Film Script Emotion Visualization - FINAL REPORT

I. Introduction - Motivation

A. Screenwriters, producers, critics and even viewers have an interest in creating great movies. But what does a "good film" resemble? There are many ways to assess a work, but at the end of the day people enjoy fiction because it evokes feeling. If we could imagine a movie as a composite of emotions over its runtime, we could gain insight into the general structures of successful movies. By comparing a given script's structure to the best-performing films, those in the industry would be able to fine-tune their projects by evaluating the mood makeup of their scripts, which could prove especially useful to executives choosing which scripts to greenlight.

II. Problem definition

A. We aim to visualize a given film by the emotions its script displays on a scene-to-scene basis. Additionally, we hope to show what some of the "best" movies look like in terms of this mood structure in terms of profitability and critical reception.

III. Survey

A. Some studies have tried extracting emotions from text, while others have analyzed the process of green-lighting scripts, trying to predict films' profits (Eliashberg et. al, "From story"). No research we found combines these themes and analyzed film mood to inform greenlighting efforts. Many efforts to aid greenlighting decisions did not analyze their scripts greatly, so we utilize natural language processing to better aid these efforts.

1. Sharda used a neural network to predict a film's profitability (243), and Ghiassi improved on this by including features such as the MPAA rating, production budget, runtime, and release date (3188). Lash and Zhao leveraged hybrid features, predicting movie profit using LASSO. Additional important predictors included the actor/director/genre profits (1), while advertising expenditures did not predict as strongly. Eliashberg analyzed scripts directly and used genre, storyline, and word frequency features ("Assessing" 2639) in a kernel-based approach. Their model identified early exposition, budget, a strong nemesis as strong predictors. Finally, Hunter found that network text analysis for screenplays was a more significant predictor than any other single factor (5). These articles are useful in identifying useful algorithms and predictors of movie success, as well as data sources. However, they often use a shooting script

instead of the greenlighting script, and frequently have small sample sizes. They also tend to use high-level features, making it difficult to use them for greenlighting decisions. Our approach addresses this concern by focusing on script structure analysis.

2. Strapparava and Giannakopoulos's studies focus on identifying emotions in text, scoring texts in six different emotions. Giannakopoulos adapts his "emotion wheel" method to movie scripts, which include more dialogue than standard text. Wang analyzes films shot-by-shot for their affective significance (689), and these principles apply to any cinematic analysis, though he does not limit himself to textual interpretation. Having two different approaches to finding the mood in a script will be helpful, eliminating mistakes that could occur with the awareness of only one.

3. For information outside of script analysis, we considered Aditya's work on developing vision-based scene description graphs (1). Although we are analyzing moods from the script, we could consider extracting graphs from film frames to bolster our algorithm. However, the authors use a deep learning vision system in their paper, requiring additional time and processing power.

4. Ignatov's paper introduces an algorithm that returns a list of recommended movies based on the previous ratings of the user and other users (1). This could be useful in mapping the mood map structures between two similar movies through user ratings.

5. Many movies are also adapted from books. We could use additional mood information from inspiring books so long as match the context and dialogue. Zhu aligned scenes to extract information about the characters and the setting. By using their system, we can add context to our script and augment our mood mapping algorithm for applicable movies.

6. It is important to understand a 'successful' film: from the perspective of the box office, critics, and the audience. Boor's study gives us further insight into the differences in these elements' perception of movies.

IV. Proposed method

A. Intuition - why should it be better than the state of the art?

1. No current tool exists to visualize the emotions of a movie over its runtime. This would be an easy-to-understand, useful point of information to someone trying to evaluate a film. Instead of relying on a third party's description, an executive can quickly get a grasp of the general layout of a movie by plotting its script's emotions.

B. Description of your approaches: algorithms, user interfaces, etc.

1. Using natural language processing and random forest regression, we score the affective meaning of text in the six generally-recognized basic emotions (happiness, sadness, anger, fear, surprise, disgust). Training an NLP algorithm for such a task requires labelled data, however, and performs best when this data is similar in format to the test data.
2. Inspecting a script on a scene-by-scene basis, we assign scenes scores in the emotional categories. These scores will be represented in proportion to each other as sections of a stacked bar chart.
3. Our web app prompts the user to select a visualization of some of the top movies' scripts, selected from the IMSDB by the highest-grossing movie list on Wikipedia and the highest-rated list on Metacritic (critical reception). In addition, we will also allow selection of some poorly-received films, and averages of our script dataset to examine overall trends across the industry. These graphs will examine relative emotion strengths within the scene.

C. List of innovations

1. We are training our algorithm on labelled data that we hope resembles the dialogue-heavy format of scripts (discussed in IV.B.1). This includes the SemEval news headline dataset, the CrowdFlower prose dataset, and the Potter classic fairytale dataset - by utilizing multiple related datasets, we hope to be able to generalize our algorithm to the script space.
2. Analyzing a movie based on sentiment from scene to scene has not been done. We analyze each sentence individually based on its emotion scores (discussed in IV.B.2), and will be able to provide a meaningful movie arc for an emotion category by aggregating the emotion scores of sentences within a scene.

V. Experiments/ Evaluation

A. Description of your testbed; list of questions your experiments are designed to answer:

1. How do we best pull emotion from text?
2. What training data will work best for scripts?
3. Are there common themes to successful movies?

4. Do movies with different structures tend to perform differently?
5. Do different structures perform better by one metric and not others?
6. How can we distinguish between movies with different characteristics?
7. Are our analyses affected by the distinction between displayed and conveyed emotion? Do our structures cope with subtlety, sarcasm, and irony?
8. How can our analyses be improved?

B. Details of the experiments: emotion strength prediction
   1. Sentences/phrases were gathered from 3 datasets and consolidated into a SQLite database
      a) Affective Text dataset
         (1) Maps 1250 headlines to strength of 6 emotions: anger, disgust, fear, joy, sadness, surprise
      b) Potter storybook dataset
         (1) Maps 1946 storybook sentences from 19 books to 8 emotions. Each sentence is classified into an emotion by 4 voters.
         (2) Emotions outside of the 6 from dataset 1 are ignored, and each vote for a label contributes 25 points to its strength to restrict the range in emotion strengths to (0, 100).
      c) CrowdFlower prose dataset
         (1) Maps 2490 sentences to 18 Plutchik emotions. Each sentence is classified into an emotion by a number of voters.
         (2) Emotions found in the Affective Text dataset are given 2 points. Other emotions provide 1 point to the emotion(s) they are closest to in Plutchik's wheel of emotion. Each sentence's emotion votes are then divided by (2*number of people voting on the sentence) and multiplied by 100 to restrict the range in emotion strengths to (0, 100).
      d) CrowdFlower tweet dataset
         (1) Maps 30368 tweets to 13 emotions: love, relief, neutral, anger, sadness, empty, surprise, fun, enthusiasm, happiness, hate, worry, and boredom.
         (2) Emotions are mapped to their closest emotion in dataset 1 (love/relief/fun/enthusiasm/happiness:joy, hate:anger, worry:fear, boredom:disgust).

2. Sentences had punctuation and stopwords removed, were converted to lowercase, stemmed (Porter stemming), and then converted to TF-IDF vectors.
3. Predictions - NRMSE: Normalized root-mean-square error
    a) Random Forest Regression predictions
        (1) NRMSE of ~15% for testing on dataset 3
        (2) NRMSE of ~16% for testing on datasets 1-3
    b) Logistic Regression predictions
        (1) NRMSE of ~18% for testing on 2490 dataset 3
        (2) NRMSE of ~19% for testing on datasets 1-3
4. A linear machine learning classifier (stochastic gradient descent, or SGD) was used on training data of text with emotional ratings.
    a) Text had stopwords and punctuation removed, were stemmed, and converted to tf-idf feature vectors.
    b) The SGD classifier was used because it works with training datasets that are normally distributed, and is flexible to the amount of data displayed.
5. Emotional scores were assigned to every scene of a script.
    a) These values allow us to track the film's content in each of the six basic psychological emotions.
    b) This allows richer analysis than a generic positive/negative textual evaluation (sentiment analysis).

VI. Conclusions and Discussion

A. Our IMSDB script database consisted of over 1100 movie scripts with varying box office grosses, product budgets, critical receptions (from both audiences and critics), and release dates. We decided to tentatively divide our movie set into different categories, based on box office, profitability, and critical reception.
    1. Evaluation of results was done both manually and by averaging the graphs of similar movies.
        a) Among films that were well-received critically, emotional intensities tend to be fairly low, with large peaks.
            (1) There are few scenes with large peaks, implying a subtler experience, which is an idea that pairs well with the fact that critically hailed movies generally are toned-down compared to high-thrill blockbusters, but have moments of great emotional impact.
            (2) Scenes also generally possess a greater variety of emotions, which is what we perceived as indicating a more complex

dialogue.

b) Among high-grossing blockbusters, we saw generally high emotional impacts, particularly in the final act.
    (1) These climaxes are often punctuated with single-emotion scenes.
    (2) These films are wild rides that aim to make the viewer feel they got their money's worth, so it is no surprise that their structure points to big finishes.
    (3) In addition, surprise is more likely to be found in larger amounts in these films

B. Because when averaging the movies, we saw relatively constant emotional makeups throughout the movie, these averaged visualizations are best used for measuring emotional impacts over runtime. Our conclusions mainly focused on the differences between a 'good' movie and a 'bad' movie, both in regards to box office gross versus budget, and critical appeal.

1. A common feature of nearly all these 'good' movies is a basic story structure, with beats that characterize a section.
    a) More specifically, the tone of the movie tends to change throughout, giving the viewer differently themed sections of the movie (areas of movies dominated by specific colors).
    b) Each one of these sections tends to have a unique combination of a primary emotion and emotion variance.
    c) Most importantly, they have a climactic, conflict-filled final portion: something that many 'poor' film" lack.
2. Many 'poor' films that we analyzed seemed to lack structure, having one emotion dominate throughout the film or keeping the same emotional tone throughout - not conducive to a basic story structure. In addition, many critically panned films have higher-than-average emotional impacts in the middle of their scripts, representing what critics recognized as overacting and improper variance of tension.

C. Next steps

1. Optimizing the machine learning models used
    a) Optimizing model parameter tuning further
    b) Utilizing other models
        (1) K-nearest neighbor
        (2) Extra tree
        (3) SVM
        (4) Multilayer perceptron
        (5) Various AdaBoost

2. Simplifying the visualization
    a) Visualization using strongest emotion of the scene only. This would allow more simplistic analysis of movies and hopefully allow us to reach more definite conclusions.
    b) Visualization using just the positive/negative sentiment of the scene only.
3. Extraction of scenes from scripts to predict from fitted models.
    a) Label scenes manually to gauge performance
4. Consider modifying consolidation of datasets
    a) While dataset 1 measures strengths of emotions, the other datasets only track votes for each emotion, which is not the same as the strength of an emotion. Predicting the strongest emotion would likely work better for all datasets.
    b) Dataset 3 is most likely to be relevant to movie scripts (as prose is more similar to scripts than headlines, storybook lines, and tweets). Using the 18 emotion labels from this dataset alone without modification might work best.
    c) Clean twitter data to use as an additional training dataset.
        (1) This dataset is currently unused as it is riddled with typos and slang - therefore it is unlikely to be common in movie.
        (2) This will be helpful for predicting dialogue emotion.
5. Consider combining emotion analysis and sentiment (positive/negative) analysis for a better model. Whenever a positive sentiment is reported, joy and sadness would be more likely to be visualized. Whenever a negative sentiment is reported, anger, disgust, fear, and sadness would be more likely to be visualized. This kind of ‘joint-approach’ would help with more ambiguous scenes that don’t have a strong primary emotion.

VII. Team member effort

A. All team members contributed equally to the project.

Research Papers

Aditya, Somak, et al. “From Images to Sentences through Scene Description Graphs Using Common Sense Reasoning and Knowledge.” 10 Nov. 2015.

Boor, Myron. "Reliability of ratings of movies by professional movie critics." *Psychological Reports* 67.1 (1990): 243-257.

Eliashberg, Jehoshua, Sam K. Hui, and Z. John Zhang. "Assessing box office performance using movie scripts: A kernel-based approach." IEEE Transactions on Knowledge and Data Engineering 26.11 (2014): 2639-2648.

Eliashberg, Jehoshua, Sam K. Hui, and Z. John Zhang. "From story line to box office: A new approach for green-lighting movie scripts." *Management Science* 53.6 (2007): 881-893.

Ghiassi, M., David Lio, and Brian Moon. "Pre-production forecasting of movie revenues with a dynamic artificial neural network." *Expert Systems with Applications* 42.6 (2015): 3176-3193.

Giannakopoulos, Theodoros, Aggelos Pikrakis, and Sergios Theodoridis. "A dimensional approach to emotion recognition of speech from movies." *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009.

Hunter, Starling David, Susan Smith, and Saba Singh. "Predicting Box Office from the Screenplay: An Empirical Model." (2015).

Ignatov, Dmitry I, and Denis Kornilov. “RAPS: A Recommender Algorithm Based on Pattern Structures.” 20 July 2015.

Lash, Michael T., and Kang Zhao. "Early predictions of movie success: the who, what, and when of profitability." *Journal of Management Information Systems* 33.3 (2016): 874-903.

Sharda, Ramesh, and Dursun Delen. "Predicting box-office success of motion pictures with neural networks." *Expert Systems with Applications* 30.2 (2006): 243-254.

Strapparava, Carlo, and Mihalcea, Rada. “Learning to Identify Emotions in Text”. 20 March 2008.

Wang, Hee Lin, and Loong-Fah Cheong. "Affective understanding in film." *IEEE Transactions on circuits and systems for video technology* 16.6 (2006): 689-704.

Zhu, Yukun, et al. “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books.” 2015.