Avengers: Age of Quantitative Text Analysis

--Movies Reviews Sentiment and Topic Analysis with the Rating Classifier

Tianzhu Qin
t.qin4@lse.ac.uk

Abstract

The essay aims at analyzing the Avengers movie reviews in order to catch views from audiences rather than the movies themselves. Firstly, the essay scrapes reviews from IMDb (Internet Movie Database) for all the four movies, including numerical ratings and related texts. A series of models are then proposed to carry text mining. Sentiment analysis introduces that the sentiment does make sense in explaining the rating, with some regular pattern. Topic modeling shows that there are evident topics people like to talk about. And the topics sometimes highly relate to the plot, except the praise to movie. Lastly, with the text features acquired from sentiment analysis and topics modelling, the essay also introduces two rating classifier, Elastic-Net and Random Forest, to give a rating to the review itself. The classifier can practically help to have deeper understanding of the text features.

1. Introduction

Avengers: Endgame, as an incredible move by *Marvel*, is getting increasingly popular these days. Tickets for the next few weeks are snapped up within several minutes, and almost everywhere we could see people talking about it. It is astonishing to see a movie could be as successful as this, and we cannot help to think if there are some special methods in the movie to arouse public discussing or criticizing. Furthermore, whether the contents of people's talking, which may focus on several topics or have special sentiment directions, will influence the people's interests is also a point worth thinking. In this sense, if we can find out the characteristics of the movie and its potential chances to be imitated by other movies, the essay can really benefit the whole movie industry, by paying more attention to the *sentiment director* and *topic maker*, so as to have a high rating and extensive popularity.

The essay would like to analyze the *Endgame* success, as well as the whole *Avengers* series, from the perspective of audiences, who express their opinions in the website, with their ratings for movies. Different from the view from the movies themselves, the essay chooses the more objective targets, who will determine the rating and the popularity at last. The essay implemented a web-scraper to acquire all reviews data before May 1st, 2019 (It should be noticed that the project is constantly updated, and the initial work is done long before the *Endgame* came out, with the data still collecting, processing and analyzing). The essay then carries sentiment analysis of the words to find the emotion directions of the reviews and its pattern changes. The topic modelling takes out main topics of the four *Avengers* movies, which are compared later. Finally, a rating classifier is introduced to further analyze which influence the most of the ratings and popularities. The classifier itself is also an important method to predict the potential value of the movie, for the producers' and investors' references to response in time.

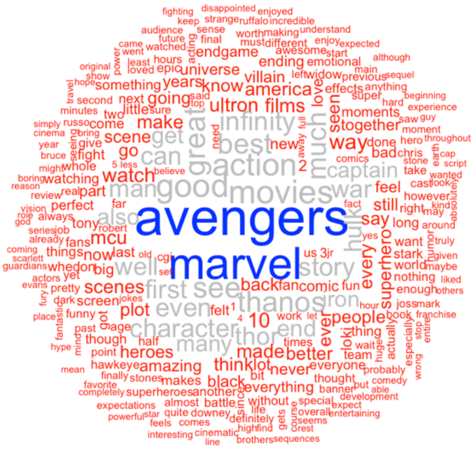
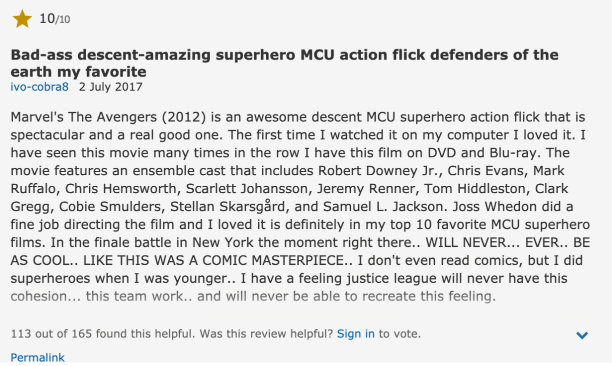


Figure 1 Word Cloud of All reviews for Avengers (Colored by Captain America's shield)

2. Data Scraping

In order to acquire data, the essay implements a data-scraper for the website of *IMDb (Internet Movie Database)*. The website holds the reviews of a movie by users who watched it. A user is able to write his comments towards the movie, and may give it a rate (not necessary).

The essay hereby takes *Avengers1* as an example. The reviews website for it is https://www.imdb.com/title/tt0848228/reviews?ref_=tt_urv. Any movie is determined by the id (i.e. *tt0848228*). In the website, all reviews are stored by the class *review-container*, in which *ratings, titles, authors, dates, texts, number of people who were helped by the reviews and the number of all people who gave it a remark (i.e. 113 out of 165 found this helpful),* as showed by the graph below.



★ 10/10

Bad-ass descent-amazing superhero MCU action flick defenders of the earth my favorite
ivo-cobra8 2 July 2017

Marvel's The Avengers (2012) is an awesome descent MCU superhero action flick that is spectacular and a real good one. The first time I watched it on my computer I loved it. I have seen this movie many times in the row I have this film on DVD and Blu-ray. The movie features an ensemble cast that includes Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johansson, Jeremy Renner, Tom Hiddleston, Clark Gregg, Cobie Smulders, Stellan Skarsgård, and Samuel L. Jackson. Joss Whedon did a fine job directing the film and I loved it is definitely in my top 10 favorite MCU superhero films. In the finale battle in New York the moment right there.. WILL NEVER... EVER.. BE AS COOL.. LIKE THIS WAS A COMIC MASTERPIECE.. I don't even read comics, but I did superheroes when I was younger.. I have a feeling justice league will never have this cohesion... this team work.. and will never be able to recreate this feeling.

113 out of 165 found this helpful. Was this review helpful? Sign in to vote.

Permalink

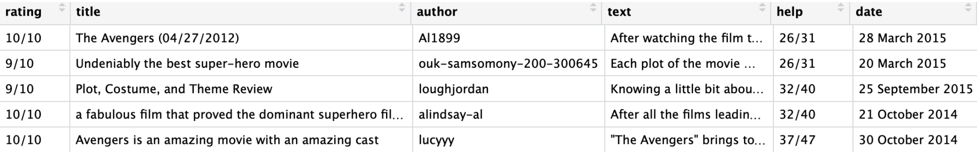
Figure 2 IMDb Review Example

What is worth saying, sometimes a container will hide some texts for presentation purpose, which result to locate the whole container to get all texts in a cell. Besides, one page of the website only returns 25 reviews, and for more we want we have to press the button below, to request more data. In this case, the scraper the essay sets will look for the inner object the button will target to, and directly request for that API. It turns out that the html code of this component holds the *key* and the *ajaxurl* in it, so that the whole process will be a request following a request, while all cells are processed and the next page API is acquired.

▾ Load More

Figure 3 Load Button

All codes of this scraper are showed in Appendix. The data we get is four *data.frames* for four movies, i.e. *Avengers* (Avengers 1), *Avengers: Age of Ultron* (Avengers 2), *Avengers: Infinity War* (Avengers 3), and *Avengers: Endgame* (Avengers 4). Example of the data is showed below.



rating	title	author	text	help	date
10/10	The Avengers (04/27/2012)	Al1899	After watching the film t...	26/31	28 March 2015
9/10	Undeniably the best super-hero movie	ouk-samsomony-200-300645	Each plot of the movie ...	26/31	20 March 2015
9/10	Plot, Costume, and Theme Review	loughjordan	Knowing a little bit abou...	32/40	25 September 2015
10/10	a fabulous film that proved the dominant superhero fil...	alindsay-al	After all the films leadin...	32/40	21 October 2014
10/10	Avengers is an amazing movie with an amazing cast	lucyyy	"The Avengers" brings to...	37/47	30 October 2014

Figure 4 Data Sample

3. Exploratory Analysis

To deeper understand the data logic, the essay carries the exploratory data analysis hereby to show the interesting pattern of the data and wishes to further find the likely breakthrough.

At the very first, the essay carries the preprocessing towards both numerical and text data. The *rating* is followed by the string "/10". The *help* including both only the number of people who are helped, and the number of all people give the judge, is divided by "/". The *date* is not in the formal type. As for texts, the essay carries "lowering", "stemming", "punctuation removing" and "stopwords removing". Most of the cases, the values are computing by adding proportion weights to reduce the bias. However, the concrete text process depends on the details of the corresponding experiments, and all codes are showed in the Appendix.

Simply drawing the number of reviews each day for four movies respectively, dramatically from the plot below, we can see during the April with *Avengers*, the reviews concentrated, while the newest movie is incredible more than before.

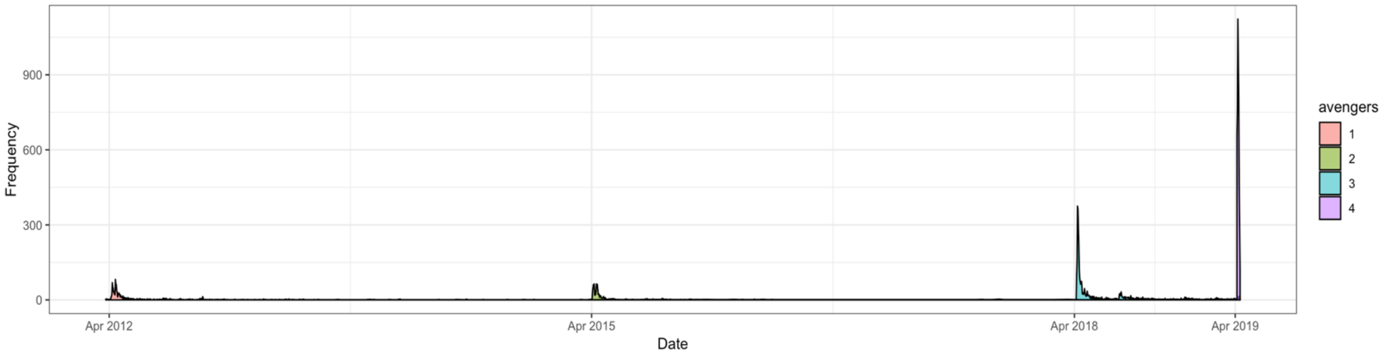


Figure 5 Number of daily Reviews

As for rating, the structure of it below (left) is obvious that except the bad Avengers 2, the 10/10 rate dominates the ratings, and especially in the Avengers 4 the proportion now hits the peak, which is over 60%. As a result, the differences between the fourth and the ones before will account a lot for the ratings and popularities.

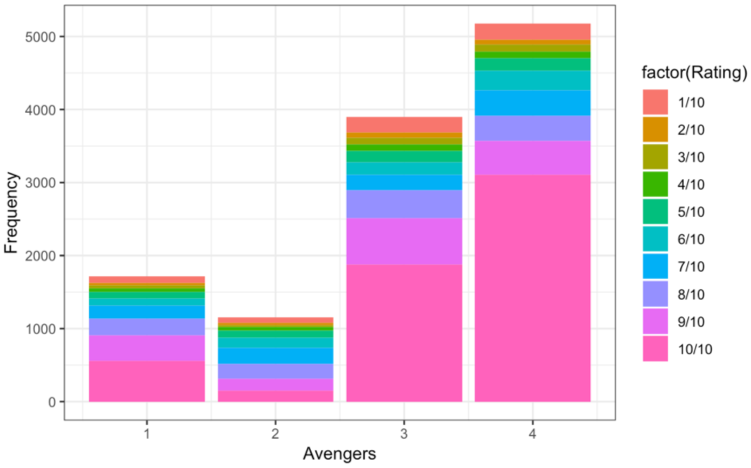


Figure 6 Rating Structure

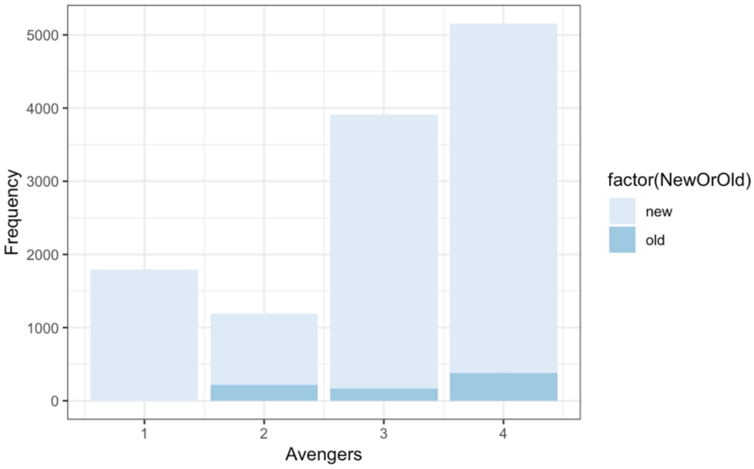


Figure 7 Audience Structure

On the other hand, the audience structure, which shows whether the user who reviews a later movie also reviews the one before, clearly show that more and more new fans are getting into this movie. It is exactly a topic maker, which can attract more ordinary people to view it and enjoy it.

Lastly, the essay shows below distributions of the lengths of texts (tokens) for the four movies. Out of expectations, we are surprised to view that it seems along with the rise of the rating of a movie, the average length of the text is less, although there are similar numbers of long comments for them. It might be result of the truth that better movie will lead to more short quick explicit comment like "A good movie!" or "Excellent!". But for the one that is not too good, it will take longer time to criticize or defend for it.

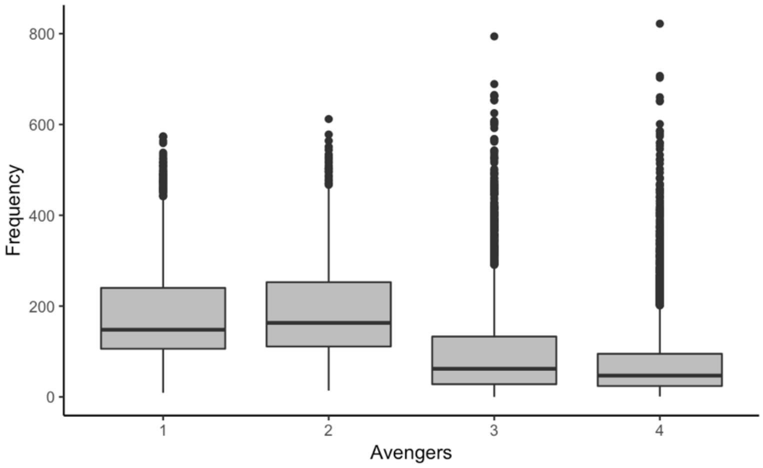


Figure 8 Review Length Distribution

4. Sentiment Analysis

The first experiment the essay makes is Sentiment Analysis. The essay goes deep into different dimension of the sentiment to see the specific difference the four movies holds, including *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*, *positive*, *sadness*, *surprise*, and *trust*. These ten attributes are from Mohammad and Charron's (2010, 2013) English version of the NRC Word-Emotion Association Lexicon. The essay utilizes dictionary methods to compute the 10 attributes. The results are showed below.

The first movie *Avengers 1* (Red) has the most praise and joys with it, and because it is the first movie assembling super heroes in the 21th century, people tend to think it is fun. Meanwhile, the third movie *Avengers 3* shows more fear than others, which might be a result of the plot that Thanos removed a half population in the world, including the main characters, which also earns more surprises.

Also as indicated roughly by the popularities, *anticipation* makes much more sense. The 3rd and 4th movie is the highest and the 2nd, which is seen to be the worst in the series, really gets a lower score among all.

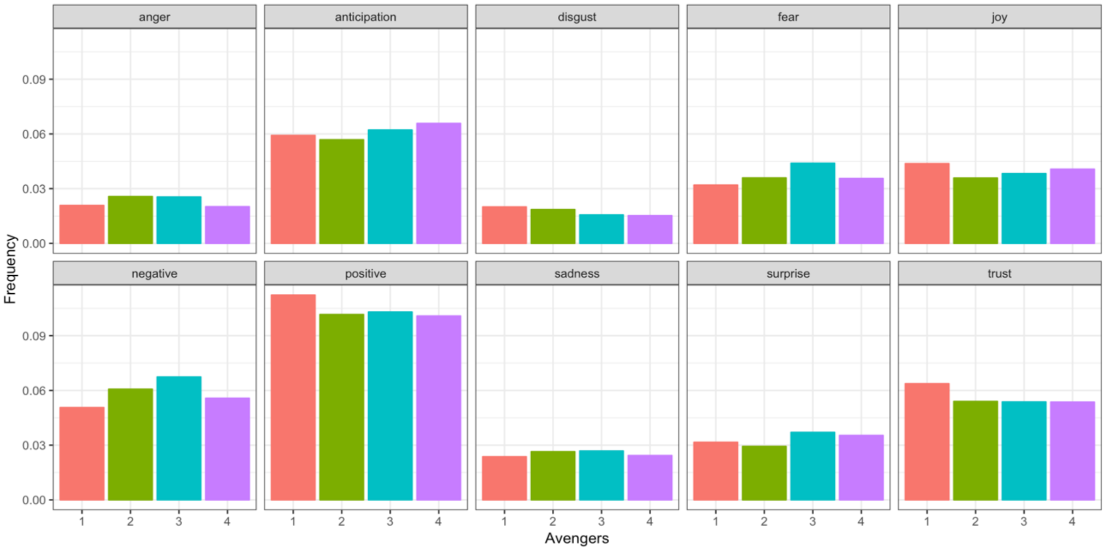


Figure 9 Emotion Result

The essay then compares the distribution solely of *positive* and *negative*, to figure out the detailed differences of the four movies. The result showed below uses density rather than the real counts. It is amazed to see that right as the rating of the 4th movie is really high and the "10/10" makes the most of percentage, the texts also account for it, with the *Negative* index to be high for left and low for right, as well as the *Positive* index to be higher for the right tail. This suggests that text does hold related information for predicting the ratings, and motives us a lot to carry on the text mining.

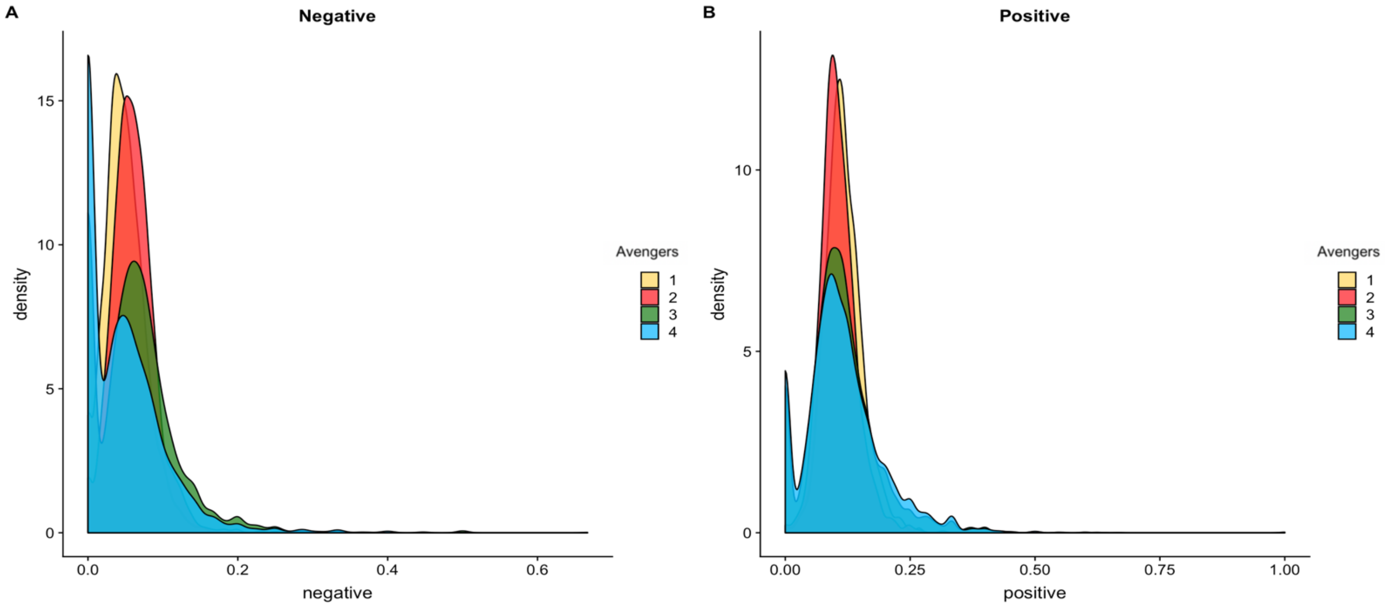


Figure 10 Sentiment Distribution

Lastly, the essay makes the time series plot for the four movies review sentiment. I take the mean by day of review sentiment for each movie respectively. The results are showed below. *Avengers 1* seems to be always more positive, and *Avengers 2* is getting worse in the later time, while Avengers 3 is just the opposite, which is better with time going by. These emotion data will be used in the later classifier model, as features to input. Hopefully they can be used as good predictors.

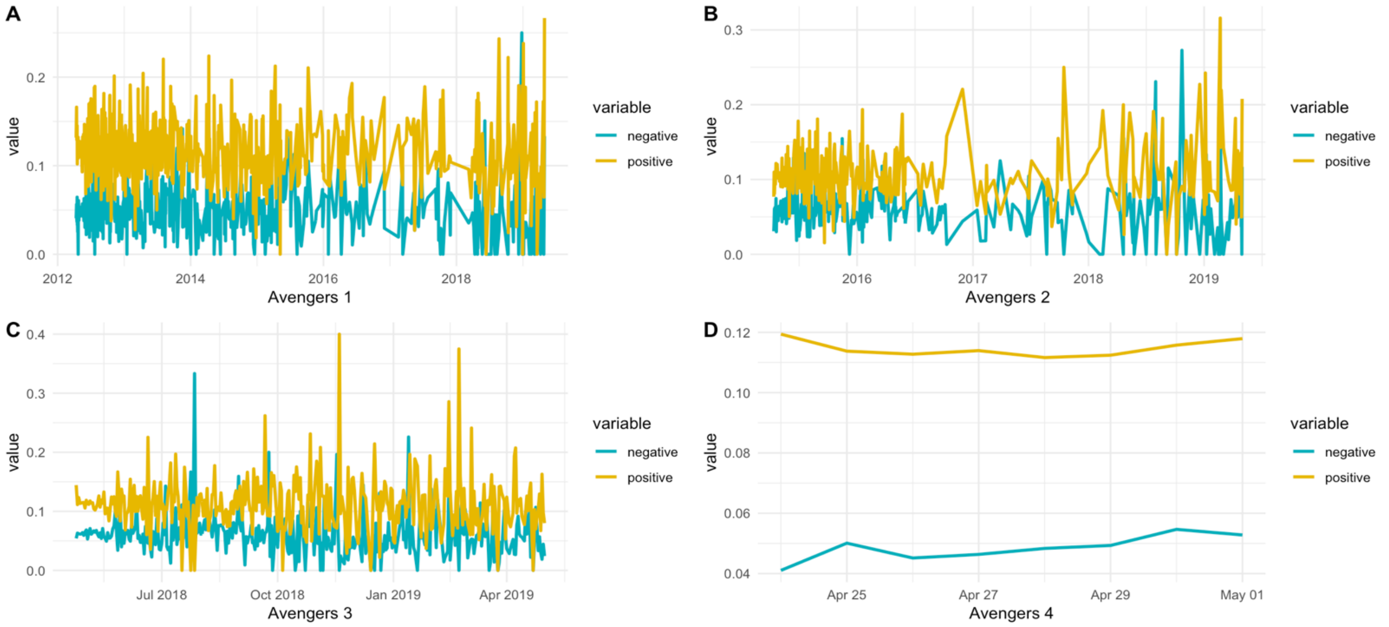


Figure 11 Sentiment Time Series

5. Topic Modelling

The essay hereby takes topic modelling to compare the topics that people would like to talk about each movie. We hereby utilize LDA model to get the topics and their scores of each text.

The number of topics the essay choose is by perplexity evaluation. I compute the value for each potential topic number, and then I choose the likely "elbow point" which has the largest gap right at that time. Here as the graph below, I could choose 20 as the topic numbers, where is not far from the value for 30, but is quite far from 10.

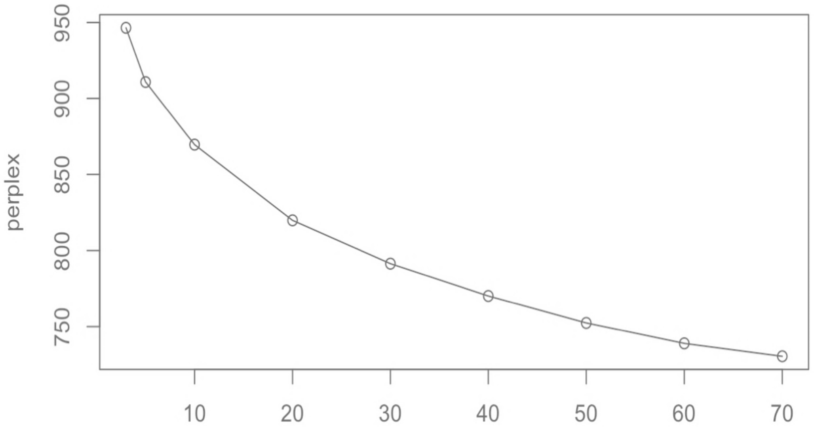


Figure 12 Perplexity Line

After 5000 iterations of the 20 topics, the result turns out as below. I use the basic movie plot knowledge and related information to divide the topics into several parts. Some are very obvious. For example, Topic 3 (orange) is mainly about *Avengers 2* (*ultron*, *age*, *witch* and *stark* are the most important figures for that movie), Topic 5 (blue) is mainly about *Avengers 3* (*thano*, *infinity*, *war* and *stone*), Topic 7 (green) is about *Avengers 1* (*loki*, *hulk*, *chris*, *downey*, and *iron*) and Topic 15 (purple) is mainly about *Avengers 4* (*end*, *final*, and *endgame* of course).

For others, Topic 1 is about hero the concept itself, including *super power* and *flight*. Topic 2, Topic 4 and Topic 19 talk about the action: watching the movie. Topic 6, Topic 10, Topic 16, and Topic 17 are series of expressing words with no sense, while Topic 9 and Topic 11 is about to analyze the movie in different views (i.e. actions, effect, scene and story). Topic 12 is many numbers about time length of the movie and ratings as well. Topic 13 is about the main roles' name in the movie. Topic 14 is related to comic fans. Topic 18 is pure praise to the movie. Topic 20 is, on the opposite, the mock or dislike to the movie.

[Spoiler Alert] Topic 8 is so interesting that it even describes the story of Avengers 4, which is very correct: Tony makes the time travel, goes back to get and use the stones, and dies at the end.

This LDA model is confirmed to be highly effective. It represents reviews in the long run quite well.

Table 1 Topics Presentation

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
hero	movi	aveng	film	thano	like	aveng	time	action	love
world	watch	ultron	charact	infin	realli	loki	back	great	end
super	made	first	aveng	war	just	whedon	toni	good	amaz
power	expect	age	much	univers	good	hulk	end	effect	perfect
come	make	new	one	stone	lot	chris	travel	well	everi
togeth	peopl	action	mani	marvel	much	man	past	scene	laugh
can	enjoy	witch	expect	strang	feel	downey	get	act	cri
fight	part	vision	first	half	felt	iron	use	job	thank
one	end	scarlet	part	guardian	littl	thor	die	special	everyth
aveng	great	stark	enjoy	aveng	think	mark	stone	stori	emot

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
charact	10	thor	comic	mcu	just	even	movi	see	like
stori	2	man	marvel	end	one	yet	best	watch	bad
mani	3	hulk	book	endgam	know	set	marvel	go	just
plot	hour	captain	aveng	war	go	manag	one	wait	plot
make	just	iron	like	infin	thing	audienc	ever	just	good
develop	1	america	superhero	emot	get	feel	seen	worth	noth
villain	review	scene	fan	year	can	deliv	superhero	next	even
well	time	fight	big	fan	say	previous	time	time	guy
main	rate	aveng	dark	marvel	even	though	everi	say	cgi
work	give	black	stori	final	think	particular	made	theater	joke

Next, the essay carries the time series of Topic 3 (orange), Topic 5 (blue), Topic 7 (green) and Topic 15 (purple), which are *Avengers 2*, *Avengers 3* , *Avengers 1* and *Avengers 4*, as showed below. It is noted that for the clear purpose, the data is aggregated monthly.

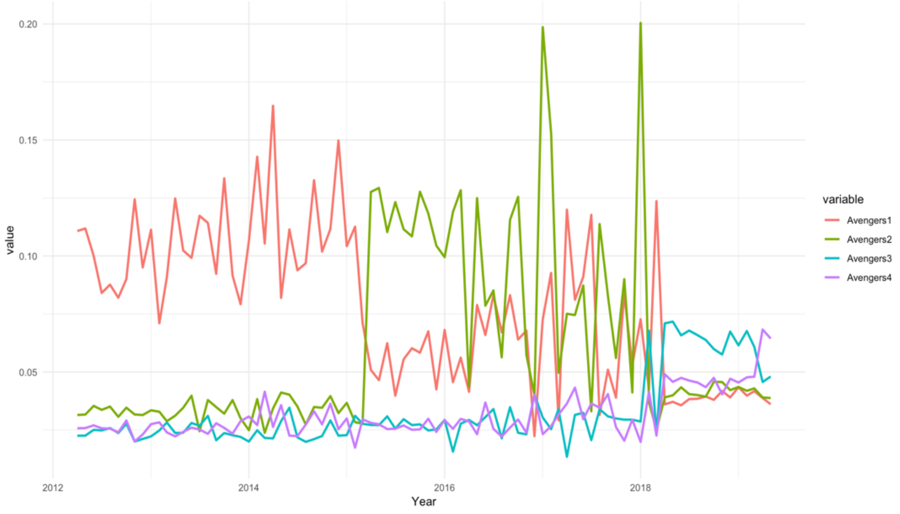


Figure 13 Topic Time Series

It turns out that the pattern is just alike what we assume. The *Avengers 1* Topic is more active in early days, while the others are not even shown up. Once the *Avengers 2* comes out, it dominates for quite a long time. At the last few days in 2019, *Avengers 4* Topic hits the peak and *Avenger 3* also shows up a bit again.

6. Rating Classifier

The essay collects the features constructed by sentiment analysis and LDA topic modelling. Each review now also has many text features, which will take more possible experiments. In order to further understand the use of the texts, the essay now introduces rating classifiers, which uses the texts input to predict the rating it may give. This classifier will not only give the rates, but also it will tell the value of the features and the importance among them. We would like to use that as an inner explanation of each text feature, what is why we choose Elastic Net and Random Forest, the two models that are highly interpretable and give suitable illustration for the text features. We will be also willing to see this classifier might have a good result in order to help producer and investor to further understand the market, which is what kind of topics that audiences like to hear and talk.

6.1 Elastic Net

The essay firstly uses elastic-net to try to pick out the useless variables in order to avoid overfitting. This method will converge some coefficients to zero, in which case the corresponding features are not going to be in the model. It combines both the variable of Lasso (L1) and the variable of Ridge (L2), which confirms the convergence and the efficiency. The loss function is as below.

$$L = \sum_{i=1}^{N}\left(y - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{J}\beta_j^2 + \lambda_2 \sum_{j=1}^{J}|\beta_j|$$

The y is the actual value of the dependent variable. β_0 and β_j is the coefficient for predictors. The N is the number of observations, and J is the number of predictors. The last two components are the L1 sand L2 term for penalty the complexity of the function, with λ_1 and λ_2 to be the weights to control how many we like to focus on penalty. The Elastic-Net, with the parameters automatically choosing the least deviance by Cross Validation result, is showed below. It should be noticed that there are two dashed lines in the graph. The one with a little bit higher deviance is within one-standard-error of the optimal one, but because it can be less computational, the essay chooses that as the final lambda.

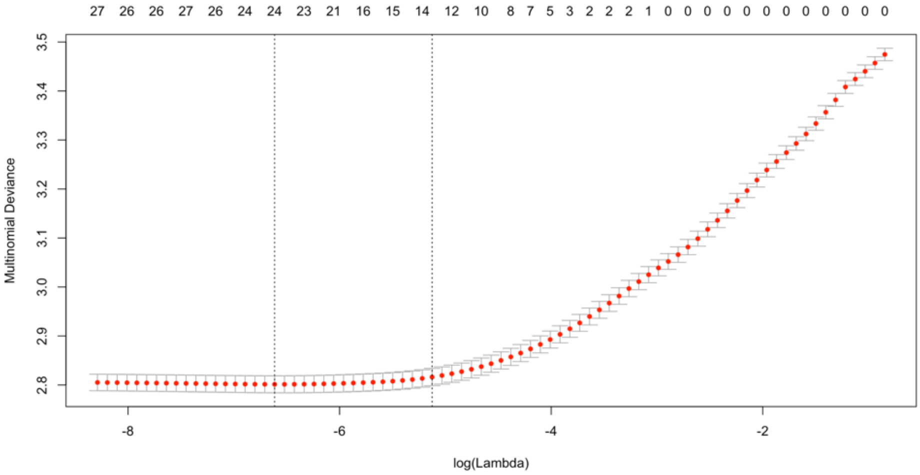


Figure 14 Elastic-Net Variable Selection

The classifier result is not too good, with recall only 0.5 and accuracy only 0.47. But since we have 10 classes to divide and text mining result does not totally care the deterministic statistics, the results for feature explanation that can be made here are rather important.

We focus on the worst case (1/10) and the best case (10/10), as indicated by the table blow. For the worst case, it turns out that *anger*, *fear*, *negative*, *sadness*, and Topic 20 (which is mocking the movie) will enhance the probability to give a bad rating (the coefficients of which are positive). On the contrary, Topic 9 which is praise to many stages of a movie will help leaving from the worst case (the coefficients of which are negative). On the other hand, *anticipation*, as showed in the EDA part, is very good for best case, and the topic 10, which are all very good praises, do extremely help the best case, while Topic 20, as always, seriously prevent rating to be best case.

From this result, it is clear to see that text data does account for the ratings. And some of the kinds of emotions and some special topics do have its special effects towards the ratings.

Table 2 Elastic Net Presentation

	1/10	10/10		1/10	10/10		1/10	10/10
(Intercept)	-1.4988061	1.93974804	V1	.	-1.6981149	V11	-0.7711135	-3.3890084
anger	**1.9041026**	**.**	V2	5.5717864	2.0550092	V12	2.8707351	-5.4728622
anticipation	**.**	**1.77893923**	V3	.	-1.8662914	V13	0.3485692	.
disgust	.	-1.6065316	V4	-1.6617243	.	V14	3.4022042	-1.6351024
fear	**0.3847649**	**0.04943408**	V5	.	1.19398438	V15	.	6.26250028
joy	.	.	V6	-1.6790824	-10.255208	V16	.	2.24152095
negative	**2.127439**	**-5.1797401**	V7	-0.8518799	2.78899184	V17	-2.7973566	-1.9074574
positive	.	.	V8	1.0187491	-4.3703359	V18	1.5242185	21.6512534
sadness	**1.7621788**	**.**	**V9**	**-3.4708001**	**2.16888073**	V19	2.0291101	5.42930327
surprise	.	-0.8813468	**V10**	**.**	**30.1673754**	**V20**	**10.6264795**	**-33.983317**
trust	.	.						

6.2 Random Forest

The essay then introduces random forest method, which is a bootstrap collection of the trees with different predictors. This is pruned by minimize its corresponding loss function. The loss of a predictor removing will account the importance of it, which in our case is the real influence of one text feature. The loss function for Gini index is as below.

$$L = \frac{1}{B} \sum_{b=1}^{B} \sum_{j=1}^{J} \mathrm{p}\,(1-p) + \alpha |T|$$

The p hereby is the proportion of training observation in the related region that exactly is from the corresponding most common class. $\sum_{j=1}^{J} \mathrm{p}\,(1-p)$ computes the Gini and the $|T|$ is the number of nodes that represents the complexity. The α here is also the weight for that. It will depend how much we are concerning the complexity.

Like the Elastic Net, the classifier ability part may not account too much, but the importance calculated by the method I mentioned above really explain the features.

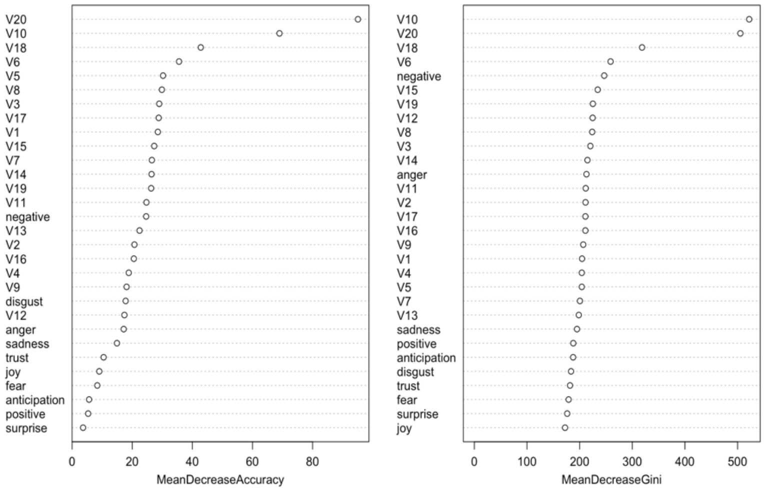


Figure 15 Importance Graph: Accuracy (left), Gini(right)

Here from two dimensions, we can conclude that Topic 20, Topic 10, and *negative* are likely to be very important predictors for the rating, which further influence the tendency and popularity of a movie. The *surprise*, *joy* and things alike seem to be not very important. These inspire that the producer should take more attention to the bad remarks, or negative thoughts of the movie, and the joy and surprise issue may also account a lot, but these are the core of a Hollywood movie, so it seems from the graph not the primary.

7. Conclusion

In conclusion, the essay finally makes the Avengers movie reviews analysis, from the very beginning, data scraper, which gives me all the reviews towards Avengers within 11 years. The essay then takes sentiment analysis, including dictionary methods and time series analysis. By sentiment conducting, the essay acquires 10 directions of emotion data for each of the text. From it, the essay looks for potential sentiment pattern that this data should have and lead to more possible experiments. Furthermore, the essay carries topic modelling, where the elbow point points number of 20, which gives 20 very informative topics. With these topics, the essay learns from it and picks out funny results, say some topics are focusing on one movie, which are then be confirmed by the time series graph.

After feature engineer, the essay utilizes Elastic Net and Random Forest to act as classifiers to try to predict the right rate that a review should have. Although from the perspective of prediction, they perform not too good, the view of text features is even clearer and tell many stories. We suggest that some features are more important than others, and their corresponding influence directions as well.

The essay completes the whole process from data acquiring, data preprocessing, feature engineering, and experiment analysis. All of them are big parts and since it is brand-new I will have to re-collect data again and again. For further research, the more data acquired might lead to the more interesting points, and some algorithms that accounts less for feature explanations can also be used.

8. Acknowledge

The essay hereby thanks Prof. Pablo Barberá for his help in need, comfort in deep, and the whole great course MY459, where I learned solid quantitative text analysis methods, systematically and fully. The essay should also appreciate Dr. Gokhan Ciflikli, who helped my R implementation and the *Quanteda* issues, which really tied me down at first. The essay would also like to thank Marvel Entertainment, LLC, Robert Downey Jr. and all the others, who devote their incredible 10 years to send all fans such a big present.

9. Reference

[1] Basari, A. S. H. , Hussin, B. , Ananta, I. G. P. , & Zeniarja, J. . (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53(Complete), 453-462.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[3] Dupuy, C. , Bach, F. , & Diot, C. . (2017). Qualitative and descriptive topic extraction from movie reviews using lda.

[4] Joshi, M. , Das, D. , Gimpel, K. , & Smith, A. N. . (2010). Movie Reviews and Revenues: An Experiment in Text Regression. Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. DBLP.

[5] Kennedy, A. , & Inkpen, D. . (2010). Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence, 22(2), 110-125.

[6] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[7] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[8] Pimwadee Chaovalit, L. Z. . (2005). Movie review mining: a comparison between supervised and unsupervised classification approaches. Hawaii International Conference on System Sciences. IEEE.

[9] Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science, 36(6), 823-848.

[10] Zhuang, L. , Jing, F. , & Zhu, X. . (2006). Movie review mining and summarization.

[APPENDIX: Reproducible codes]

Scrape Man.Rmd : Data scraper

Captain Explore.Rmd: Explore the data

Doctor Analyze.Rmd: Sentiment analysis and topic modelling

Classifier.Rmd: Classifier of elastic-net and random forest