Visualizing Storytelling using Sentiment Analysis

JohnHenry Ward

Abstract

Storytelling is a powerful tool that all of humanity shares, but can these stories be translated into a graphical form. Furthermore, can these graphs be categorized as one of the six basic shapes of storytelling? Using sentiment analysis with a siding window over a movie script, we will calculate the average happiness of that window, and eventually display a graph that can be used to tell the emotional story arc of that movie. Beyond that, can this graph be described as one of the six emotional arcs, or some combination of them? For each of the movies we focus on, we will attempt to categorize it, as well as discuss why, or why not, it fits into one of the six categorizes.

1 Introduction

The goal of this project is to see if it is possible to graphically visualize the plot of a movie. In his rejected master thesis, American author Kurt Vonnegut came up with a method to plot the arcs of stories. Multiple papers have taken his work, and have come to a conclusion that there are six basic plot arcs to describe stories. These plot arcs track the stories relative fortune, from good to ill, across the entire story. The six arcs are: Rise, Fall, Rise, Fall Rise, Rise Fall, Rise Fall Rise, and Fall Rise Fall. These arcs are given names based on common stories that follow the arc. Respectively, they are; Rags to Riches, Riches to Rags, Man in Hole, Icarus, Boy Meets Girl (or Cinderella), and Oedipus. Using sentiment analyzing, is it not only possible to see a graph, but define it as one of the six plot arcs. The six plot arcs can be seen in Figure 1. This project will perform sentiment analysis on any given movie script, and produce a graph that can then be studied and compared to these plot arcs, as well as graphs produce by other movie scripts. The graphs will be plotted on a graph with the x-axis representing time, and the y-axis representing the happiness score.

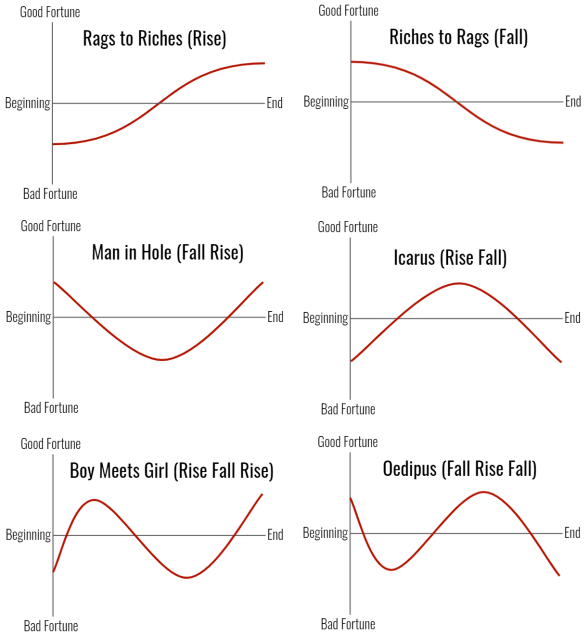

Figure 1: The Six Basic Plot Arcs of Storytelling

2 Related Work

This first related work, *The emotional arcs of stories are dominated by six basic shapes* from the University of Vermont, focuses on categorizing books into one of the 6 basic story arcs. In this research paper, they analyzed over 1,000 books from Project Gutenberg's collection, performed sentiment analysis on these texts and plotted the graphs to reveals 6 basic shapes. They were similarly inspired by Kurt Vonnegut's rejected master thesis which discusses the emotional arcs of storytelling. The University of Vermont's research paper was a big inspiration to try and identify key plot points in the output graphs, as they similarly did. This was found to be a bit more difficult than anticipated, as their approach to creating the graphs may have been a more refined and worked out then this projects. Another related work is *Visualizing the Emotional Arcs of Movie Scripts Us-*

ing Rule-Based Sentiment Analysis. This is more of a passion project, created by Nayomi Chibana, that uses a different metric to analyze the sentiment of movie scripts. The metric used here is The NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon. Chibana pulled scripts from The Internet Movie Script Database, and used this lexicon to analyze the intensity of the action, and how much of it there is. The result is an interesting display of graphs that can be compared to graphs of similar movies. This work has inspired some future work ideas that will be discussed later, as well as helped the approach in this project.

3 Data Used

3.1 Happiness Dictionary

The source for analyzing words and getting corresponding scores is from labMT: Language Assessment by Mechanical Turk. This data set contains over 10,000 words. Each word has a corresponding rank, score, and standard deviation, among other metrics that aren't pertinent to the project. The score that corresponds to a specific word is described as the happiness score. For the purposes of sentiment analysis, only the word and the corresponding happiness score will be used. The scores range from 8.5 for the word "happiness" to a 1.30 for the words "suicide" and "terrorist". This data set, and others that were considered, seem to be more suited towards social media but this labMT data did work well with movie script sentiment analysis. The data set is simply stored in a text file, that is read into a python dictionary for easy access, using a word as key, and the corresponding happiness score as the value. An example of the data set can be seen in Figure 2.

word	rank	score
laughter	1	8.50
happiness	2	8.44
love	3	8.42
happy	4	8.30
laughed	5	8.26
...	...	...
terror	10206	1.76
die	10207	1.74
killing	10208	1.70
arrested	10209	1.64
deaths	10210	1.64

Figure 2: Snippit from labMT data set

3.2 Scripts

Along with the data set, multiple movie scripts will be required. This project uses a variety of movie scripts which can be found in The Internet Movie Script Database. These scripts vary from being a very late stage draft of the script, which includes setting description and detailed actions, or just be the dialog of characters in the story, with little to no descriptive wording. There are six scripts that I will use in this projects results section. Those scripts are from The Wizard of Oz (1939), Psycho (1960), Jurassic Park (1993), Finding Nemo (2003), Iron Man (2008), and Avengers: Infinity War (2018). Of these six, Jurassic Park, Psycho, and The Wizard of Oz are the true scripts, while Iron Man and Avengers: Infinity War is the dialog with less descriptive actions, and Finding Nemo is only the dialog.

4 System Description

The system can be broken up into three parts. First, the data set that contains words and their corresponding happiness score needs to be read from a text file and stored in a data structure that could be easy and quick to access. Since python is being used, a dictionary is the optimal choice to store the word as well as the happiness score for that word. The text file containing the happiness scores is read in word by word, assigning the word as the key, and the happiness score as the value. Once that is complete, the second task can begin which involves processing the script. A script is stored in a text file, read into the program and tokenized. Non-alphanumeric symbols are removed, and each word is appended to a list. Once the whole script has been read, a list will contain the entire script, with one word per index. Next is calculating and plotting the happiness scores. Rather than plotting each words happiness score, which was an initial approach discussed in the What Didn't Work section, a sliding window was used. The window size is specified as being 10% of the entire text. The window starts with it's first word being the first word in the script. Every word in the window is read and the corresponding happiness score is calculated. Then the average is calculated, and that value is plotted. Once it is plotted, the window shifts over, and the new window's average happiness score is calculated. This continues until we reach the end of the script, meaning that the once the last word in the script

is read once, we finish. This means that the graph will be missing the first and last 5%, but this is a small enough amount that it does not sway the results.

5 Results and Discussion

5.1 Initial Test

Before jumping in and trying to categorize a movie script as one of the six shapes, the program needed to be tested to see if overall sentiment is in fact being calculated in an accurate way. To test this, the script from Avengers: Infinity War was used. This script specifically was chosen because of the nature of the film. It is not a typical super hero film in that the heroes come out on top at the end and declare victory, but rather the exact opposite. By the end of the film the antagonist, Thanos, erases half of all life from existence as the protagonists, The Avengers, fail to stop him. This should show a clear decline towards the end of the graph, with little to no incline after the fact. As can be seen in Figure 3, there is an obvious and steep decline towards the end of the graph. This preliminary test of the sentiment analysis system shows that, to at least a certain degree, the arc of a story can be visualized. Based on the graph, Avengers: Infinity War would be categorized as the Riches to Rags arc, also known as Fall.

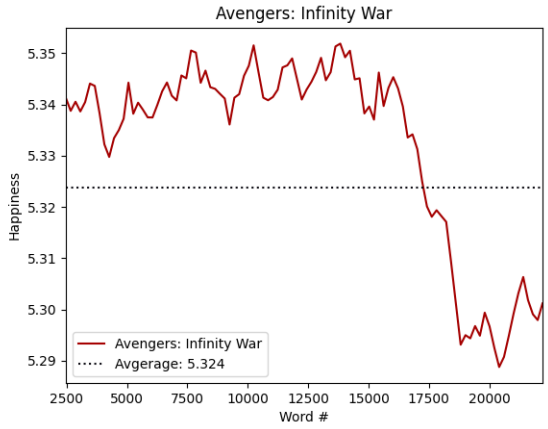

Figure 3: Avengers: Infinity War as Riches to Rags

5.2 Further Findings

Now that it's been seen that the approach seems to have potential, it should be seen how well it does with multiple movies that follow different plot arcs. First, one can see in Figure 4 that the animated Pixar movie, Finding Nemo, would be categorized as a Man in Hole (or Fall Rise) story arc. This makes sense to those familiar with the movie. Marlin's son Nemo, is captured and taken away, as can be seen by the steep decline in the beginning of the graph. This prompts Marlin to set off on a journey to find him. This is the bulk of the film, and can be seen as part of the graph that is below the average line. It is worth mentioning that sub-plots also seem to be visible to some extent, as seen in the middle section of the graph. Eventually. Marlin does find Nemo, and they return home, visualized as the incline towards the end of the graph. This is one of the better visualizations that was produced from this project, as it is distinctly obvious which of the six plot arcs it resembles the most, that being Man in Hole.

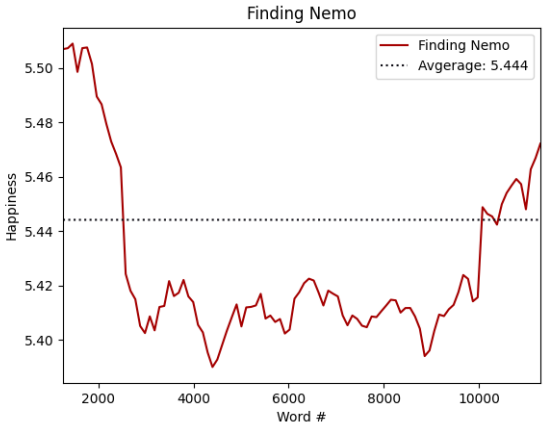

Figure 4: Finding Nemo as Man in Hole

Two more examples of movies that have fairly reasonable graphs are those of Psycho and Jurassic Park. While each of these graphs, seen in Figures 5 and 6, show clear peaks and valleys, it may not be instantly obvious as to which of the six arcs they follow. Psycho, whose main character Norman Bates, would be described as a having an Oedipus complex [1]. With this knowledge, one would assume that the story arc should follow that of the Oedipus plot arc, a Fall Rise Fall. As can be seen in Figure 4, the graph begins with a steady fall, before a rise that passes the scripts average score, before falling again, and eventually ending at the lowest score.

Perhaps a bit less obvious, and one that is harder to define, Jurassic Park, whose graph is seen in

[1] An Oedipus Complex is defined as: the complex of emotions aroused in a young child, typically around the age of four, by an unconscious sexual desire for the parent of the opposite sex and wish to exclude the parent of the same sex.

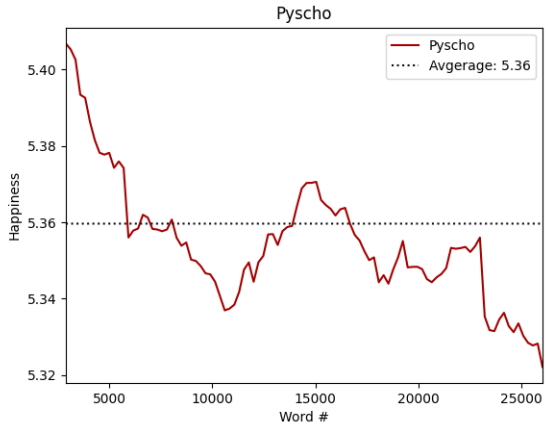

Figure 5: Psycho as Oedipus

Figure 6, could be described as a movie that has humans pushing the boundaries of science, before eventually having it all blow up in their face. This is similar to the story of Icarus [2], so one would assume that the plot arc would have similarities to that plot arc graph. With an initial look at Figure 6, this movie, and countless others like it, show that a movie has much more nuance than just a single plot arc. The Icarus plot arc could be categorized as the first half of the movie, while the second half may be of a totally different arc. Movies can be complex, intricate, and follow a variation of one of the six arcs, or just not follow one of the six arcs all-together.

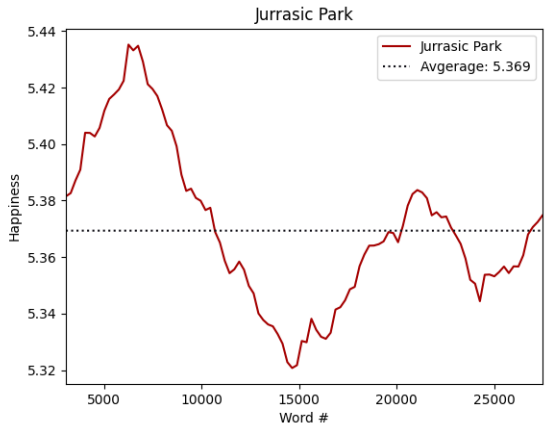

Figure 6: Jurassic Park as Icarus

Expanding on this idea that movies are complex, and as has been seen here, somewhat difficult to visually graph, one can see similarities of graphs from movies that have nothing in common. Take for example The Wizard of Oz and Iron Man, the graphs for these can be seen in Figures 7 and 8. Notice how both these graphs seem to follow the same pattern, a fall to start, with a rise that gets the movie past the half way point, then another fall, with an ending of a rise. The intensities may be different, but the arcs are similar in shape. These could be categorized as double Man in Hole arcs, a Fall Rise followed by another Fall Rise. This is interesting to note because it shows how this project could be used to find similar movies arcs, even if the movies may not seem similar on the surface.

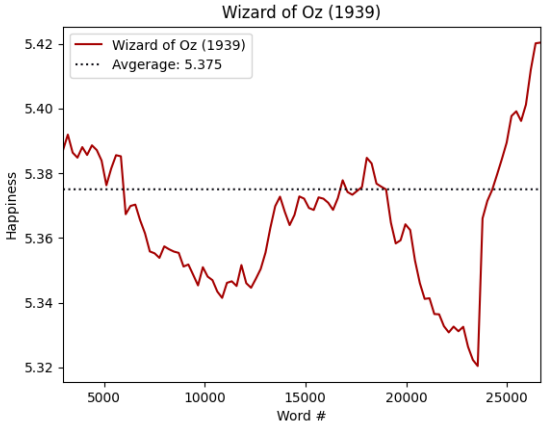

Figure 7: The Wizard of Oz as Double Man in Hole

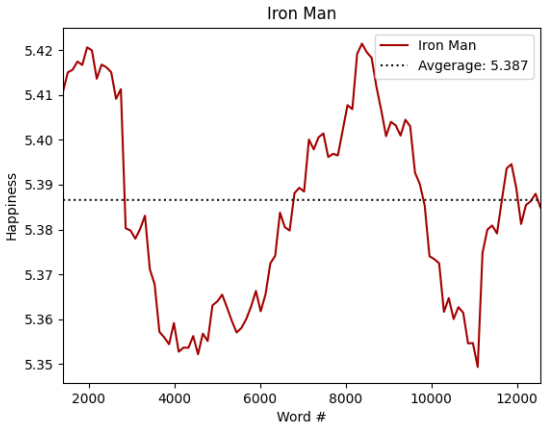

Figure 8: Iron Man as Double Man in Hole

5.3 Discussion

As is evident by the graphs shown here, the plot arcs can be visualized to some extent. While a more nuanced movie such as Jurassic Park or Psycho is a bit more difficult to categorize, movies that follow basic plot arcs are obvious, such as

[2]The Greek Story of Icarus involves Icarus flying higher and higher using wax wings, before flying too high where his wings melt and he crashes towards Earth.

Finding Nemo. Furthermore, some movies combine plot arcs, as is seen in Iron Man and The Wizard of Oz. While some of the graphs shown don't fit a plot arc exactly, the results are promising. Perhaps with a larger, more diverse lexicon that is better suited for movie scripts, the results could become increasingly obvious as to what plot arc they reflect. It also became apparent as the project continued that larger texts seemed to work better overall. The related work, *The emotional arcs of stories are dominated by six basic shapes* which uses books instead of movie scripts, found better success in categorizing the plot arcs. While other factors influenced this result such as more thorough research and approach by those involved, it should be noted that books are much more descriptive than movie scripts, and therefore should be easier to graph and categorize.

6 What Didn't Work

The initial approach to plotting the scores was to plot each words score as a single data point. This resulted in some erratic graphs that were not useful and did not show anything remotely close to what was expected. The sliding window approach was much better suited for what I wanted to accomplish, but the window size choice of 10% was not the initial choice. At first, a fixed window size of 500 words was used. This quickly broke down as it would not scale well with larger bodies of text, so a percentage of the text was chosen to be the window size. 10% of the text was chosen because it seemed to capture large enough plot arcs while still keeping the graphs interesting and detailed. The academic paper *The emotional arcs of stories are dominated by six basic shapes*, which was discussed earlier, also used a sliding window, which helped construct the approach for this project.

The choice of the labMT data set was not the first choice, and is still not a perfect data set. Initially, the NLTK Vader Sentiment Analyzer was used. This lexicon is more suited toward social media sentiment analysis, as it contains phrases, such as "the bomb", emoticons such as ":)", and acronyms like "lol". Since movie scripts are the focus of analysis, this lexicon did not seem like a very suitable set to use. Luckily the labMT data set is more suited towards generic text, and while is useful for social media, worked well with this project.

7 Conclusion and Future Work

As shown in this project, analyzing sentiment in movie scripts to visualize plot arcs is not perfect. It suffers from movie scripts not being as long or descriptive as most books, but is able in some cases to show a clear plot arc from beginning to end. The sentiment analysis used here is also fairly simple, and could benefit from a more rigorous analysis. Even so, the sentiment analysis used here had some successful outcomes, and shows clear and obvious plot arcs, similar to one of the six arcs discussed by Kurt Vonnegut. Storytelling is a powerful tool, and using sentiment analysis to visualize it is just one way to understand it more.

This project has multiple potential future developments. The first being an added automated similarity system, that can compare the output graph to the six plot arcs, and give a percentage score on each one based on the similarity of the graphs. This tool could be used to identify movies that, while on the surface appear very different, follow similar plot arcs based on the sentiment analysis scores. This leads to a second possible future work. K-means clustering could be used to plot movies, and view how similar multiple movies are to each other. It would also show the frequency of certain movie plot arcs. For example, a Man in Hole plot is could be more popular than an Icarus plot arc, and that could be visualized via K-means clustering. The third and most simplistic future work is comparing retellings of classic stories to see how closely they resemble the original text. For example, The Lion King is a retelling of Shakespeare's Hamlet. These two graphs should, in theory, be very similar as they are the same story, just told in different ways. There could be an issue of The Lion King being a movie, and Hamlet being a play written in an older English style. A similar issue and comparison could be used again for Shakespeare's The Taming of the Shrew, and the movie 10 things I hate about you, which is again a modern retelling of Shakespeare's story. This, as well as the other two possible future works, can show just how similar many movies are to each other, as well as the types of stories that are most popular, and how common some arcs are compared to others.

8 References

Chibana, Nayomi. (18 Oct. 2019) *Visualizing the Emotional Arcs of Movie Scripts Using Rule-*

Based Sentiment Analysis, Medium, Towards Data Science towardsdatascience.com/visualizing-the-emotional-arcs-of-movie-scripts-using-rule-based-sentiment-analysis-1016b4b1af5a

Kurt Vonnegut on the Shapes of Stories, Youtube, https://www.youtube.com/watch?v=oP3c1h8v2ZQ&ab

Reagan et al, University of Vermont. (2016) *The emotional arcs of stories are dominated by six basic shapes* [PDF] https://cdanfort.w3.uvm.edu/research/2016-reagan-epj.pdf

Rinker, Tyler. (2 May. 2019) *LabMT: Language Assessment by Mechanical Turk*, rdrr.io/cran/qdapDictionaries/man/labMT.html

"The Internet Movie Script Database (IMSDb)." The Internet Movie Script Database, imsdb.com