Raymond Pettit

Table of Contents

1. Abstract **3**

1.1. Business Problem 3

1.2. Deliverables and Outcomes 4

1.3.1. Brand Score 4

1.3.2. Text Analysis Techniques 5

1.3.3. Movie Success Prediction 6

1.3.4. Aspect-Based Opinion Mining 7

1.3.5. Summary of Methods 7

2. Data **9**

2.1. Datasets 9

2.3. Exploratory Data Analysis 10

2.3.1. Movies 10

2.3.2. Rollup 11

2.3.2.1 General 11

2.3.2.2 Drilldown 13

2.3.3. Success vs Flop 16

3. Modeling **18**

3.1. Brand Score 18

3.1.1. Step 1: Cluster 18

3.1.2. Step 2: Scoring 22

3.1.3. Step 3: NLP 24

3.1.4. Discussion 27

3.2 Movie Success Prediction 28

3.2.1 Step 1: General Models 28

3.2.2 Discussion 35

3.3. Aspect-Based Opinion Mining 36

3.3.1. Step 1: Label 36

3.3.2. Step 2: Train 37

3.3.3. Step 3: Sentiment Analysis and Topic Model 38

3.3.4. Discussion 39

4. Implementation **40**

5. Conclusion **41**

1. Abstract

The following describes the procedure of how the solution was fully implemented. Three models were designed - brand score, movie success prediction and aspect mining. These models were then put in a pipeline for deployment. Ultimately, these models are used to help guide decision makers to understand their audience and make better and more targeted approaches to their marketing campaign. The pipeline is illustrated in Figure 1.1, and will be explored again in section 4, ROI.

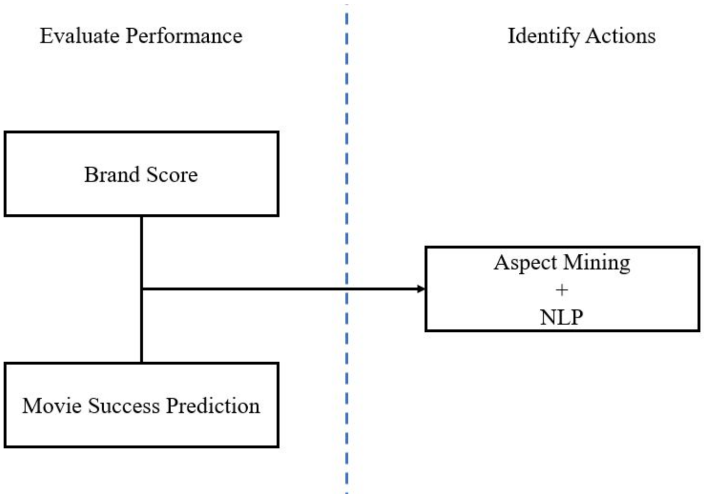


Figure 1.1: Implementation Pipeline

1.1. Business Problem

Listenfirst had tasked the team with answering the following question:

> “*How can a client use their customer’s social media interaction data to influence their social media strategy?*”

Based on meetings with Listenfirst, the team decided to approach this question from two angles. First, Listenfirst had requested a scoring methodology that could help their clients understand the effectiveness of their current social media campaign. Second, based on the fluctuations in the score, the client could take a deeper look into what caused the changes in score over time as well as see a prediction of the success of their product based on the current score trend. The team's solutions were based on the idea of adding value to a client's decision-making process before the movie's release date.

1.2. Deliverables and Outcomes

Following the initial meetings with Listenfirst as well as the data transfer process, the team began to map out a direct approach. While the project had been split in two, it was still necessary to determine the actual methodologies that would be used to address the business problem. To address the first half of the project, the team would use an unsupervised clustering algorithm to cluster clients based on their attributes and create a custom scoring formula based on the attributes contained within Listenfirst's data. This part was to be completed before the midterm. Following the midterm, the team would address the in-depth analysis of text data provided by Listenfirst using both NLP algorithms and simple exploratory text analysis, as well as the prediction of product success based on trends of the score.

1.3.1. Brand Score

With the team's approach to the business problem being split in two, the motivation behind the brand score was to create a scoring formula that would be solely based on raw interactions in the form of clicks. In this case, the team defined two different kinds of social media interactions: raw and engaged interactions. Raw interactions could be defined by any interaction that is solely click-based. With Listenfirst collecting data from three social media platforms (Facebook, Twitter, Instagram), the type of clicks varied between the three. For example, data on Facebook interactions included clicks such as "angry" and "sad" reactions, while Instagram and Twitter data only tracked clicks such as "likes/favorites". Engaged interactions could be defined by anything that involved user input, such as comments/replies.

An important caveat in the scoring process was that a client would have the ability to compare their score against other brands in the same domain. In order to make this possible, before scoring the client, the client would be grouped with other brands with similar attributes.

The specifics of this clustering approach will be detailed in later parts of the report. For each group of brands, their scores would be calculated over a range of 1 year before release to 1 year after release. By comparing scores with similar brands, the client would be able to make more insightful comparisons, rather than making comparisons with brands that are vastly different.

The first complication in creating the score formulas came in the different kinds of raw interactions tracked between the three social media channels. Another complication lies in the fact that some clients did not have social media pages on some of the channels. For these reasons, it was not possible to create a single formula to score movies on their performance from all three channels. Instead, the team created a separate formula for each channel. Further, trends and fluctuations in a movie's brand score, whether it is an increase or a decrease, would be indicative of audience reactions towards marketing actions. For example, a decrease might indicate that the audience are reacting negatively towards a new trailer or poster. Hence, brand score accompanied with exploration through a sentiment analysis or topic model will be able to guide decision-makers on how to adjust a marketing campaign.

Another important caveat in the scoring approach was to group the brands based on date, and create an aggregate score for each date. Since the movies in the data spanned a wide range of years, it was not as effective to aggregate by given date, as it was to aggregate by "days after release". This way, all the movies could be grouped together, no matter which year they were released in. For each day "after" a movies release, its score would be compared with the average score of a cluster on that particular day.

By scoring a client based on their raw interactions, the two parts of the approach to the business problem would be kept separate. In this case, it was important to create a score that could accurately capture changes in customer sentiment so that the insights of the analysis of the score trend could be independent of the score itself.

1.3.2. Text Analysis Techniques

With the score taking care of the first part of our approach to the business problem, the next set of steps involved a thorough analysis of the social media interactions. On a given day, a client would be able to see a wide range of different visualizations of text data. These included word frequency charts (both for general comments and comments of positive/negative sentiment), word clouds, and a topic model that could draw out topics of discussion on a given day. The quality of these analyses would of course depend on the volume of interaction on said day.

1.3.3. Movie Success Prediction

After analyzing the engaged interactions, the movie-success prediction solution allows the client to see if their product would be a ‘success’ or ‘flop’ given the score trends up until that point. With the scope of this project, a product was to be called a success if its profit exceeded the production budget, as these were the features available to us. In terms of prediction methods, the team settled on using a range of machine learning models, such as supervised binary classification using Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, and a simple Neural Network. In addition, another method the team tried was a Recurrent Neural Network using LSTM/GRU layers. This final method would incorporate the time based/sequential data Listenfirst had provided, while the first methods would use aggregated interaction features.

This is a possible solution as profit and total post interactions exhibit medium positive correlation as illustrated in Figure 1.3.2.1 and Table 1.3.2.2. Each individual graph in Figure 1.3.1 is denoted by months before the movie's release date. Given this information, decision-makers will be able to decide to increase marketing efforts and engagement given current levels social media interactions.

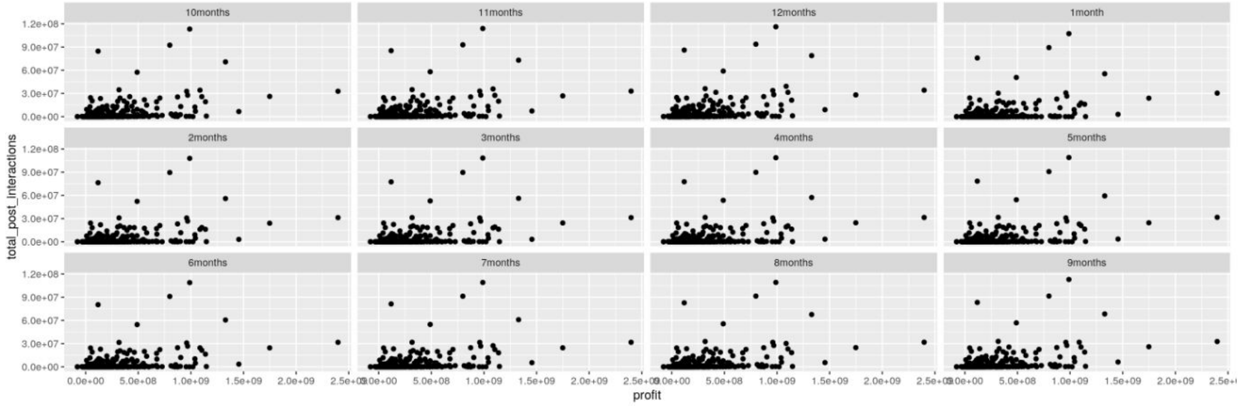
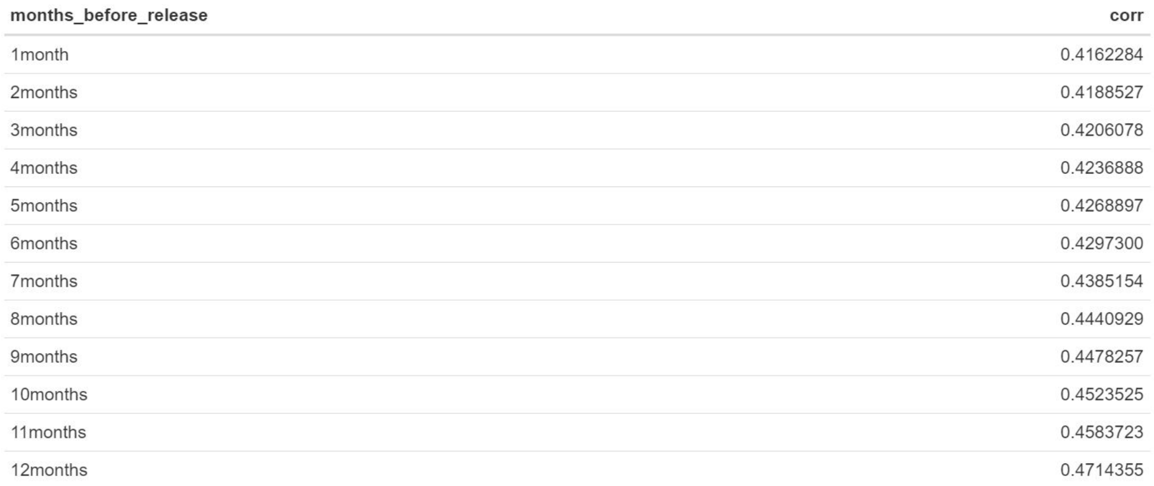


Figure 1.3.2.1: Total post interactions against profit plot



months_before_release	corr
1month	0.4162284
2months	0.4188527
3months	0.4206078
4months	0.4236888
5months	0.4268897
6months	0.4297300
7months	0.4385154
8months	0.4440929
9months	0.4478257
10months	0.4523525
11months	0.4583723
12months	0.4714355

Table 1.3.2.2: Correlation Graph

1.3.4. Aspect-Based Opinion Mining

This solution classifies each comment based on certain aspects like ‘social#mentions’ or ‘movie#characters’. After classifying these comments, decision-makers can do a deep dive into these aspects through topic modeling or sentiment analysis to be able to uncover what users are saying or feeling about that particular aspect. Decision-makers can then adjust their marketing strategies by taking in the audience feedback into account.

1.3.5. Summary of Methods

Table 1.3.4.1 broadly summarizes the three solutions and its use cases.

Business Problem How can a client use their customer’s social media interaction data to influence their social media strategy?				
Solution	Scoring brands in clusters	Text Analysis	Success/flop prediction	Classify comments based on aspects (e.g. movie, social)
Details	Score movies within clusters and drill down during variations to diagnose the reason for fluctuation.	Use exploratory analysis and models to allow clients to gain a deeper understanding of social media interactions	Use social media trends to predict success of movie	Classify comments on aspects and understand sentiment, topics discussed in each aspect.
Use Case	Movie is doing badly, as the brand score is going down. Use social media strategy from periods of good scoring for future posts	Topic models during those periods show what people are talking about. Can adjust product strategy going forward	Given current social media trends, the movie is predicted to be a flop. Marketing efforts can be intensified.	Find out what people are talking about regarding the characters in a movie and make adjustments on marketing material.
Data	Movie data	Comments	Rollup Movie Data	Comments

Table 1.3.4.1: Summary of Solutions

2. Data

This section contains a discussion regarding the data provided by ListenFirst along with any additional datasets the team compiled as well as the exploratory data analysis.

2.1. Datasets

The datasets came from three primary social media sources, Facebook, Instagram and Twitter. For each source, there were three types of datasets - comments, deltas and rollup. A dataset with general information about each movie and media channel information was also provided by ListenFirst. In addition, a dataset with information about movie attributes was scraped by the team from various sites containing movie attributes (IMDB, The-Numbers.com). Table 2.1 details what kind of information each dataset provides.

Dataset	**Information Provided**
Brand Source Affiliation	Brand ID, social media channel information
Movie	Movie attributes (e.g. genre, budget)
Deltas	Unaggregated interaction data (e.g. facebook likes)
Rollup	Daily rolled up data for interaction (e.g. total likes on a given day)
Comments	Social media comments

Table 2.1.1 Dataset Details

More information about how the data can be joined can be found in Appendix A. Most of the team's analysis was completed using data found in the Brand Source Affiliations, Movie, Rollup and Comments tables. Deltas was not used by the team due to the scope of the project, but can provide valuable insight if explored further.

2.3. Exploratory Data Analysis

This section will explore and give preliminary insights to the data.

2.3.1. Movies

In total, there was data on approximately 600 movies provided by ListenFirst. After filtering movies that were not released and old movies, there were 526 movies used for analysis. Figure 2.3.1.1 shows visualisations of the movies dataset and Table 2.3.2.1 shows the insights derived.

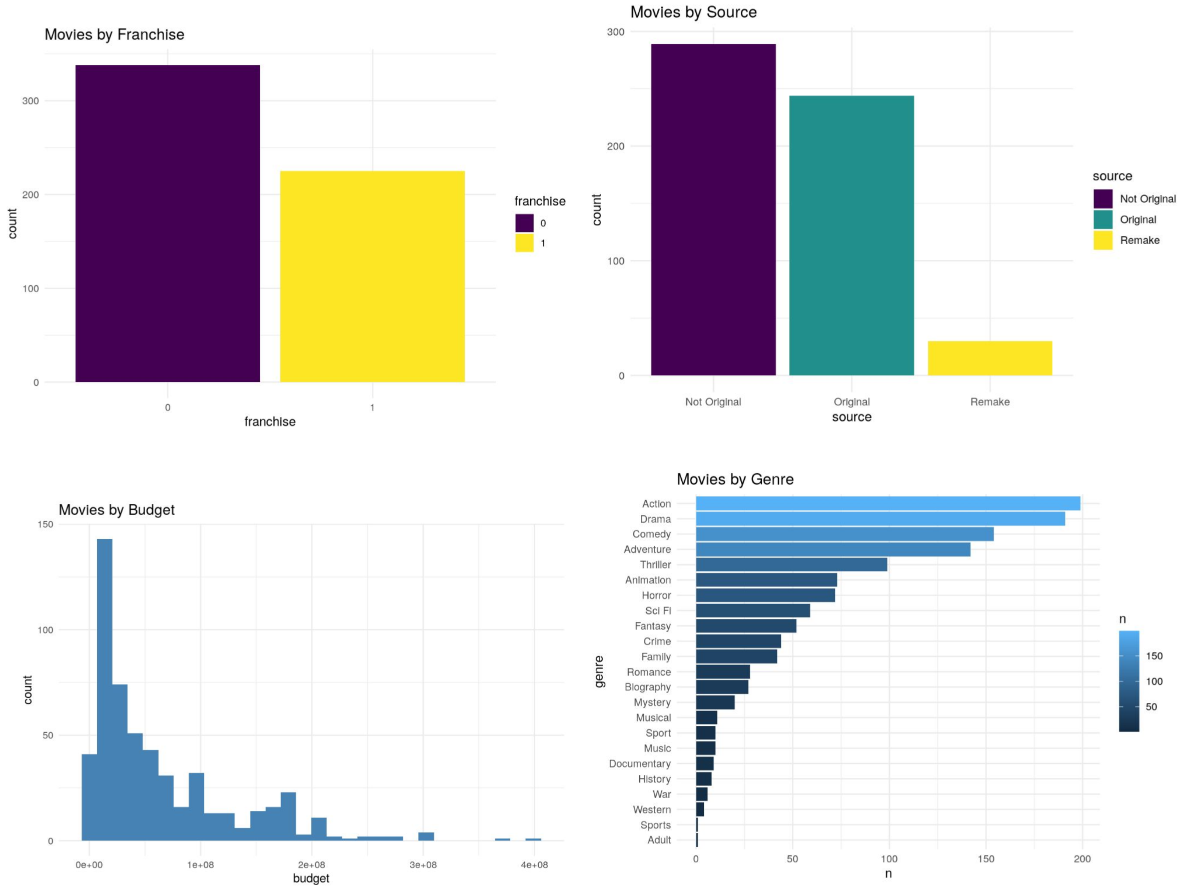


Figure 2.3.1.1: Movie Visualisations

Table	Insight
Movies by Franchise	Most of the movies are not from a franchise, however, a good deal of them still are.
Movies by Source	Most of the movies are derived from a source (e.g. television, comic)
Movies by Budget	Most of movies are made with a budget lower than 10 million
Movies by Genre	Most movies fall within the action, drama and comedy genres

Table 2.3.1.2: Movie Insights

2.3.2. Rollup

This section explores data aggregated one year before and after the release of a movie. Daily interaction information was averaged across all movies to identify a general trend across movies. Each social media platform has different metrics that are used for measurement. The trends are generally consistent across different social media platforms.

2.3.2.1 General

Facebook

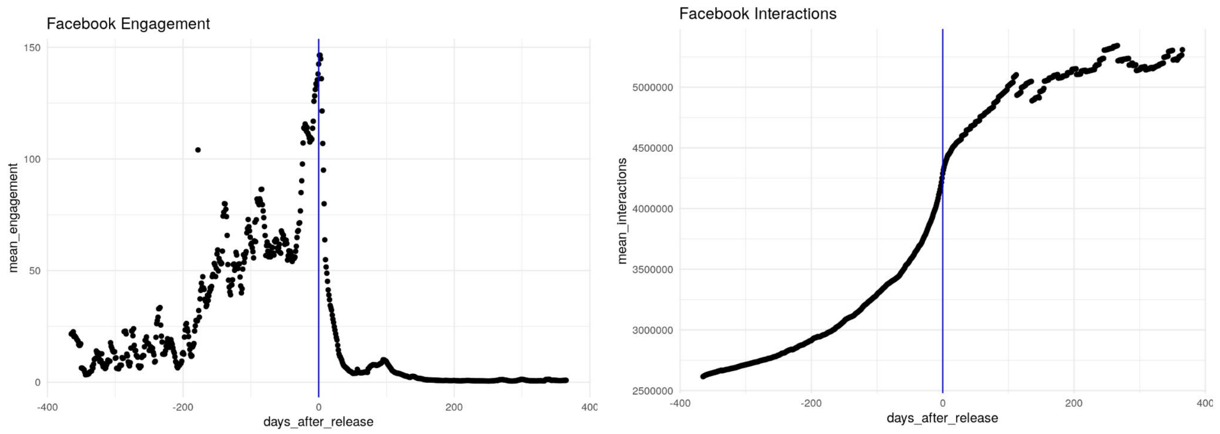


Figure 2.3.2.1: General Facebook Engagement and Interaction Trends

In general, engagement and interactions seem to spike up exponential approaching the date of movie release, at around the 90 day mark. Further, right after the release of the movie, engagement and interaction volume decrease exponentially as well. Before the release of the movie, Facebook interactions exhibit a consistent upward trend while engagement has a lot of noise. A reason for this can be that the engagement metric is very dependent on the quality of posts by the brand and captures positive interactions while interactions simply measure raw interaction volume, both positive and negative. Figure 2.3.2.1 shows both graphs.

<u>Instagram</u>

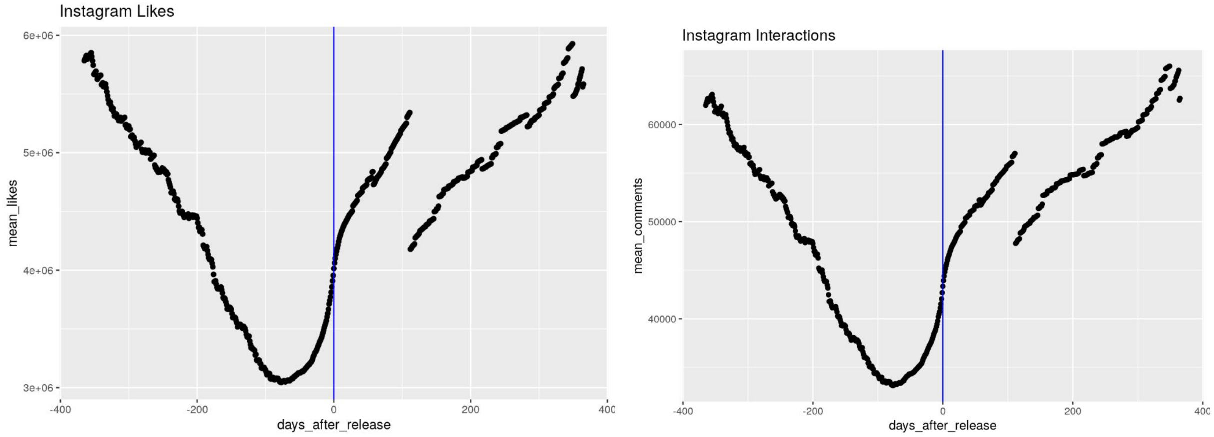


Figure 2.3.2.2: General Instagram Likes and Interaction Trends

Instagram likes and interaction exhibit very similar trends. Likes and interactions go on a downward decline at an exponential rate until approximately 90 days before the release of the movie, where it increases exponentially from that point. After the release of the movie, the trends show an exponential decline. This can be explained by the fact that marketing and 'hype' for the movie is usually built up from 90 days before the release of the movie. After the release as people watch it, engagement would decrease sharply as marketing activities decrease as well. The decrease in interaction and likes until the 90 day can be explained in the following drilldown section.

<u>Twitter</u>

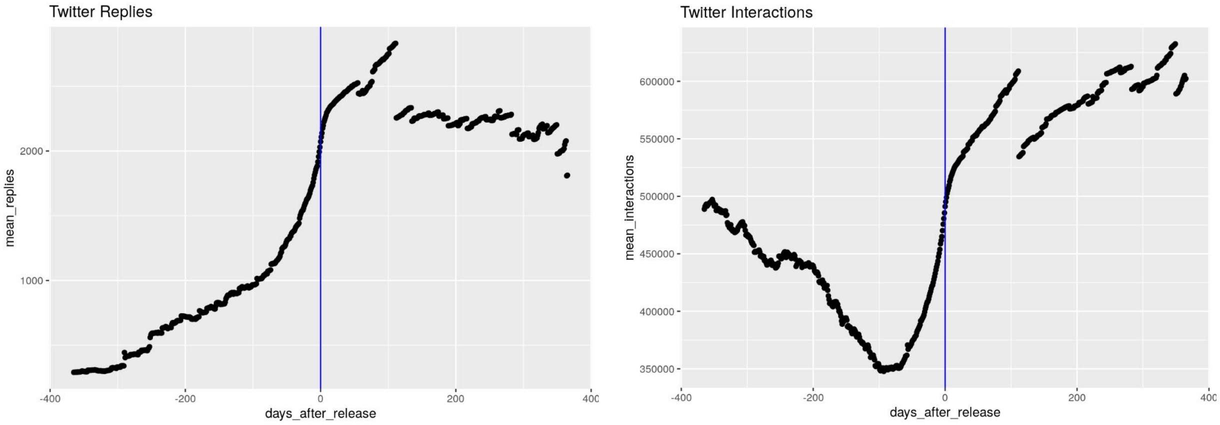


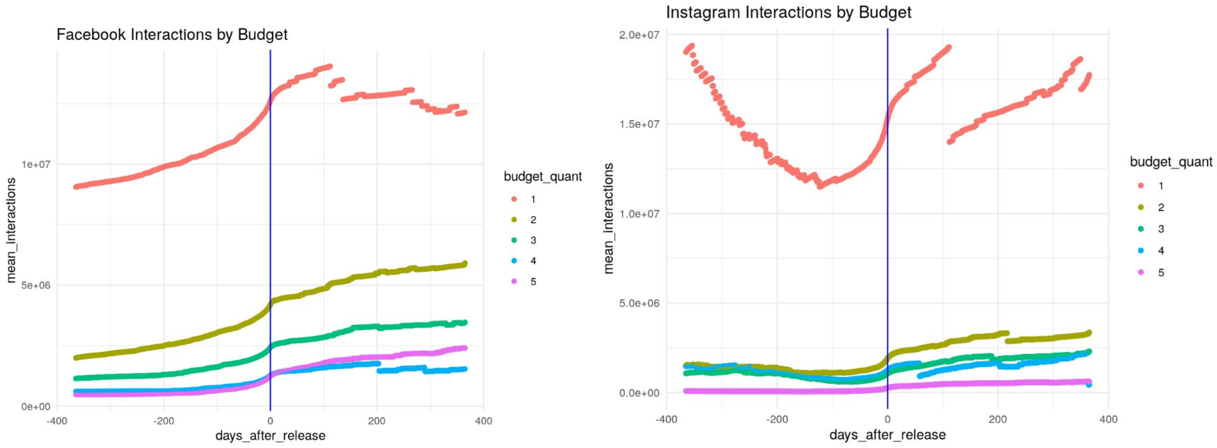
Figure 2.3.2.3: General Twitter Reply and Interaction Trends

Similar to Instagram and Facebook, Twitter reply and interaction data exhibit very similar trends. There is an exponential increase around 90 days before the release of the movie in both replies and interactions. After the movie, there is an exponential decrease and generally tapers off after some time.

2.3.2.2 Drilldown

<u>Budget</u>

In this section, movies were sorted into bins based on the budget of the movie, where budget_quant = 1 being the highest budget movies and 5 being the lowest budget movies. Figure 2.3.2.4 shows the interaction trends for all three social media platforms.



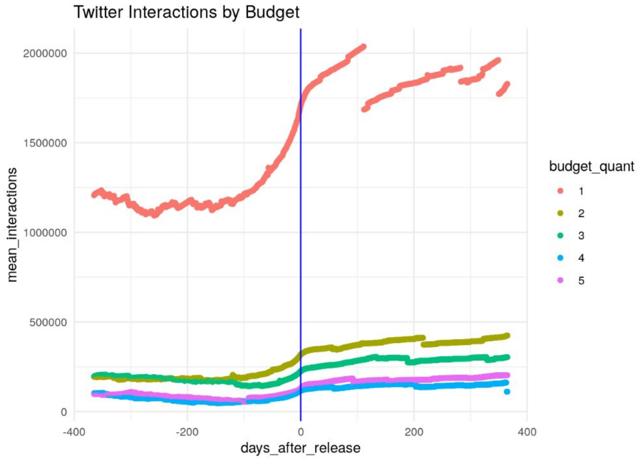


Figure 2.3.2.4: Facebook, Instagram & Twitter Interaction Trends by Budget

Generally, movies with a higher budget have more interaction, with the highest budget movies having substantially more interactions. The general exponential trend right before at approximately 90 days before a movie's release is still present, although more pronounced in higher budget movies. Interestingly, the decline in interaction is most prominent in the highest budget movies, as seen in the Instagram interactions. Hence, the decline in interactions seen in general interactions can be explained by the presence of a franchise. A franchise shares one social media account for multiple movies, hence, interactions would decline as it follows a declining trend after an earlier part of the series, and then ticks up again when a sequel is going to be released. Big budget movies are usually associated with franchises (movies in franchises get higher budget), hence, this finding is not surprising. The next section will explore interactions by franchise.

Franchise

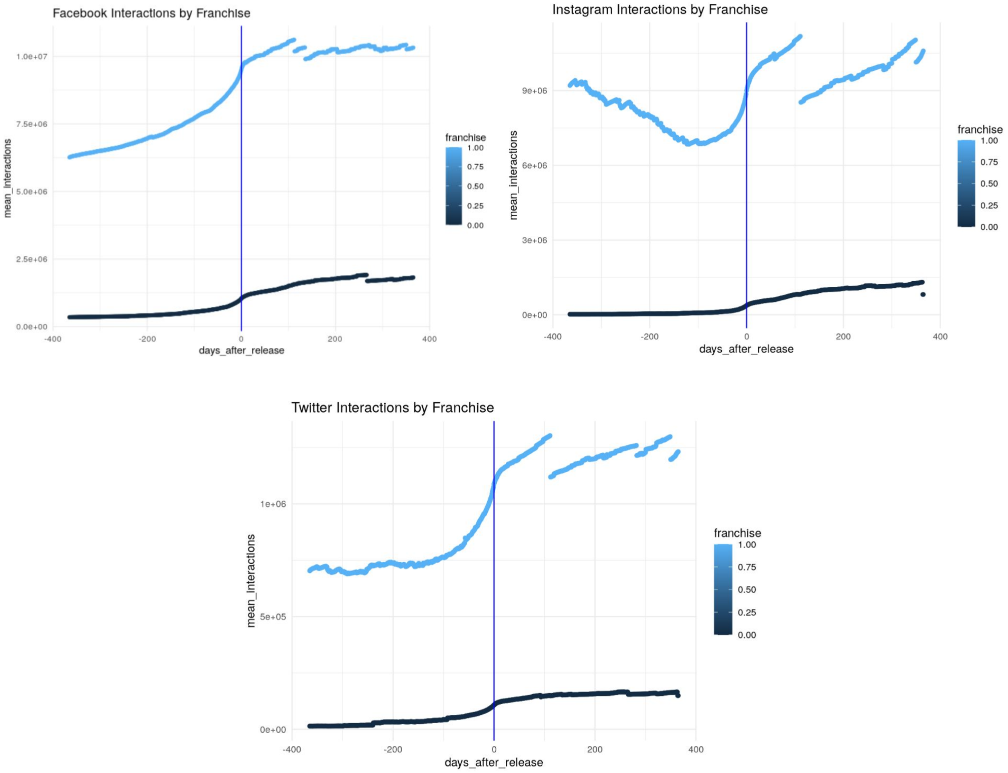


Figure 2.3.2.4: Facebook, Instagram & Twitter Interaction Trends by Franchise

Twitter and Instagram graphs correspond to prior analysis about how interactions decline due to presence of previous movies. An explanation as to why Facebook does not exhibit the same trend as while brands tend to share one social media account for Twitter and Instagram for multiple movies, brands actually open up multiple social media accounts for different entries of a franchise. This might be due to the way these social media platforms function, where it is more convenient to separate accounts on Facebook while combining them on Twitter and Instagram.

2.3.3. Success vs Flop

Drilling deeper, into interactions, there are key differences in interactions between successful movies and flops. These interactions will be able to demonstrate that there is potential in predictive modeling using social media interaction data.

Since distribution of total_post_interactions_c (daily interaction) does not follow any particular distribution, non-parametric statistical analyses were performed to compare success/flop movies. Taking Facebook data as an example, total_post_interaction (cumulative interaction) 30 days and 60 days before the movie release date were extracted. Three statistical tests were performed. Below are the results for scores 30 days before movie release date:

1. Two sample ks test to compare distributions.

   Null hypothesis (H0): both samples come from the same score distribution.

   Result: KS score = 0.169, p-value = 0.004 < alpha = 0.05. Reject the null hypothesis and conclude that the two score distributions are different.
2. Kruskal Wallis Test to compare medians.

   Null hypothesis (H0): population score medians are equal.

   Result: Median in success = 5968.363 > Median in flop: 3985.273, KW = 9.499, p-value = 0.002 < alpha = 0.05. Reject the null hypothesis and conclude that score medians are the same.
3. Mann Whitney U test to compare means:

   Null hypothesis (H0): population score means are equal.

   Result: Mean in success = 12887.585 > Mean in flop: 7770.630, MW = 23962, p-value = 0.00158 < alpha = 0.05. Reject the null hypothesis and conclude that score means are the same.

ROI analysis for interaction data only:

For the same reason, non-parametric statistical analyses were performed to compare success/flop movies.

Below are the results for total_post_interactions 60 days before movie release date:

1. Two sample ks test to compare distributions.

   Null hypothesis (H0): both samples come from the same interaction distribution

   Result: KS score = 0.143, p-value = 0.021 < alpha = 0.05. Reject the null hypothesis and conclude that the two interaction distributions are different.

2. Kruskal Wallis Test to compare medians.

   Null hypothesis (H0): population interaction medians are equal.

   Result: Median in success = 372811.5 > Median in flop: 230431.5, KW = 5.421, p-value = 0.02 < alpha = 0.05. Reject the null hypothesis and conclude that interaction medians are the same.

3. Mann Whitney U test to compare means:

   Null hypothesis (H0): population interaction means are equal.

   Result: Mean in success = 3854622.035 > Mean in flop: 3533274.604, MW = 24951, p-value = 0.00996 < alpha = 0.05. Reject the null hypothesis and conclude that interaction means are the same.

This data shows that successful movies generally have higher interaction and higher scores. This shows that there is potential in a predictive model to predict for a movie's success or failure through social media interaction data alone. Although interaction data is not the only dependent variable to determine whether a movie would be successful or not, other variables like sentiment cannot be easily changed by changing social media marketing strategies. Hence, one needs to be careful when interpreting this information.

3. Modeling

This section will discuss the implementation of predictive modeling, along with a discussion on the results and deployment.

3.1. Brand Score

As stated earlier, the full brand scoring process involved training clustering models on the brands provided by Listenfirst. A new client would then have their cluster predicted by the pretrained models, and the score of this client would be compared to the scores of the existing cluster. In order to compare the social media performance of similar movies, clustering methods based on movie features (budget, genre, MPAA rating, etc.) were used to categorize similar movies together. Different scoring methods were created for each social media channel to reflect different features of each platform. With no feature assigning brands to clusters, this was an unsupervised clustering problem. In order to better understand trends and fluctuations of the scores as well as acquire more insights, NLP was applied on the clusters to further analyze users' comments. An interactive HTML file is also generated for further exploration.

3.1.1. Step 1: Cluster

The first step is to create movie clusters based on movie features. Different combinations of the movie features were tested to determine which one led to the optimal clustering result.

Clustering Features:

The features that Listenfirst provided to the team included genre and release date. While these were strong indicators of movie similarity, these two features do not tell the full story of what makes movies similar. For this reason, the team required an additional data gathering process. Some features that were gathered included: production budget, MPAA rating, source (whether the movie was an original creation or based on another source), and whether or not the movie was part of a franchise. While this is not the full extent of features that can be collected in regard to movie attributes, we felt that this set would provide a stronger baseline for the clustering models.

In terms of changes to the features, the team felt that some features, such as genre and budget, required additional feature engineering. As shown in feature 2.3.1.1, the existing distribution of genre was quite spread. With only ~550 movies available for clustering, the team grouped the genres with the use of “domain knowledge”. As mentioned earlier, some movies were removed from the dataset due to their release date. However, even among the movies that remained, the monetary value of the budget was offset due to inflation. For this reason, the team adjusted all the budgets for inflation as of the year 2020.

Clustering Overview:

While most of the popular clustering algorithms involve numeric features, the data provided by Listenfirst includes solely categorical features. The team decided to focus on clustering algorithms that could use categorical features as well as a mix of both categorical and numeric features.

Method 1: K-Modes

The K-Modes algorithm is one of the earliest examples of a clustering algorithm for categorical features. While the rest of the algorithms chosen by the team involve some sort of numeric features or transformations, K-Modes can be directly applied to categorical features. Originally introduced as an extensions of the K-Means paradigm, Huang (1997)[1] introduced K-Modes and K-Prototypes as methods for incorporating categorical features into clustering with the use of mode as the statistical measure of distance and similarity measures that would count the number of differences between each point and a cluster center. Similar to K-Means, the algorithm’s hyperparameters could be adjusted by changing the number of clusters and number of computational iterations.

Method 2: K-Prototypes

[1] Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* **2,** 283–304 (1998). https://doi.org/10.1023/A:1009769707641

The second algorithm introduced by Huang (1997) was the K-Prototypes algorithm. As suggested by the name, this algorithm incorporates both numerical and categorical using the same similarity measure as K-Modes, while also incorporating Euclidean distance for numerical distance.

Method 3: MCA + K-means

Similar to Principal Component analysis, Multiple Correspondence Analysis (MCA) is used to reduce dimensions and reveal underlying relationships among categorical columns and transfer variables to numerical data types. As shown in Adbi & Valentin (2006)[2], MCA is an extension of simple Correspondence Analysis that allows for data with greater dimensionality. By representing high-dimensional data in a low-dimensional Euclidean space, categorical features can be converted to numeric features and used in popular clustering algorithms such as K-Means. The team found optimal results using four principal components. K-means was used following MCA, with Silhouette score and the elbow method used to determine the optimal number of clusters.

Method 4: FAMD + Hierarchical Agglomerative Clustering (HAC)

Factor Analysis for Mixed Data (FAMD), or Factorial Analysis for Mixed Data is a principal component method dedicated to analyze a data set containing both quantitative and qualitative variables (Pagès 2004[3]). It makes it possible to analyze the similarity between individuals by taking into account a mixed types of variables. The FAMD algorithm can be seen as a combination of principal component analysis (PCA) and multiple correspondence analysis (MCA). In other words, it acts as PCA quantitative variables and as MCA for qualitative variables. Quantitative variables are normalized during the analysis in order to balance the influence of each set of variables.

[2] Abdi, Hervé and Dominique Valentin. “Multiple Correspondence Analysis.” (2006). https://personal.utdallas.edu/~herve/Abdi-MCA2007-pretty.pdf

[3] Pagès Jérôme (2014). Multiple Factor Analysis by Example Using R. Chapman & Hall/CRC The R Series London 272 p

After reducing dimensions to the desired number of principal components and transferring variables to numerical data type, Hierarchical Agglomerative Clustering (HAC) was applied to cluster similar movies. Each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained (Rokach, Lior, and Oded Maimon, 2005[4]). The Silhouette score was used to determine the optimal number of clusters.

After thorough examination of results found from the four clustering algorithms chosen, it was found that optimal results came from Methods 3 and 4. It can be inferred that Methods 1 and 2 could have been improved using a larger feature set, however, due to limitations in data collection, this was not possible within the scope of the project. When comparing Methods 3 and 4, it is important to compare the optimal number of clusters found by the two clustering algorithms. Method 3 found three optimal clusters, while Method 4 found ten optimal clusters. With the goal of the clustering being to create insightful comparisons, a clustering algorithm with a larger number of optimal clusters would minimize the number of "outliers" in each cluster. For example, with only three clusters, it is inevitable that a cluster composed of mostly family-friendly movies would contain horror movies. While these kinds of errors exist even with the results in Method 4, it was more just as important for the team to minimize these errors, as it was to minimize the different clustering metrics (silhouette, inter cluster distance). For these reasons, Method 4 was chosen. Figure 3.1.1.1 shows the distribution of Silhouette scores for different numbers of clusters using Method 4.

[4] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

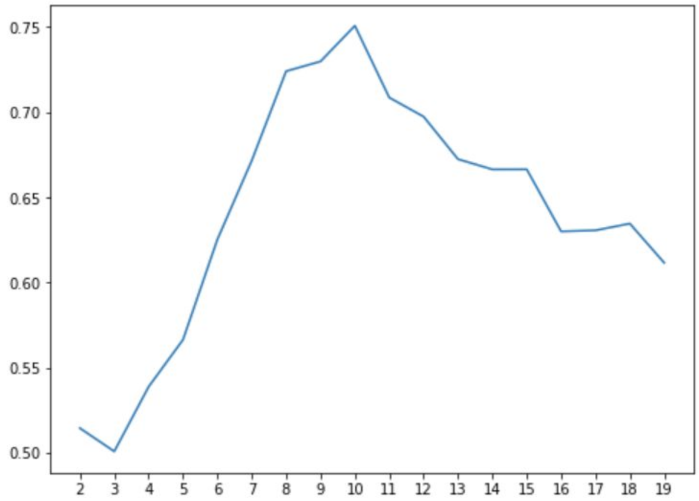


Figure 3.1.1.1: Distribution of Silhouette scores using Method 4

3.1.2. Step 2: Scoring

As mentioned above, scoring methods were created for each social media channel, which reflects different ways that people interact with each channel. For instance, the scoring method for Facebook counts in emotions, while the scoring method for Twitter counts in hashtags and keywords.

The three formulas used are as follows:

$$FacebookScore = \frac{(postInteractions - postComments) - (\frac{1}{2}sad + \frac{1}{2}angry)}{totatPost} \forall totalPost \neq 0$$

$$TwitterScore = \frac{(postInteractions - totalReplies) + (hashtagVolume + keywordVolume + cashtagVolume)}{(tweets * (\frac{totalReplies}{followers})} \forall (tweets, totalReplies, followers) \neq 0$$

$$InstagramScore = \frac{(postInteractions - totalComments)}{mediaCount * (\frac{totalComments}{followedByCount})} \forall (mediaCount, totalComments, followedByCount) \neq 0$$

Most differences among the three formulas can be explained in the features contained in the original data, with the variables coming from the “Rollups” data introduced previously. With the motivation of the score being to measure the raw interactions, it was important to exclude engaged interactions (comments/replies) from the formulas. In the three datasets, the postInterations columns contain both raw and engaged interactions, and for this reason replies/comments were subtracted. “Sad” and “Angry” reactions were subtracted from the Facebook interactions since these were attributes related to sentiment. While these variables are related to “negative” sentiment, it is important to remember that “any publicity is good publicity”, and these reactions still lead to exposure. The decision to include hashtag, keyword and cashtag volume was made after seeing their relative importance in the feature engineering process described in the next section.

In each formula, we divide the interactions by some variation of “posts” multiplied by “engaged interactions” over “followers”. By doing this, we normalize the raw interactions by the number of posts and the number of engaged interactions per follower at that point of time. In this sense, we are scoring a client based on their ability to draw in clicks and subsequent views. Based on an arbitrary social media post, the smaller the fraction of engaged to raw interactions is, the more negative the reaction to the post is. To examine this in depth, take into account the following screengrab from two random tweets. The first tweet was received poorly while the second was not. When social media users see a post they feel negatively about, they are more compelled to comment on the post, rather than liking it.

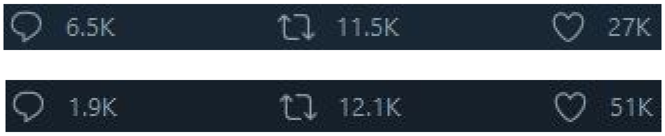


Figure 3.1.2.1: Score Example for Sonic on Facebook over time

A ScoreGenerator module was written in Python to cluster a assign a new brand to an existing cluster, compute a score for that new brand, and create semi-interactive plots using Plotly. Figure 3.1.2.2 shows the score fluctuation for the movie “Sonic the Hedgehog”. The release date of the movie is denoted on each subplot using a vertical black line. Comparing score trends of movies within each cluster would give the clients an overview of how their movies are doing compared to others.

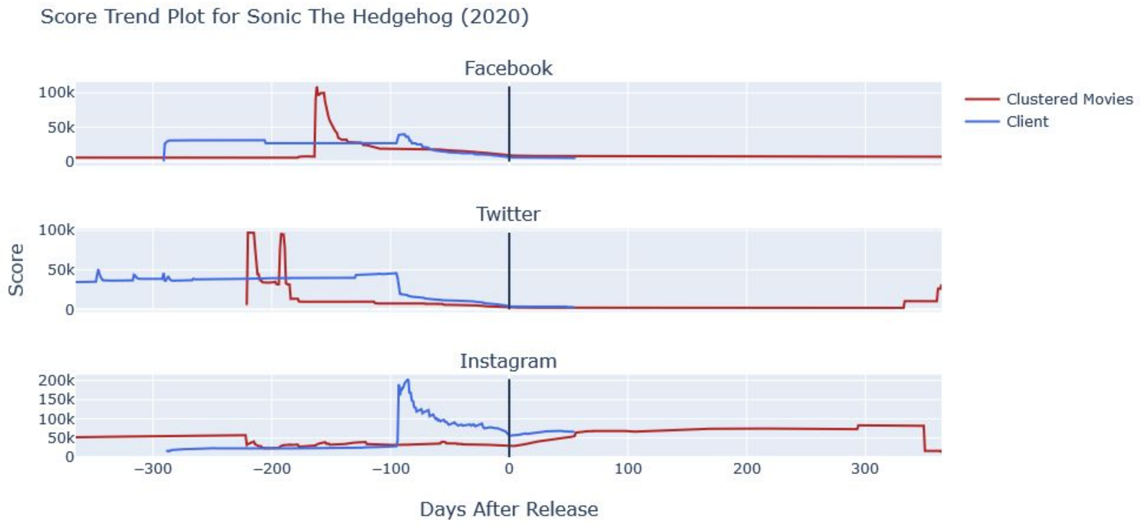


Figure 3.1.2.1: Score Example for Sonic on Facebook over time

3.1.3. Step 3: NLP

After the clustering and score generation process, natural language processing was used to gain deeper insight on social media interactions. In this case, the team decided on both descriptive and statistical analyses of text. Similar to the ScoreGenerator module described above, another module would be written to allow for seamless integration between the two. Clients could see a visualization of their score trends and immediately see visualizations on a given day.

For each specific social media channel and each cluster, the team applied descriptive methods such as word frequency plots, both for the general set of comments and comments split by sentiment (positive, neutral, negative). Additionally, a word cloud could be generated for a given window (since word clouds are generally more insightful with more data, using a window provided optimal results compared to a word cloud based on daily comments). For example, the results of the visualization module for the film “Sonic the Hedgehog” can be seen below:

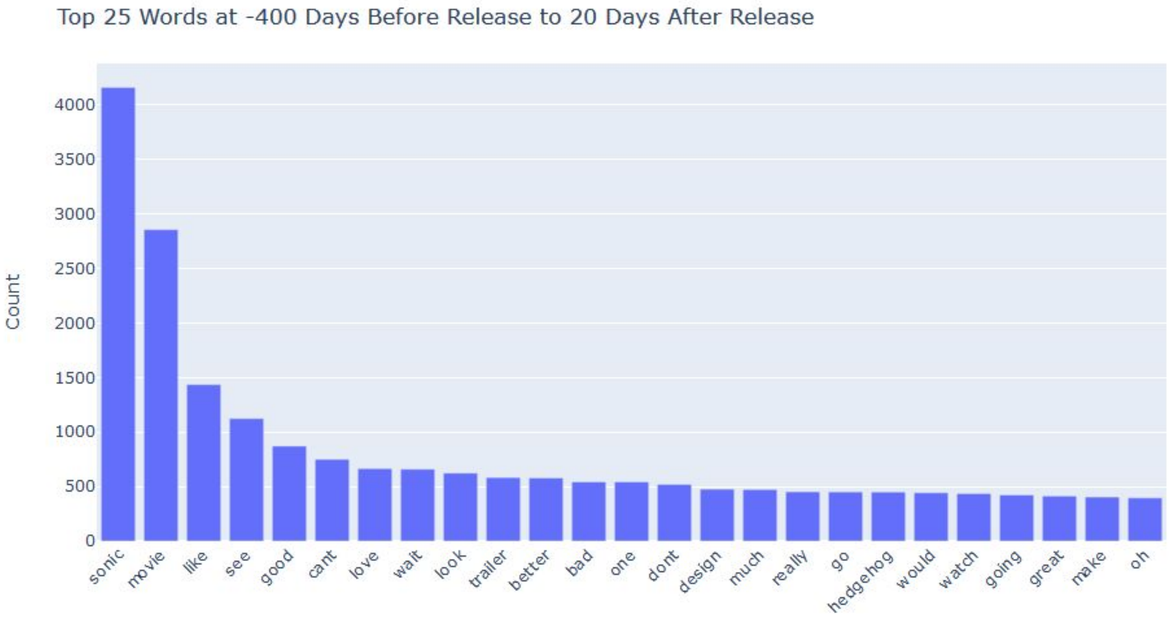


Figure 3.1.3.1: Word Count Plot for "Sonic the Hedgehog"

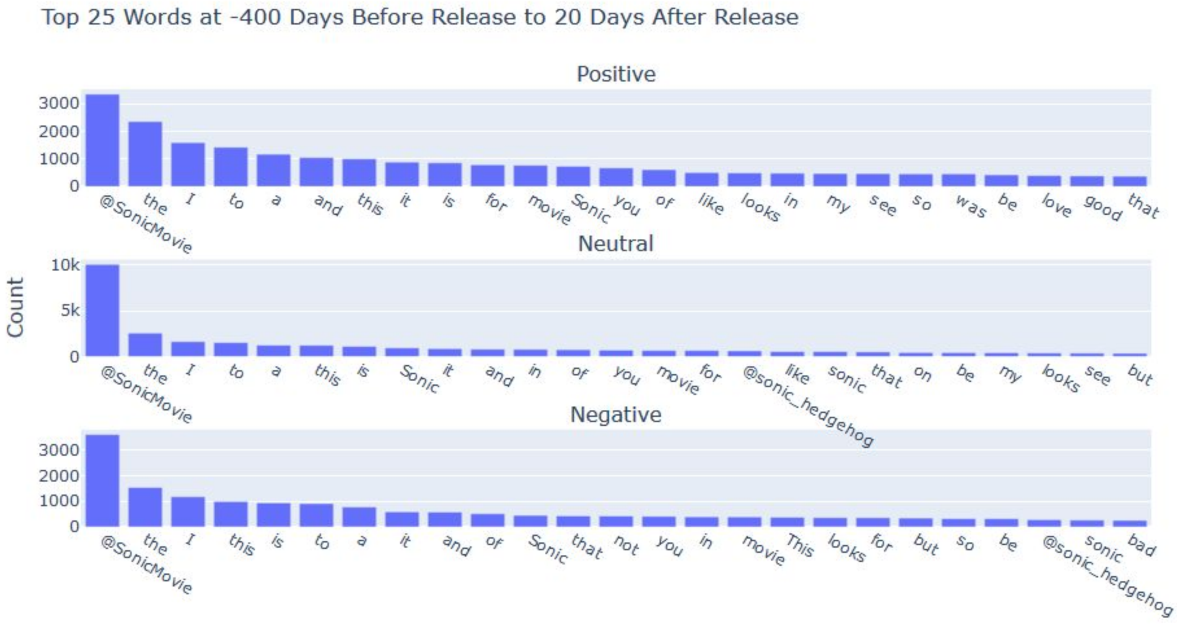


Figure 3.1.3.2: Word Count Plot by Sentiment for "Sonic the Hedgehog"

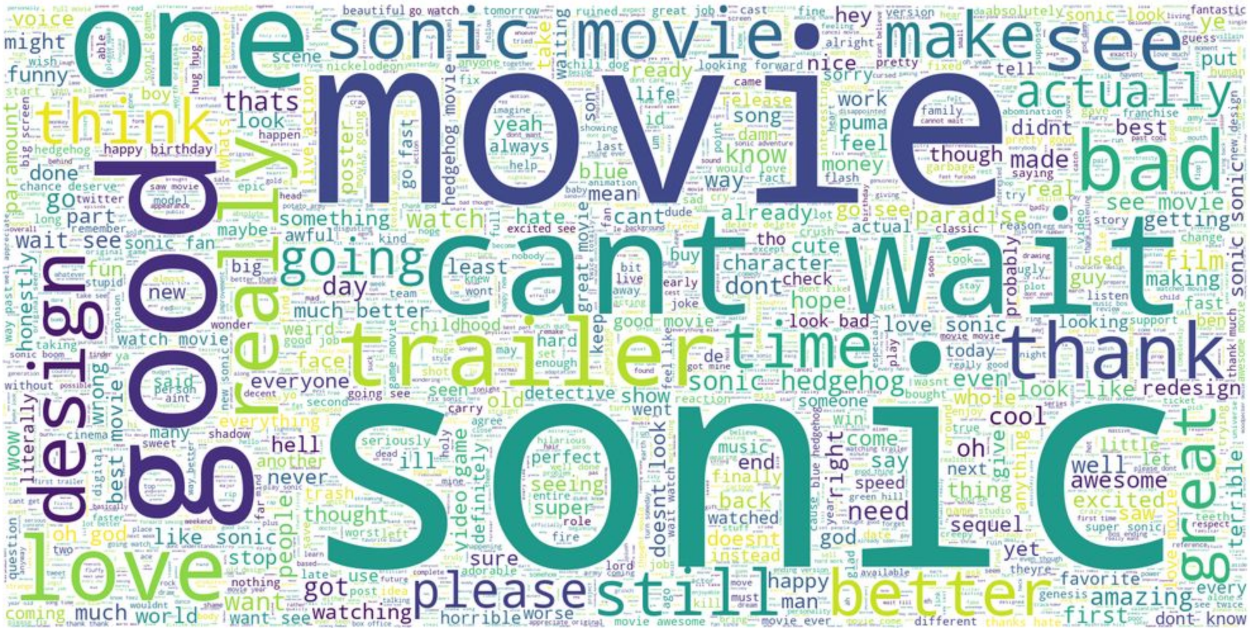
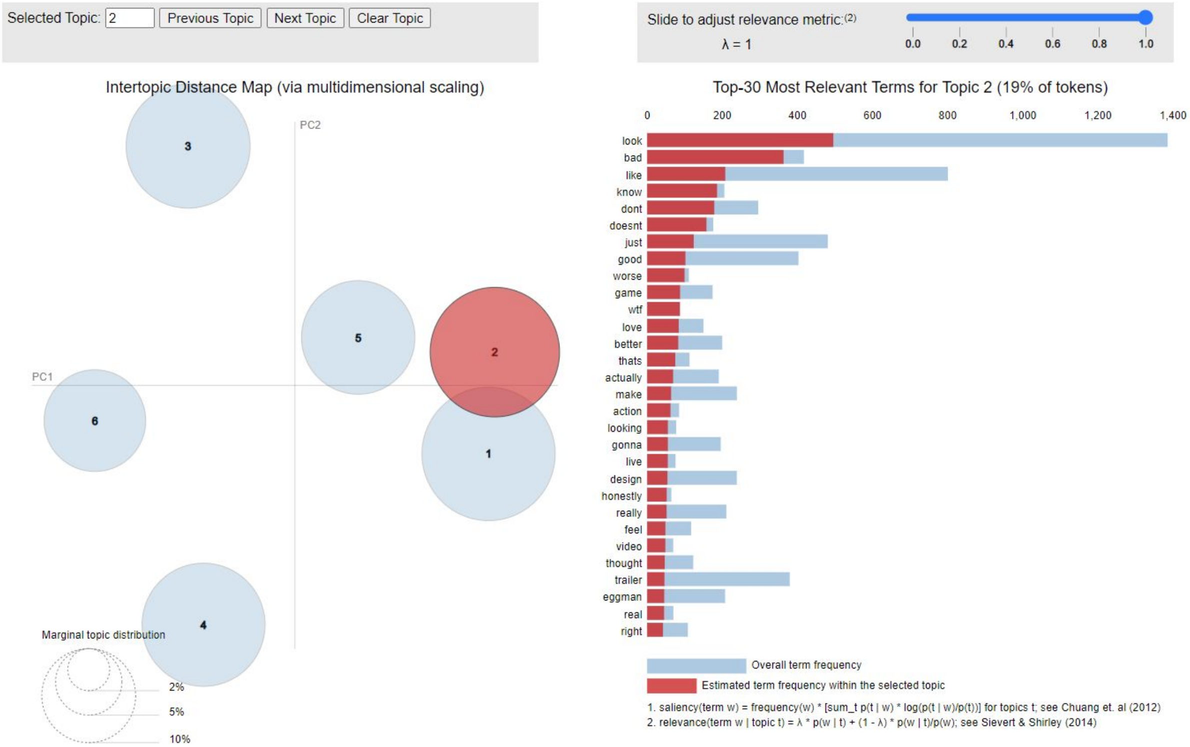


Figure 3.1.3.3: Word Cloud for “Sonic the Hedgehog”

Following an exploratory analysis, topic modeling was performed on the comments. Figure 3.1.3.1 shows the Topic Model of the Sonic movie from 2019-03 to 2019-06-30 on Twitter. Topic Models attempt to extract topics from text by finding word and phrase patterns and clustering them together. We picked this period because it is a good example of how topic models can help companies better understand what people are saying about their product at a certain time. The first Sonic trailer released in April 2019 to the horror of many Sonic fans. Many people were outraged by the art style because it was very different from the original and looked bad. We can see that the topic model caught on to this in topic 2 which specifically discusses the look of the character.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3.1.3.4: Interactive topic model and ngram chart

To use these NLP tools effectively in conjunction with brand score, these charts and topic models can be generated when the score dips or increases to get insights into what is working and what is not. This will be able to effectively advise clients on how to proceed forward and adjust their marketing strategies.

3.1.4. Discussion

By creating a new metric to evaluate social media performances of each movie on different channels, clients would be able to know how their brand is doing compared to movies with similar category, budget, MPAA rating, etc. NLP, Topic Modeling and Sentiment Analysis were performed to further diagnose the reasons of score fluctuations. Using this method in conjunction with the NLP tools, clients would be able to adjust social media strategies to increase brand awareness and enable better marketing campaigns. Some clients can even increase their box office sales by adjusting certain aspects of the movies to meet the needs of the public.

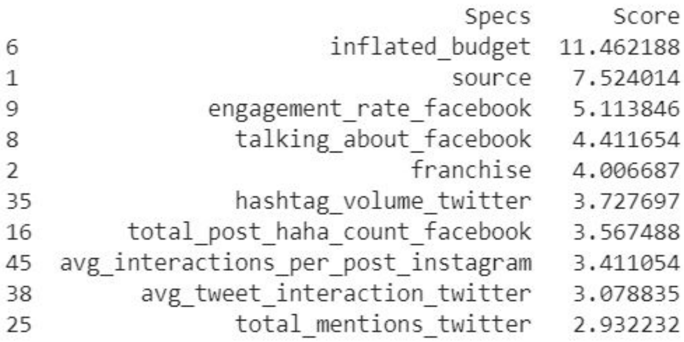
3.2 Movie Success Prediction

After addressing the exploratory aspects of social media strategy evaluation, this section of the report will cover the predictive methods the team employed to attempt to give clients a reliable prediction of how their product will perform in the market, given the current trends of score. With the data involving only movie brands, the team decided to use box office data to assign classes to each film. If a film's box office profit was greater than its budget, it was labeled as a success. If the budget was greater than the profit, the film was labeled as a "flop". With only 526 brands available to the team, a binary classification provided the most realistic predictive approach. As mentioned above, the models the team chose were split between normal statistical models, and a time-based RNN approach. The team used similar feature engineering and feature selection methods to approach the models.

3.2.1 Step 1: General Models

Before applying models, the first step is to determine what features are important to the final "success/flop" label, so that the models aren't polluted by useless features that would harm performance. To aggregate the years of data, we took the mean value of each feature across social media for each movie. After aggregation we also added in the categorical and numerical movie attribute features bringing the total number of features to 89.

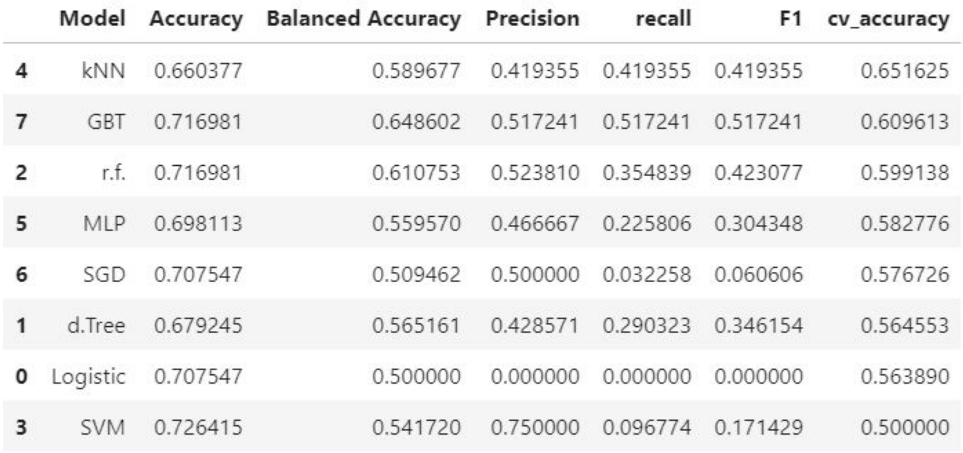
The team applied different feature selection methods such as ANOVA F-value, KS Score, manually selecting features, and Recursive Feature Elimination. These methods were tested and their effect on model performance on a set of only numerical features as well as a set of both numerical and categorical features. In the end, the team found that the ANOVA F-value top 10 numerical (without daily metrics) and categorical features performed the best. ANOVA or analysis of variance uses the F-test to check if the means between several groups are different. In this case it is comparing the distribution of the concerned feature between the label variable "success/flop". The F test measures the between group variance over the within group variance, meaning that the greater the F value, the more difference there is between the two groups, which would help us differentiate between successes and failures. The figure below shows the top features produced by using ANOVA F-value.



```
                                   Specs      Score
6                        inflated_budget  11.462188
1                                 source   7.524014
9                engagement_rate_facebook   5.113846
8                  talking_about_facebook   4.411654
2                              franchise   4.006687
35                  hashtag_volume_twitter   3.727697
16          total_post_haha_count_facebook   3.567488
45     avg_interactions_per_post_instagram   3.411054
38             avg_tweet_interaction_twitter   3.078835
25                    total_mentions_twitter   2.932232
```

Figure 3.2.1.1 Top 10 Numerical and Categorical Features from ANOVA method

By only using these variables and normalizing the numerical variables by themselves, we got the following performance.



	Model	Accuracy	Balanced Accuracy	Precision	recall	F1	cv_accuracy
4	kNN	0.660377	0.589677	0.419355	0.419355	0.419355	0.651625
7	GBT	0.716981	0.648602	0.517241	0.517241	0.517241	0.609613
2	r.f.	0.716981	0.610753	0.523810	0.354839	0.423077	0.599138
5	MLP	0.698113	0.559570	0.466667	0.225806	0.304348	0.582776
6	SGD	0.707547	0.509462	0.500000	0.032258	0.060606	0.576726
1	d.Tree	0.679245	0.565161	0.428571	0.290323	0.346154	0.564553
0	Logistic	0.707547	0.500000	0.000000	0.000000	0.000000	0.563890
3	SVM	0.726415	0.541720	0.750000	0.096774	0.171429	0.500000

Figure 3.2.1.2 Model Performance based on Normalized Top 10 Numerical Features

The model with the best test set performance was the Gradient Boosted Trees model. Although it has a lower Cross validation mean score than K-Nearest Neighbors, GBT has a much better performance on the test set. With such a small sample size it is possible the KNN model only performs very well on the training data and performs poorly on any new data. For this reason, we chose the Gradient Boosted Trees model to be the best "General Model".

Gradient Boosted Trees is an ensemble method using decision trees. First, the model fits an initial “weak learner” decision tree model to the data. Then another model is fitted, focusing on improving the predictions the first model mislabeled. Gradient boosting determines the weaknesses of the previous model using the gradient of the loss function. In this case the loss function is a measure of how good/bad that model is at predicting movie success/flop. This goes on for a set number of rounds called boosting rounds. The idea is that by combining these models that improve on the errors made by the last model, we can get an ensemble model that predicts well on the dataset. Each tree in a GBT model is called a “weak learner” in the sense that the tree performs poorly on its own, but a chain of many weak learners improving based on the errors of the previous tree produces notably good results.For this dataset, we found the following setup the most effective in predicting success/flop:

Parameters	**Gradient Boosted Classifier**
Number of Boosting Rounds	140
Max Depth of Trees	14
Learning Rate	0.2
Metric	Balanced Accuracy

Table 3.2.1.3 GBT Parameters

Using these parameters we got a balanced accuracy of **64.9%** and overall accuracy of **71.7%** on the test set with the details of the confusion matrix shown below in Figure 3.2.1.4:

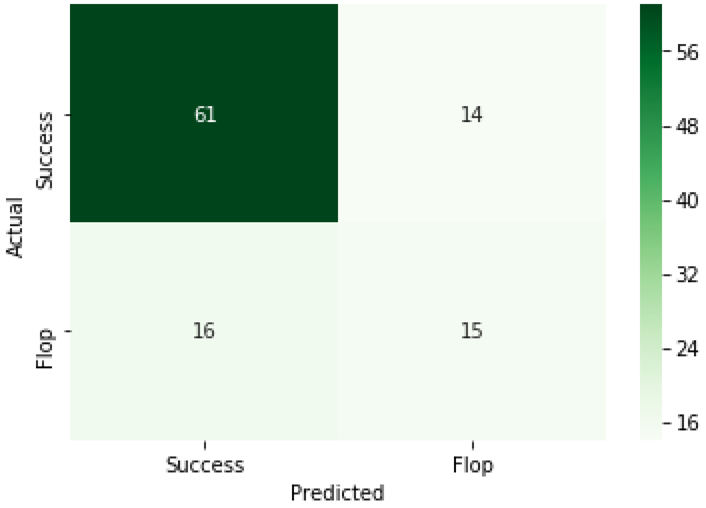


Figure 3.2.1.3 GBT Confusion Matrix

3.2.2 RNN

While using a simple MLP classifier had been explored in the previous section, the existence of time-based data allows for the use of Recurrent Neural Networks (RNNs) for classifying successes or flops. RNNs aim to address one of the more glaring issues of neural networks - fixed-size input and output. While vanilla neural networks are incredibly powerful for classification, RNNs allow for learning based on sequences of data. The underlying principle of RNNs is a “state”, or the ability for layers to remember information about the input as the neural network moves through the sequence of data. The most popular applications of RNNs have been text and image based learning/classification, however, as shown below, the different combinations of RNNs make them flexible for a variety of use cases.

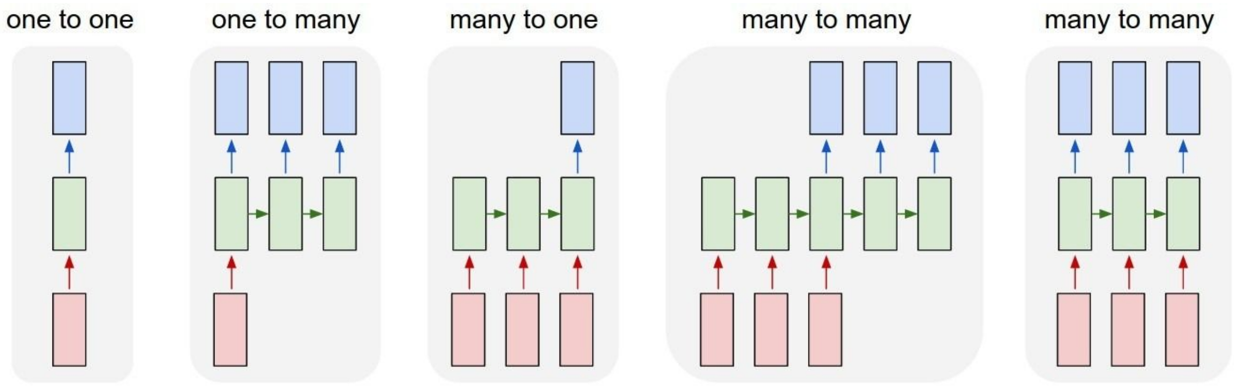


Figure 3.2.2.1 RNN Architectures[5]

Among the four images above (ignoring “one to one” as it follows the architecture of a typical neural network), our data closely followed the “many to one” architecture: many days before the release of a movie leading to a singular box office value. As it stands, the two most popular RNN “types” are Long Term Short Memory (LSTM) and Gated Recurrent Units (GRU). The main difference between the two lies in the simpler structure of GRU models, making their computation time much faster. For the purpose of this project, both types of RNNs were attempted.

The features of the RNN were adapted from the results of the above feature selection. Since the RNN involves a sequence of data, it would not make sense to use inflated_budget. Also, some features (total_post_comments_facebook and total_post_haha_facebook) had negative values which interfered with the log transformation. For these reasons, the final set of features selected is presented in the table below:

Facebook	**Twitter**	**Instagram**
talking_about	hashtag_volume	avg_intractions
engagement_rate	total_mentions	
	avg_tweet_interaction	

Table 3.2.2.1 RNN Features

As with the general machine learning models that the team attempted, the problem of a small training dataset persisted with RNNs. While the volume of data was larger, as transforming the input data to a sequence creates data with larger dimensions, the fact that only 300-400 (depending on train/test split) samples were available for training severely limited the effectiveness of both kinds of RNNs (LSTM and GRU). After training and evaluating both kinds

[5] https://karpathy.github.io/2015/05/21/rnn-effectiveness/

of RNN models, the following confusion matrices were achieved using the following hyperparameters.

Parameters	**LSTM**	**GRU**
Layers	1	4
Nodes	100	50 -> 20 -> 10 -> 5
Optimizer	Stochastic Gradient Descent	Stochastic Gradient Descent
Learning Rate	0.00001	0.00001
Metric	Accuracy	Accuracy

Table 3.3.2.2 RNN Parameters

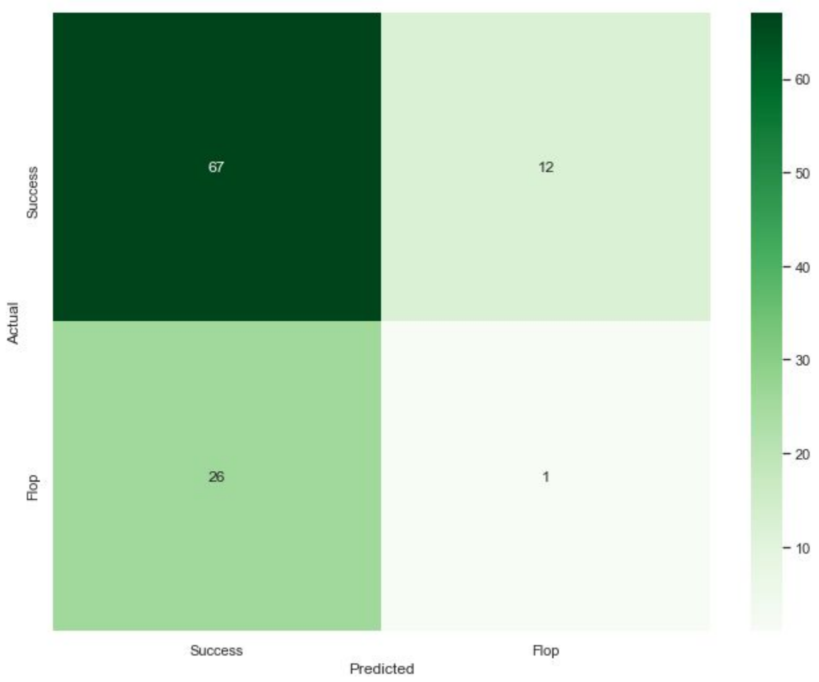


Figure 3.2.2.2 LSTM Confusion Matrix

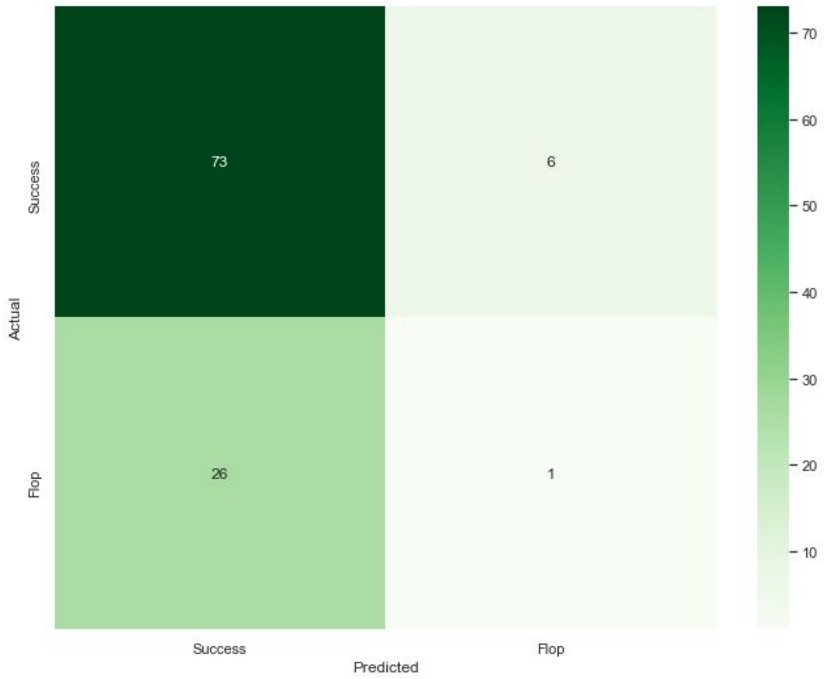


Figure 3.2.2.3 GRU Confusion Matrix

With the fast computation time of the GRU model, we can add more layers while adding more nodes to the LSTM model. As shown in the confusion matrices above, the model performance is not much different and this can be expected due to their similar structure "under-the-hood". However, the issue with both models seems to be the large number of "Success" predictions. Overall, the LSTM and GRU models achieved **73%** accuracy and **68%** accuracy on the test set, respectively. Judging by the low amount of samples, it can be inferred that the model was not able to generalize the variations among the samples and predicted "Success" for almost all movies. Since the LSTM model predicted more "Flop" movies, it can be said that this model is more optimal. However, before its deployment, training on more samples is necessary.

3.2.2 Discussion

After comparing both General Models and RNN models, the team chose the General Gradient Boosting Classifier as the final model. Unlike the RNN models, the team was able to optimize model search by balanced accuracy instead of accuracy. Because the available dataset is slightly imbalanced in favor of successful movies, optimizing by accuracy has the danger of choosing a model that essentially only predicts one value. As shown in Figures 3.2.2.2 and 3.2.2.3, the RNN models essentially only predict success and have a decent accuracy. The regular model hyperparameters were optimized based on balanced accuracy, which is the average of the individual class accuracies, and allows for a more balanced outcome. As shown in Figure 3.2.1.3, the GBT model is better able to differentiate between successes and flops. Since the more simplistic and understandable GBT model has better performance, this is the final model we chose.

The team also wanted to better understand the impact of features on the probability that the movie would be a flop through Partial Dependence Plots (PDP). Figure 3.2.2.1 shows the PDP plot for the Gradient Boosted Trees Model.

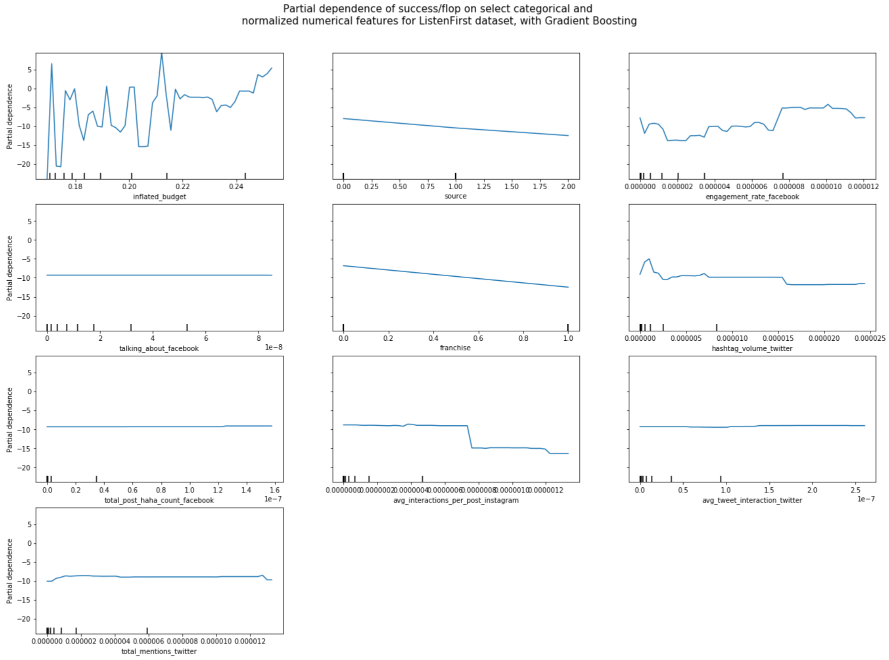


Figure 3.2.3.1 Partial Dependence Plots for the 10 selected features on GBT model

A higher inflated budget seems to be trending towards a higher probability of flop, however, the trend is quite volatile. Source, where 0 is Not Original, 1 is Original, and 2 is Remake, seems to have a slightly lower chance of flop when Original or Remake. If a movie is franchised it also helps decrease the chances of flop. The last trend worthy of mentioning is the average interactions per instagram post, past a certain amount the chances of flop.

3.3. Aspect-Based Opinion Mining

Aspect-based opinion mining (ABOM) can be interpreted as a multiclass classification problem that is used to classify comments into categories. ‘Aspects’, in this instance, represents a type of category that the comment falls under. For Listenfirst, allowing clients to understand certain aspects of their marketing campaign at a glance, see the sentiment and understand audience feedback in these aspects, would be a huge value add and allow the clients to make targeted changes to their marketing campaigns.

This solution requires a dataset of comments and aspects to be trained on. The datasets provided by ListenFirst only had comments, and no predetermined labels. Hence, the following demonstration of how such a model can be implemented was done with a subset of comments from the franchise ‘John Wick’ that the team manually labelled.

3.3.1. Step 1: Label

The first step involves creating a dataset to train a model on. Comments were put in one of 5 aspects, as seen in Table 3.3.1.1. A total of 500 comments were labelled.

Aspect	**Example**
MOVIE#GENERAL	shouldve trailer replica hope soon
MOVIE#CHARACTERS	rafi look like bad as
SOCIAL#HASHTAGS	#johnwick3
SOCIAL#MENTIONS	@manabyte awesome @russobrothers @marvelstudios

OTHERS	do you change my mind to love cat

Table 3.3.1.1: Aspects and Examples

Comments were pre-processed and cleaned, along with lemmatizing and removing stop words.

3.3.2. Step 2: Train

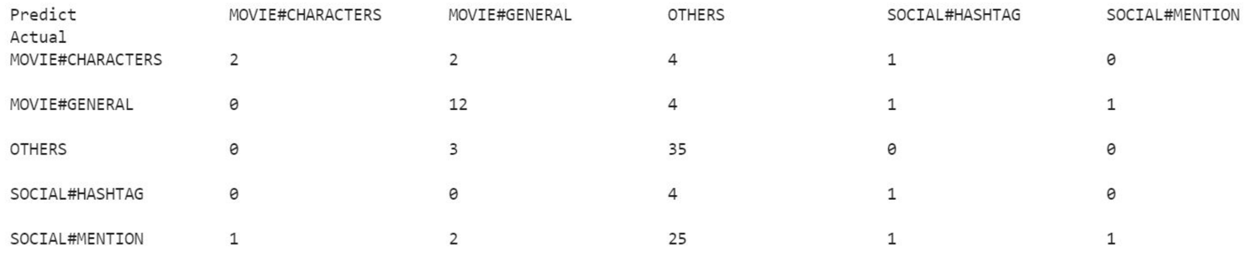
The model was trained with a four layer sequential neural network. This model works by tokenizing each comment and creating a vector of words. Each comment is represented in this vector of words in a one-hot encoding. The dataset was split 80-20 train and test set. 5-fold cross validation was conducted with the best parameters saved. Parameters were tuned and the list of tuned parameters can be found in Table 3.3.2.1.

Parameter	**Selection**
No. of Layers	4
Vocab Size (number of different words expected)	6000
No. Nodes Per Layer	Layer 1: 512 Layer 2: 256 Layer 3: 128 Layer 4: 5 (output based on number of aspects)
Activation Function	Layer 1: relu Layer 2: relu Layer 3: relu Layer 4: softmax
Optimizer	Adam

Table 3.3.2.1: Tuned Parameters

Early stopping was implemented as well with patience at 10 epochs.

The accuracy for this model was 0.9950 on the training set and 0.51 on the test set. Figure 3.3.2.2 shows the confusion matrix for the model.



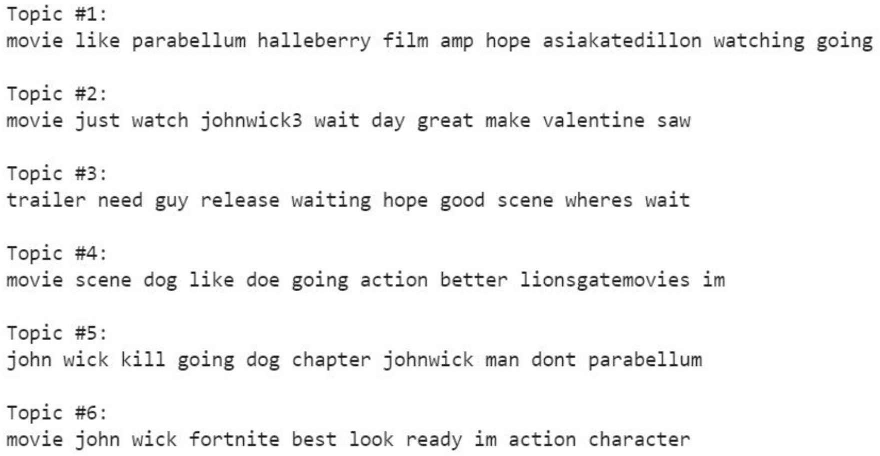
```
Predict           MOVIE#CHARACTERS    MOVIE#GENERAL    OTHERS        SOCIAL#HASHTAG    SOCIAL#MENTION
Actual
MOVIE#CHARACTERS  2                   2                4             1                 0

MOVIE#GENERAL     0                   12               4             1                 1

OTHERS            0                   3                35            0                 0

SOCIAL#HASHTAG    0                   0                4             1                 0

SOCIAL#MENTION    1                   2                25            1                 1
```

Figure 3.3.2.2: Confusion Matrix

3.3.3. Step 3: Sentiment Analysis and Topic Model

The model was then used to predict the aspects for the movie “John Wick: Chapter 3 – Parabellum” comments. After identifying the aspects, sentiment analysis and topic modeling was run on each aspect. Sentiment analysis was run with the ‘Textblob’ package and an LDA model was used to run a topic model. As a sample, this section will examine the section MOVIE#GENERAL.

The sentiment was generally positive at 0.141. Figure 3.3.3.1 shows 5 topics that were extracted from the LDA model.



```
Topic #1:
movie like parabellum halleberry film amp hope asiakatedillon watching going

Topic #2:
movie just watch johnwick3 wait day great make valentine saw

Topic #3:
trailer need guy release waiting hope good scene wheres wait

Topic #4:
movie scene dog like doe going action better lionsgatemovies im

Topic #5:
john wick kill going dog chapter johnwick man dont parabellum

Topic #6:
movie john wick fortnite best look ready im action character
```

Figure 3.3.3.1: Topic Model

From the 6 topics above, the Topics 3, 5 and 6 can be explored further. For example, Topic 3 suggests that people are waiting for the release of a trailer and looking forward to good scenes. After a quick Google search, it can be inferred that Topic 5 refers to an emotional scene involving the main character’s dog. Topic 6 refers to the introduction of John Wick as a playable character in the game Fortnite, and seems to suggest a positive reaction to his appearance. Even a quick exploration of 3 topics reveals different directions that the marketing team for this particular movie can take.

3.3.4. Discussion

Given the small dataset with multiclass labels, an accuracy of 0.51 is acceptable as the main emphasis on the model is the ability to run sentiment and topic models on aspects. Having a more accurate ABOM model will allow for more nuanced and targeted approaches. To further improve this model, there is a need to train this on a larger dataset.

4. Implementation

The true worth of these tools is realized when they are used synergistically. Figure 4.1 illustrates how these tools can be used together in a pipeline.

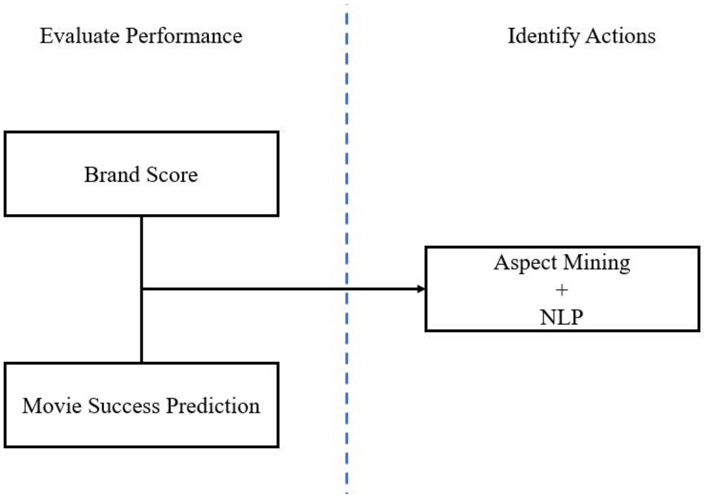


Figure 4.1: Implementation Pipeline

Looking at trends of brand score along with the results of movie success prediction will be able to advise clients on whether the movie, based on the trends, is going to perform well or poorly. If these tools reflect a negative performance, actions can be taken to identify what follow up actions to take via aspect mining, followed by a NLP drill down based on aspects and time. By identifying what is going well and what is not based on what the audience is saying, clients can be reactive and adjust their marketing strategies to respond to the audience's needs. Further, this can be done even for movies that have shown good performance to further capitalize on good progress. For instance, if the audience is responding well and requesting for merchandise like toys, clients could possibly think of releasing merchandise to tie-in with the release of the movie.

5. Conclusion

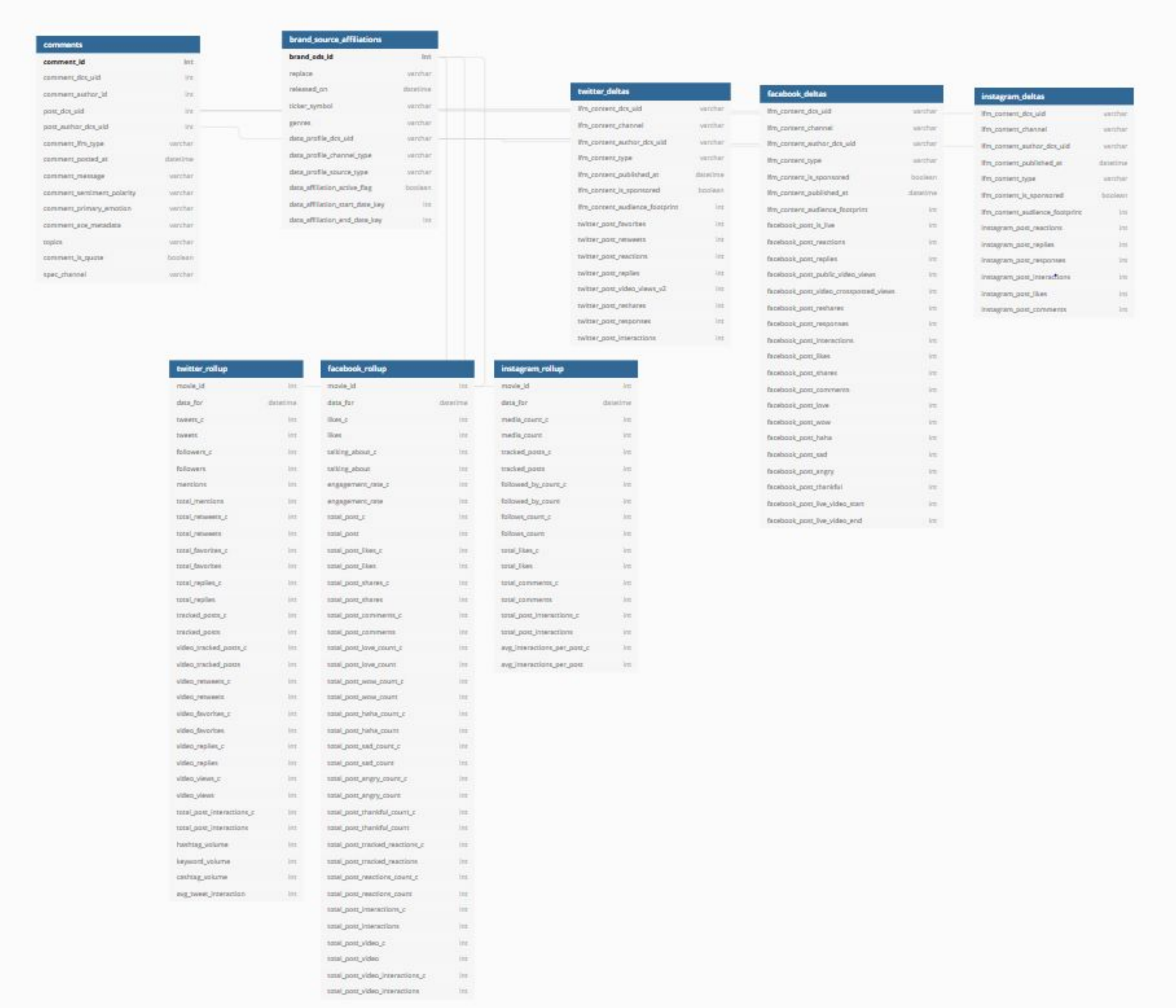
Through an in depth exploration of social media data provided by Listenfirst, the team was able to create a unique, two-fold approach to help clients better understand the information available on social media platforms. As discussed earlier, social media provides a streamlined approach for clients to reach their customers, and allows clients to observe a raw, unfiltered view of how their customers discuss products. While the interactions on social media platforms can never be taken at face value, through the use of careful analytical methods as presented above, customer interactions can be aggregated and presented in a clean and efficient manner

The team’s first object was to create a brand scoring method that would allow clients to evaluate their social media performance at a glance. With a wide range of movies available to the team, an overall comparison could result in vastly different movies with regards to attributes budget, release year, and genre (among others) being compared against each other. In order to address this point, the team pursued unsupervised clustering algorithms. Through an exploration of four different clustering approaches, the team settled on an approach involving transformation of data using Factor Analysis of Mixed Data (FAMD) followed by Agglomerative clustering. We found that this combination of linear transformation to a clustering algorithm provided the team with optimal results after a comparison among the four clustering approaches chosen. The optimal hyperparameters for both the FAMD algorithm and Agglomerative clustering were chosen after evaluation of model metrics. After clustering similar movies, brand score formulas were applied to the cluster to find the average score within each cluster, aggregated by days before/after release. Finally, the client using this approach has the ability to explore comments about their movie over a range of days, in order to understand the fluctuations in their score.

After creating the brand score, the team began to approach the creation of predictive modeling. The team hoped to find a correlation between social media interactions and a movie’s profit/success, and as shown in the introduction and ROI section, this was found to be true. The team used a multitude of models, ranging from linear models, to neural networks, and to recurrent neural networks. While the team found promising results, the models were handicapped due to a lack of data. As shown above, the model with the best performance happened to be a boosted tree. The team concluded that training a model on a larger set of brands would be necessary before considering deployment.

The second part of the team's modeling approach was the creation of an aspect mining model. As shown in Table 3.3.1.1, aspect mining is a laborious process, and for this reason, a set of 500 comments was labeled. The aspect mining results paired with a topic model (as shown in Figure 3.3.3.1), show the potential of using such an approach on social media comments. The six topics shown reveal distinct topics of discussion among consumers of the film series "John Wick", and provide avenues of product development for clients.

Appendix A - Data Relation Table



comments	
comment_id	int
comment_dcs_uid	int
comment_author_id	int
post_dcs_uid	int
post_author_dcs_uid	int
comment_lfm_type	varchar
comment_posted_at	datetime
comment_message	varchar
comment_sentiment_polarity	varchar
comment_primary_emotion	varchar
comment_ace_metadata	varchar
topics	varchar
comment_is_quote	boolean
spec_channel	varchar

brand_source_affiliations	
brand_ods_id	int
replace	varchar
released_on	datetime
ticker_symbol	varchar
genres	varchar
data_profile_dcs_uid	varchar
data_profile_channel_type	varchar
data_profile_source_type	varchar
data_affiliation_active_flag	boolean
data_affiliation_start_date_key	int
data_affiliation_end_date_key	int

twitter_deltas	
lfm_content_dcs_uid	varchar
lfm_content_channel	varchar
lfm_content_author_dcs_uid	varchar
lfm_content_type	varchar
lfm_content_published_at	datetime
lfm_content_is_sponsored	boolean
lfm_content_audience_footprint	int
twitter_post_favorites	int
twitter_post_retweets	int
twitter_post_reactions	int
twitter_post_replies	int
twitter_post_video_views_v2	int
twitter_post_reshares	int
twitter_post_responses	int
twitter_post_interactions	int

facebook_deltas	
lfm_content_dcs_uid	varchar
lfm_content_channel	varchar
lfm_content_author_dcs_uid	varchar
lfm_content_type	varchar
lfm_content_is_sponsored	boolean
lfm_content_published_at	datetime
lfm_content_audience_footprint	int
facebook_post_is_live	int
facebook_post_reactions	int
facebook_post_replies	int
facebook_post_public_video_views	int
facebook_post_video_crossposted_views	int
facebook_post_reshares	int
facebook_post_responses	int
facebook_post_interactions	int
facebook_post_likes	int
facebook_post_shares	int
facebook_post_comments	int
facebook_post_love	int
facebook_post_wow	int
facebook_post_haha	int
facebook_post_sad	int
facebook_post_angry	int
facebook_post_thankful	int
facebook_post_live_video_start	int
facebook_post_live_video_end	int

instagram_deltas	
lfm_content_dcs_uid	varchar
lfm_content_channel	varchar
lfm_content_author_dcs_uid	varchar
lfm_content_published_at	datetime
lfm_content_type	varchar
lfm_content_is_sponsored	boolean
lfm_content_audience_footprint	int
instagram_post_reactions	int
instagram_post_replies	int
instagram_post_responses	int
instagram_post_interactions	int
instagram_post_likes	int
instagram_post_comments	int

twitter_rollup	
movie_id	int
data_for	datetime
tweets_c	int
tweets	int
followers_c	int
followers	int
mentions	int
total_mentions	int
total_retweets_c	int
total_retweets	int
total_favorites_c	int
total_favorites	int
total_replies_c	int
total_replies	int
tracked_posts_c	int
tracked_posts	int
video_tracked_posts_c	int
video_tracked_posts	int
video_retweets_c	int
video_retweets	int
video_favorites_c	int
video_favorites	int
video_replies_c	int
video_replies	int
video_views_c	int
video_views	int
total_post_interactions_c	int
total_post_interactions	int
hashtag_volume	int
keyword_volume	int
cashtag_volume	int
avg_tweet_interaction	int

facebook_rollup	
movie_id	int
data_for	datetime
likes_c	int
likes	int
talking_about_c	int
talking_about	int
engagement_rate_c	int
engagement_rate	int
total_post_c	int
total_post	int
total_post_likes_c	int
total_post_likes	int
total_post_shares_c	int
total_post_shares	int
total_post_comments_c	int
total_post_comments	int
total_post_love_count_c	int
total_post_love_count	int
total_post_wow_count_c	int
total_post_wow_count	int
total_post_haha_count_c	int
total_post_haha_count	int
total_post_sad_count_c	int
total_post_sad_count	int
total_post_angry_count_c	int
total_post_angry_count	int
total_post_thankful_count_c	int
total_post_thankful_count	int
total_post_tracked_reactions_c	int
total_post_tracked_reactions	int
total_post_reactions_count_c	int
total_post_reactions_count	int
total_post_interactions_c	int
total_post_interactions	int
total_post_video_c	int
total_post_video	int
total_post_video_interactions_c	int
total_post_video_interactions	int

instagram_rollup	
movie_id	int
data_for	datetime
media_count_c	int
media_count	int
tracked_posts_c	int
tracked_posts	int
followed_by_count_c	int
followed_by_count	int
follows_count_c	int
follows_count	int
total_likes_c	int
total_likes	int
total_comments_c	int
total_comments	int
total_post_interactions_c	int
total_post_interactions	int
avg_interactions_per_post_c	int
avg_interactions_per_post	int