From network to sentiments and emotions: Analysing the Avengers movie transcripts

Sin Sze Li

Abstract— The transcripts of the four Avengers movies were analysized to study the relationship between characters, difference in the emotional arcs of the movies, and characters' emotions. Network analysis and sentiment analysis were utilized for the tasks. Networks were built based on the locations of the characters in the transcripts, and key characters were identified. In sentiment analysis, VADER was used to evaluate the polarity of the lines throughout the movies. However, the emotional arcs of the four movies show no significant difference. NRC lexicon was also employed to dissect the characters' emotions. The main Avengers characters' emotional development from the first to last Avengers movies was shown, and the emotional profile was used to distinguish villains from heroes.

1 Problem Statement

The Marvel Cinematic Universe is a media franchise that has won billions of fans around the globe. From 2008 to 2021, Marvel has released 27 movies in total. Most of Marvel's movies feature a particular superhero, while in the 'Avengers' movies, those superheroes come across each other and fight side by side. There are four Avengers movies: The Avengers, Avengers: Age of Ultron, Avengers: Infinity War and Avengers: Endgame. The latter two received remarkable ratings of 8.4 out of 10 on IMDb. Avengers: Endgame, the final movie in the series, has a box office of 2.798 billion USD, making it the highest grossing movie of all time before March 2021 (when Avatar was re-released in China after 12 years and retake the box office crown).

In this study, text analysis is used to answer the following research questions:

- Can social network analysis be used to correctly identify key characters?
- Using sentiment analysis, is there a significant difference in the emotional arcs among the movies?
- By sentiment analysis, can the characters' emotional development throughout the four movies be displayed and explained?
- Can heroes and villains be distinguished by their emotion profiles?

The analysis of this study is based on the Avengers movie transcripts. From these movie transcripts, lines for each character can be found, which are useful for sentiment analysis. The sequence of the lines also has its temporal importance, which is crucial in the first two analytical questions.

2 State of the Art

The extraction of social network in movies was studied by Park et al. [1]. Social networks in movies were introduced in a sense that 'Characters can develop relations through dialogues and actions between themselves. These relations form a social network constructed by all the characters in a movie'. They named this type of social network as 'Character-net'. In their paper, they tried to classify characters in 10 different movies, including Avatar, Die hard, Spider-Man2, as protagonist, tritagonist and antagonist based on emotion. Their methodology includes social network modelling based on dialogues, extraction of emotional words and clustering of characters based on emotion. In the aid of building a character-net, they first found out the character names and lines for each dialogue in a scene, then decided the weight of the line in a network based on the number of characters in each dialogue in the scene. The important characters were then determined by the degree of centrality of each node. The concept and method of building character-net can be employed in this study because the transcript of the Avengers movie provides similar information. However, the number of characters in each scene in the Avengers movies could not be determined because the dataset does not contain information about the scenes.

Lee et al. [2] collected over 900 English movies to extract and compare their positive sentiment scores. They gathered the movie scripts, segmented the movie scenes into blocks containing 20 lines, and determined the sentiment value using SentiWordNet word set. When the sentiment scores were plotted against time, the graphs were spiky and difficult to interpret. Therefore, they applied a moving method to smooth the graphs. The moving method regards several blocks as a window and return the average sentiment value in that window. They also tested a few signal smoothing algorithms such as spline interpolation, locally weighted scatterplot smoothing, Savitzky-Golay filter and Gaussian smoothing filtering. The moving method and smoothing algorithms had helped provoke some ideas on how to handle spiky and all pointed graphs generated from the Avengers movies transcripts.

The use of TextBlob for sentiment analysis in movie scripts was discussed by Rahmen et al. [3]. In their paper, they chose transcripts from several adventure movies, performed tokenization, removed punctuation marks and eliminated stop words as the data cleaning process. In their exploratory data analysis, they used word clouds to find the most said words in each movie. TextBlob was then applied to compare the movies' subjectivity and polarity. The polarities of the movies were plotted against time to show the

fluctuations in emotion. Their step-to-step detailed description on the sentiment analysis process serves as a guide on how to implement such analysis on text data. Word clouds and plots of emotion polarity derived from Avengers movie transcripts could be used to answer the research questions.

3 Properties of the Data

The data used in this study can be found on Kaggle [4]. The author of the dataset obtained transcripts of all Marvel movies from the website 'Fandom's Transcripts Wiki'. The transcripts were copied and pasted into txt files, then processed and stored in csv files.

Transcripts and scripts hold different information about the movies. Transcripts only contains the lines that actors say in dialogues, while scripts also include setting of scenes, actors' movements and more. Since the script of Avengers: Endgame cannot be found publicly on the Internet, only the transcripts were used for this study. This, however, has its limitation which is discussed in the Critical Reflection part of this report.

The dataset on Kaggle has many csv files, two of them were used in the analysis. First, the movie overviews file. It was used only to get a general idea of the data in the movie transcripts. It contains statistics that summarises different characters involvement in Marvel's movies. There are 24 columns including character names, their number of lines in each movie, their number of appearances in all movies, their average words per line and their average lines per movie. This is a clean dataset that it has no missing values. However, the maximum average words per line is 235, which is peculiar. After investigation, this data comes from the character 'TV Reporter's Voice' in Iron Man. In that scene, Iron Man was watching TV, and the reporter's voice is captured as a line, which explains this exceptionally long line.

The number of distinct characters in each movie were counted and visualised as in Figure 1. Avengers: Endgame has the greatest number of characters, not only because it is an ensemble movie, but also the finale of the Avengers series, so most of the characters that appeared in the previous Marvel movies were in it.

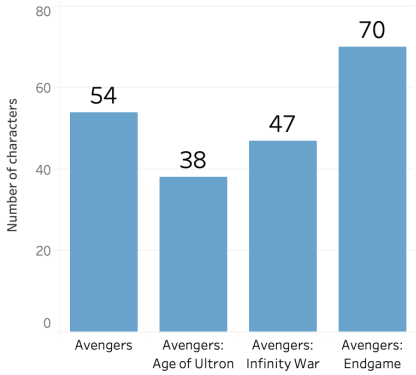

Fig. 1. Number of characters in each Avengers movie.

Also, the top characters with most lines were found in each Avengers movie, as shown in figure 2. Tony Stark, Steve Rogers, Bruce Banner, Thor, Natasha Romanoff and Clint Barton were the original six characters in Avengers. Hence, all of them are top characters with most lines in at least three of the four Avengers movies. Their emotions were traced along the four movies to answer analytical question 3. Loki, Ultron and Thanos are the villains in the movies, their emotions were compared with the Avengers to answer analytical question 4.

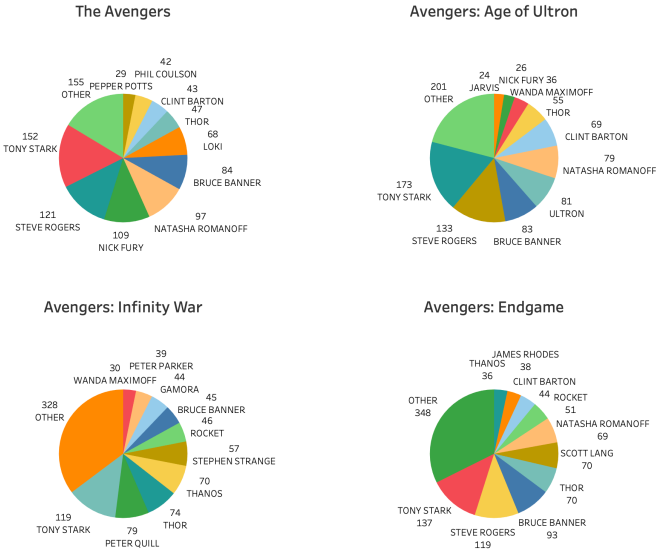

Fig. 2. Top characters with most line.

The second dataset used was the transcript csv file, which contains two columns: character and line. It is the main dataset for analysis. Each row represents a character and his/her respective line. Since the rows are arranged chronologically, sentiment analysis could be utilized line by line to find out the emotion fluctuation in the movies for analytical question 2. Network analysis could also be used to investigate the relationship between these characters for analytical question 1. Note that this dataset includes lines from all Marvel movies, so only those from the Avengers movies were selected for the analysis.

4 Analysis

4.1 Approach

The analysis was divided into three parts: data cleaning, analysis, and visualization (Figure 3).

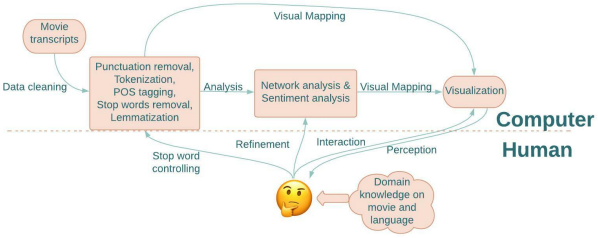

Fig. 3. Workflow of the analysis.

After data cleaning, word clouds were constructed using the python package [5] to give a general idea of the clean words present in the dataset. From the word clouds, words that gave insignificant information to the analysis were manually deleted. This process was iterated several times before the lines are clean and ready for further analysis.

Network analysis was applied in attempt to find out the relationships between characters according to their appearance in the movies. In books, the occurrence of characters can be determined when the characters' names appear in the sentences. However, characters do not call their names every time a character appear in a movie scene, so their appearance cannot be extracted simply from the dialogues. To tackle this problem, the name and line columns were concatenated into a single string, so that the character's name appears before each line. Then, all the lines were combined, and the location of each character's name was found. From there, interactions were defined as occurrences of characters within a certain threshold distance. Networks were subsequently built based on the interactions using python package NetworkX [6]. The threshold was then fine-tuned manually based on the visualized network and domain knowledge on the movies. Importance of characters could be perceived from the network as characters with highest degree centrality.

In sentiment analysis, VADER and TextBlob were both tried out to find the emotion arcs of the movies. The transcripts were evaluated on a line-level using both models and the returned sentiment scores were visualized with respect to their temporal location. The compound score returned by VADER and the polarity returned by TextBlob were chosen as the indicators for emotion. It was found that the emotion fluctuated less and skewed more towards positive with TextBlob. From the viewer's perspective, the Avengers movies are full of excitements and ups-and-downs, so results from VADER were picked to compare the four Avengers movies.

To study characters' emotions in depth, the library NRC lexicon [7] was employed to analyze character's lines. Their emotional affects in the four Avengers movies were returned and compared. Moreover, emotional affects among characters were also compared to distinguish villains from heroes. Throughout the analysis, human's prior knowledge on the movies was integrated with the visualization generated by computer in verifying and explaining the results.

4.2 Process

The analysis started with data cleaning process, which included punctuation removal, tokenization, stop words removal, POS tagging, and lemmatization. In punctuation removal, punctuation marks in sentences were removed. Contractions such as I'm, you're, it's and isn't were separated as 'I m', 'you re', 'it s' and 'isn t'. Next, the sentences were tokenized, meaning they were separated into individual words. After that, stop words, the most common words in a language, were removed according to the NLTK stop words list. At this stage, words that were variants of the same word, e.g. 'lost' as the past tense of 'lose', 'thinking' as the gerund of 'think' were considered as separate tokens, but in fact they corresponded to the same words. Therefore, stem words, the original form of the words, had to be found. To achieve this, the part of speech of each token was determined. Each token was classified as noun, adjective, adverb, verb or other. Then each token was converted into a tuple with the form (word, POS tag). Lemmatization could then be applied such that word variants like 'lost' was converted to stem words like 'lose'.

The results after data cleaning were shown by word clouds that illustrates words with different sizes according to their frequencies in the cleaned words. However, from the word clouds, there were still many words with insignificant meaning. In conversational English, words such as 'right', 'okay', 'hey', 'like', 'yeah', 'look' came up very frequently. They carried little meaning but they took up part of the word clouds so they should not be included in the analysis. Originally, only the NLTK stop word list was used, which seemed to be insufficient. In light of this, lists of stop words from other libraries, for example, Gensim, Spacy and scikit-learn were also put in use. A list with extra words that were not included in any of the libraries was created. It contained words such as 'hey', 'yeah', 'right', 'yes', 'uh' and 'oh'. Words that were included in the libraries and the extra list were deleted, and the final word clouds showed the main themes in the movies (Figure 4). For example, Loki, Ultron and Thanos were the names of the villains in the movie, while the first Avengers movie focused on the Tesseract and the last two spotlighted on the Infinity Stones, so they were mentioned frequently in the movies and appeared larger in the word clouds.

Fig. 4. Word clouds of the movies after data cleaning.

Next, for network analysis, as mentioned in the last sub-session, the characters' names were added to the front of each line. The lines were combined, and the locations of the characters' names were found. Finally, characters' interactions were defined as names that appeared within a certain distance. A challenge in this step was the number of words in characters' names. Some of the character names had one word, e.g. Thanos, Hulk, Gamora, which could be located easily, but some of them had two, e.g. Peter Quill, Steve Rogers, Natasha Romanoff. Since the names were located

word by word, characters sharing the same first name would be mistakenly considered as the same character, e.g. Peter Quill and Peter Parker. It was also not appropriate to use the last names because for example, Tony Stark and his daughter shared the same last name. Therefore, the combined line was tokenized using NLTK's multi-word expression tokenizer that tokenizes the combined line into single words except pre-defined multi-words, which in this case were the multiple-word character names. The distinct characters were then correctly located and plotted according to the time they appeared. An example of Avengers: Infinity War is shown in Figure 5. The y-axis shows different characters, and the x-axis shows their corresponding appearance in terms of word number in the whole transcript.

Once the occurrences of characters were found, their interactions can be derived. If two characters appeared within 200 words in the transcript, they were considered as having an interaction. From this definition of interaction, characters' interactions were counted and recorded in a matrix, which could be visualised using a heatmap, as shown in Figure 6. The higher the number, the more the interactions occurred between characters.

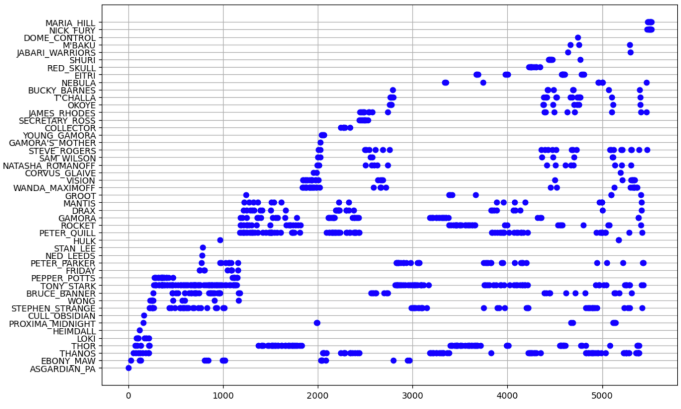

Fig. 5. Locations of characters' appearances in Avengers: Infinity War. The x-axis is the word number in the whole transcript.

Based on the matrix, networks were subsequently built, which were at first messy and unreadable because all most of the nodes representing characters were positioned at the centre of the network and overlapped. When the interaction was defined, an assumption was made that two characters' names appearing within a certain distance implied they had interaction, which was not always the case. For example, the last character speaking in the first scene and the first character speaking in the second scene could be considered as interacting by this definition, but they were in fact in two different scenes. Also, Stan Lee appeared in most of the Marvel's movies as cameo, but according to the network, he interacted with many other characters. To filter out these wrongly defined interactions, interactions that occurred less than 30 times were deleted. The resulting networks showed there were characters who had no interactions at all, so only the largest subnetworks were generated (Figure 7).

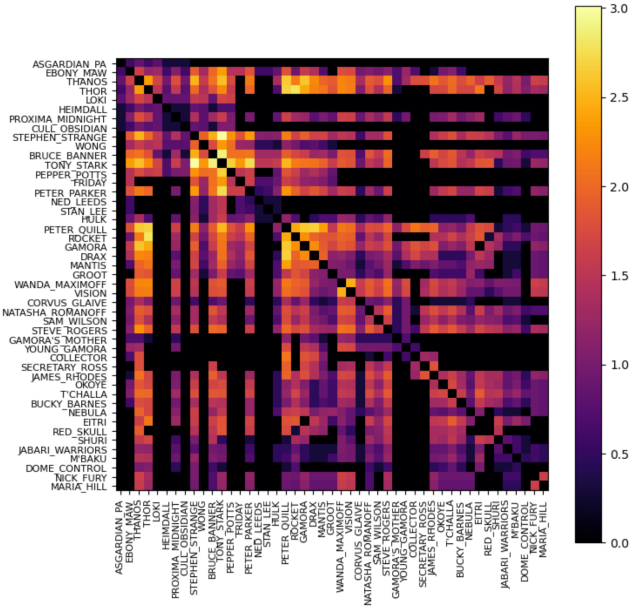

Fig. 6. Heatmap showing interactions between characters in Avengers: Infinity War. Note that log was taken because the interactions were mainly focused on some key characters only.

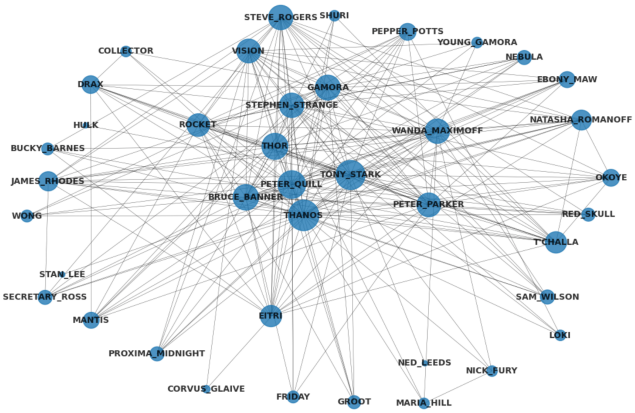

Fig. 7. Network of Avengers: Infinity War after deleting ill-defined interactions.

Characters in the center of the network generally had more interactions with other characters. Moreover, the size of the nodes represents their centrality degree. However, it is difficult to compare the importance of characters because the nodes are arranged randomly. If the nodes were arranged according to the chronological order of characters' appearances, maybe the plot of the movie could even be interpreted, as shown in figure 8.

The first character in Avengers: Infinity war, Asgardian PA, is located on the right. The nodes are arranged according to the time the character first appeared in the movie in anti-clockwise direction. This network can better show the importance of characters in a sense that it also includes when the characters were introduced in the movie.

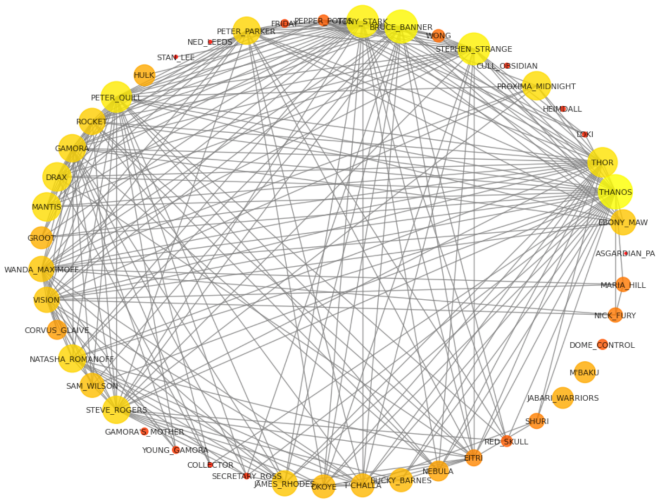

Fig. 8. Network on Avengers: Infinity War. The nodes are arranged according to the chronological order of characters' appearances.

In sentiment analysis, the lines were analyzed in three different ways: line by line, as a whole and by characters. To find out the emotion arcs of the movies, each line was evaluated as a compound score of positive, negative and neutral emotions. The compound scores were than plotted according to the line number. To smooth the graphs, the rolling average with window size of 5 was computed and a median filter of window size of 15 was applied. The window sizes were determined based on the effect they had on the visualization. The general fluctuation of the score should be easily seen, yet the insignificant variations could be omitted. The results are shown in figure 9.

There is no significant difference in the emotional arcs among the four movies. In order to conduct a detailed sentiment analysis, NRC lexicon was used to evaluate the lines. In NRC lexicon, words are assigned with their associated emotions. Eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, disgust, and two sentiments: positive and negative are covered in NRC lexicon. Using NRC lexicon, the lines were analyzed as a whole for each movie, and the frequency of words that are associated with these emotions were computed and visualized (Figure 10). Avengers: Endgame has the highest frequencies in negative and anger. This is because of the setting of the movie itself that Thanos had wiped out half of the living things in the universe.

Sentiment analysis on characters was also carried out. The same technique with NRC lexicon was applied to lines of different characters so that the emotional development of the original Avengers characters throughout the four Avengers movies could be studied, as shown in figure 11. Their change in emotion can be explained. For example, Natasha Romanoff's negative frequency decreases in Avengers: Age of Ultron because she had built genuine and close relationships was other superheroes. In Avengers: Infinity War, she lost many of her loved one because of Thanos and the negative frequency increases drastically.

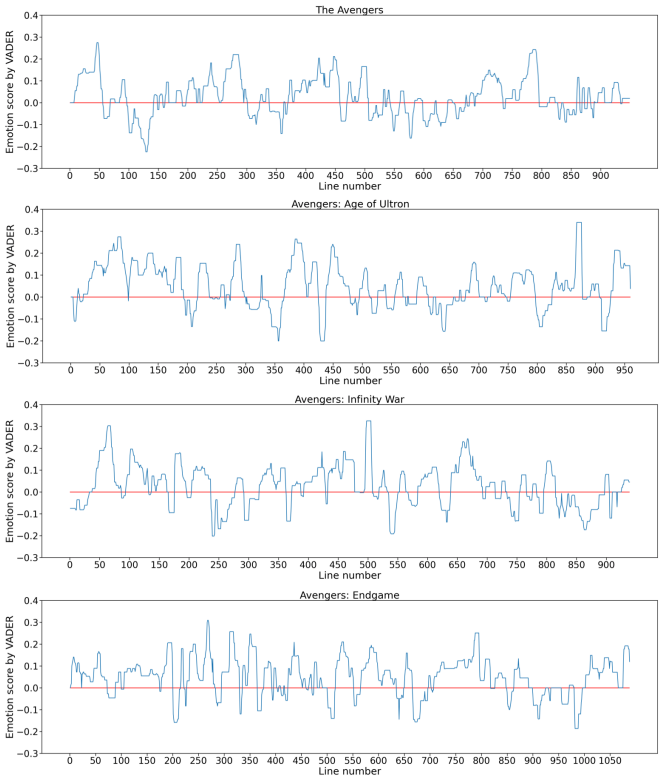

Fig. 9. Emotional arcs of the four Avengers movies.

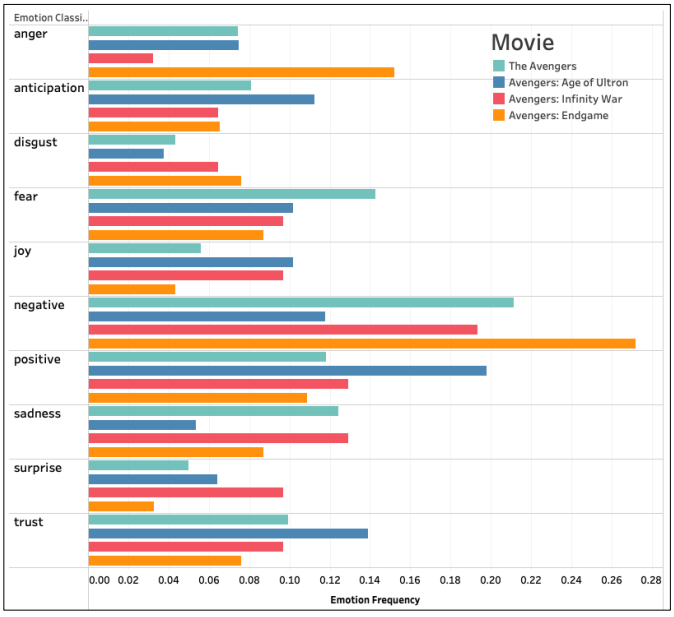

Fig. 10. Sentiment analysis by NRC lexicon on the four Avengers movies.

Finally, the emotion profiles for heroes and villains in different movies were created, so that their character traits could be compared (Figure 12). The bars for villains were represented in red to emphasize the difference between heroes and villains.

4.3 Results

Based on the analysis and visualisation, the research questions are answered in this session.

In figure 8, the key characters in Avengers: Infinity War were successfully identified by the larger nodes such as Thanos, Stephen Strange and Tony Stark. Besides, it can also be shown in figure 9 that the key characters in the movie appeared in different stages but not all at the beginning.

The emotional arcs of the four Avengers movies do not indicate significant difference. However, Avengers: Endgame has a noticeably higher frequencies on anger and negative than the other three Avengers movies in figure 10.

The emotional development of the characters can be derived by sentiment analysis and illustrated by visualization. The change in their emotions over the four movies matches the storyline.

According to figure 12, the three villains – Loki, Ultron and Thanos have one thing in common. They are generally more positive than most of the heroes, except in Avengers: Endgame which is the second part to Infinity War. In movies, villains are usually very confident and believe that they can win in a fight. Other than this, there is no major difference in the emotion profile between villains and heroes.

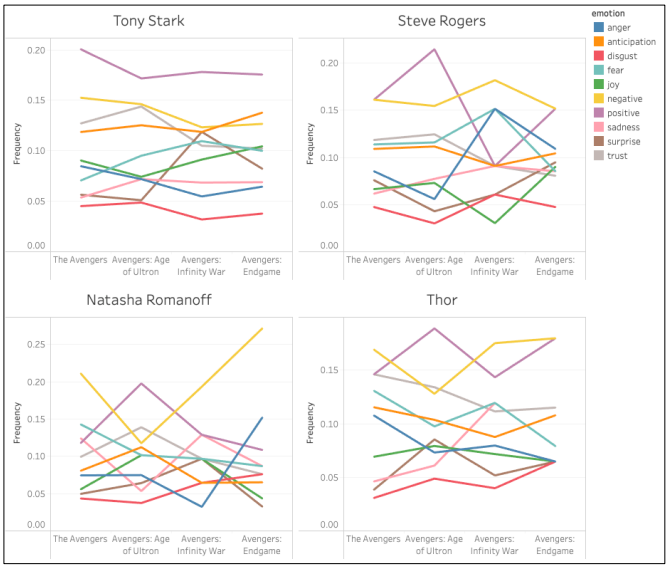

Fig. 11. Emotion development for some characters throughout the Avengers movies.

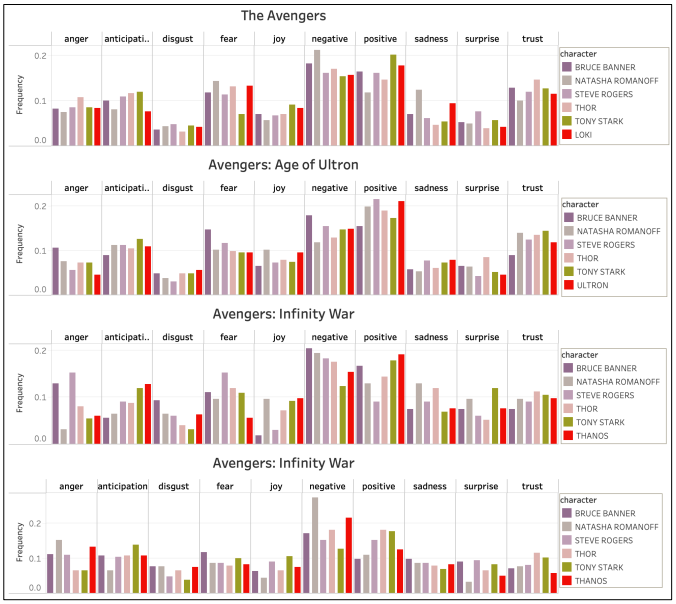

Fig. 12. Emotion profiles for heroes and villains in the Avengers movies.

5 Critical Reflection

Human reasoning was involved constantly throughout the analysis. For example, in the data cleaning process, conversational words such as ‘yeah’, ‘oh’, ‘uh’, ‘hm’ are not included in the common stop words libraries. They had to be manually added to the stop words lists. Besides, the locations of characters could not be extracted successfully in the first run because some characters could only be differentiated by their full names, which contained multiple words. The process of determining characters’ locations had to be re-examined and a new tokenization method had to be put in place. The original social network created was impossible to interpret because of the nodes overlapping each other. The reason behind was figured out by reviewing the definition of interaction, and certain interactions were then omitted to obtain interpretable networks in figure 7 and figure 8.

The emotional arcs among the four Avengers movie did not indicate significant difference. Smoother curves may give some new insights. However, by increasing the window sizes for rolling average and median filter, the graphs became flat topped. Other smoothing algorithms such as Gaussian smoothing filtering or Savitzky-Golay filter can be applied for further investigation on this topic.

Analysing movie transcripts instead of movie scripts has its limitation. As mentioned before, movie scripts carry much more inform than movie transcripts. In fact, the sentiments and emotions conveyed by a scene in the movie does not solely come from the dialogues, but also the body languages, facial expressions, scene settings and more. Just by analysing the transcripts of movies do not give the full picture of the sentimental and emotional development of the plot. Moreover, analysing movie scripts can also separate the scripts into different scenes, so that when the social networks are constructed, there can be a definitive point to tell where

characters do not have any interaction regardless of their location in the script itself. For example, interactions can be defined as characters that appear in the same scene instead of the current definition in this study.

The emotion classification in figures 10, 11 and 12 made use of the NRC lexicon. It is observed that positive and negative sentiment take up most of the affect frequency. There are 8 emotions but only 2 sentiments in the lexicon. Therefore, words are frequently associated with positive and negative sentiment than detailed emotions. Other available lexicons such as LIWC, CBET and Word-Emotion Mapping lexicon can be applied to the same dataset for comparative analysis.

To further investigate the traits of characters, unigrams and N-grams features can be extracted. Unigrams and N-grams are the words and word-clauses that are made up of 1 or N words. Using these features, characters' choice of words can be found, which may be an interesting topic to work on.

Table of word counts

Problem statement	242
State of the art	468
Properties of the data	496
Analysis: Approach	499
Analysis: Process	1341
Analysis: Results	197
Critical reflection	459

REFERENCES

[1] S.-B. Park, E. You, and J. J. Jung, "A cinemetric approach to sentimental processing on story-oriented contents," *Qual. Quant.*, vol. 48, no. 1, pp. 49–62, Jan. 2014, doi: 10.1007/s11135-012-9748-6.

[2] S.-H. Lee, H.-Y. Yu, and Y.-G. Cheong, "Analyzing Movie Scripts as Unstructured Text," in *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, Apr. 2017, pp. 249–254. doi: 10.1109/BigDataService.2017.43.

[3] R. Rahman, M. Abdul Masud, R. Jahan Mimi, and Mst. Nusrat Sultana Dina, "Sentiment Analysis on Adventure Movie Scripts," in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dec. 2020, pp. 1–6. doi: 10.1109/STI50764.2020.9350525.

[4] "Marvel Cinematic Universe Dialogue | Kaggle." https://www.kaggle.com/pdunton/marvel-cinematic-universe-dialogue (accessed Jan. 06, 2022).

[5] A. Mueller, *wordcloud: A little word cloud generator*. Accessed: Jan. 09, 2022. [Online]. Available: https://github.com/amueller/word_cloud

[6] *NetworkX*. NetworkX, 2022. Accessed: Jan. 09, 2022. [Online]. Available: https://github.com/networkx/networkx

[7] metalcorebear, *NRCLex: An affect generator based on TextBlob and the NRC affect lexicon. Note that lexicon license is for research purposes only*. Accessed: Jan. 09, 2022. [Online]. Available: https://github.com/metalcorebear/NRCLex

[8] A. van Cranenburgh and D. Benelux, "An Empirical Evaluation of Sentiment Analysis on Movie Scripts," p. 4.

[9] "Analyzing Star Wars Movie Scripts." https://kaggle.com/xvivancos/analyzing-star-wars-movie-scripts (accessed Jan. 09, 2022).

[10] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–33, Mar. 2018, doi: 10.1145/3057270.

[11] J. Kim, Y. Ha, S. Kang, H. Lim, and M. Cha, "Detecting Multiclass Emotions from Labeled Movie Scripts," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jan. 2018, pp. 590–594. doi: 10.1109/BigComp.2018.00102.

[12] J. J. Jung, E. You, and S.-B. Park, "Emotion-based character clustering for managing story-based contents: a cinemetric analysis," *Multimed. Tools Appl.*, vol. 65, no. 1, pp. 29–45, Jul. 2013, doi: 10.1007/s11042-012-1133-x.

[13] "NLTK :: Natural Language Toolkit." https://www.nltk.org/ (accessed Jan. 09, 2022).

[14] metalcorebear, *NRCLex: An affect generator based on TextBlob and the NRC affect lexicon. Note that lexicon license is for research purposes only*. Accessed: Jan. 09, 2022. [Online]. Available: https://github.com/metalcorebear/NRCLex

[15] M. Jockers, "Revealing Sentiment and Plot Arcs with the Syuzhet Package | Matthew L. Jockers." https://www.matthewjockers.net/2015/02/02/syuzhet/ (accessed Jan. 09, 2022).

[16] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Movie Analysis from the Perspective of Social Networks," *IEEE Trans. Multimed.*, vol. 11, no. 2, pp. 256–271, Feb. 2009, doi: 10.1109/TMM.2008.2009684.