Data Capture/NLP Project

Ishan Ambike
3/18/2022

Introduction:

The focus of this project is to capture data and perform NLP on the data to gain insights. For this project, I am doing an analysis of The Batman (2022) movie. This movie was released recently and has been a popular topic of discussion on social media. To perform the analysis, I am going to use recent tweets from the general public and also scrape critic reviews from the Rotten Tomatoes website. By analyzing both I can get the overall reaction to the movie and also see if critics' reactions differ from the general public.

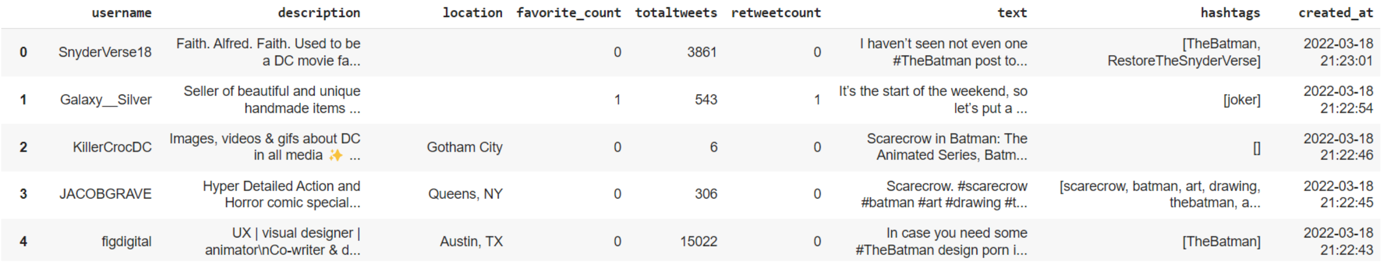
Data Captured:

To analyze the reaction to The Batman movie, I have collected public reactions from tweets and critics' reactions from the Rotten Tomatoes website.

Tweets collection:

As Twitter is a platform where people express their emotions and reactions, I have collected tweets regarding the movie to see what general reactions to the movie are. *#TheBatman* is the most popular hashtag on Twitter used for this movie. So, I have collected the most recent 2000 tweets containing this hashtag and excluded the retweets. Since the movie was released on 4th March 2022, I have collected the most recent tweets after this date. For each tweet, I have collected data like username who posted the tweet, user description, user location, favorite count(likes), user total tweets, number of retweets for the tweet, tweet text, hashtags used, and created time.

Sample Tweets collected:

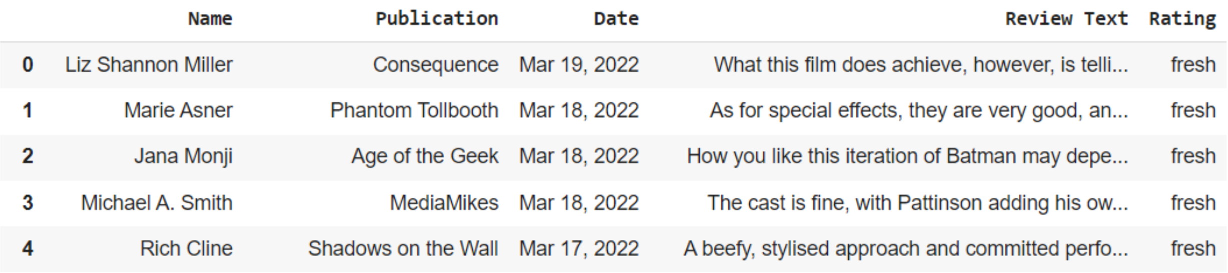


	username	description	location	favorite_count	totaltweets	retweetcount	text	hashtags	created_at
0	SnyderVerse18	Faith. Alfred. Faith. Used to be a DC movie fa...		0	3861	0	I haven't seen not even one #TheBatman post to...	[TheBatman, RestoreTheSnyderVerse]	2022-03-18 21:23:01
1	Galaxy__Silver	Seller of beautiful and unique handmade items ...		1	543	1	It's the start of the weekend, so let's put a ...	[joker]	2022-03-18 21:22:54
2	KillerCrocDC	Images, videos & gifs about DC in all media ✨ ...	Gotham City	0	6	0	Scarecrow in Batman: The Animated Series, Batm...	[]	2022-03-18 21:22:46
3	JACOBGRAVE	Hyper Detailed Action and Horror comic special...	Queens, NY	0	306	0	Scarecrow. #scarecrow #batman #art #drawing #t...	[scarecrow, batman, art, drawing, thebatman, a...	2022-03-18 21:22:45
4	figdigital	UX | visual designer | animator\nCo-writer & d...	Austin, TX	0	15022	0	In case you need some #TheBatman design porn i...	[TheBatman]	2022-03-18 21:22:43

Rotten Tomatoes reviews scraping:

I have scraped data from the page https://www.rottentomatoes.com/m/the_batman/reviews. It contains reviews of 437 critics in brief. If a critic writes a good review for the movie then it is rated as fresh or else rotten for a bad review. The publication in which the detailed review is posted and the date of review is also provided.

Sample reviews:

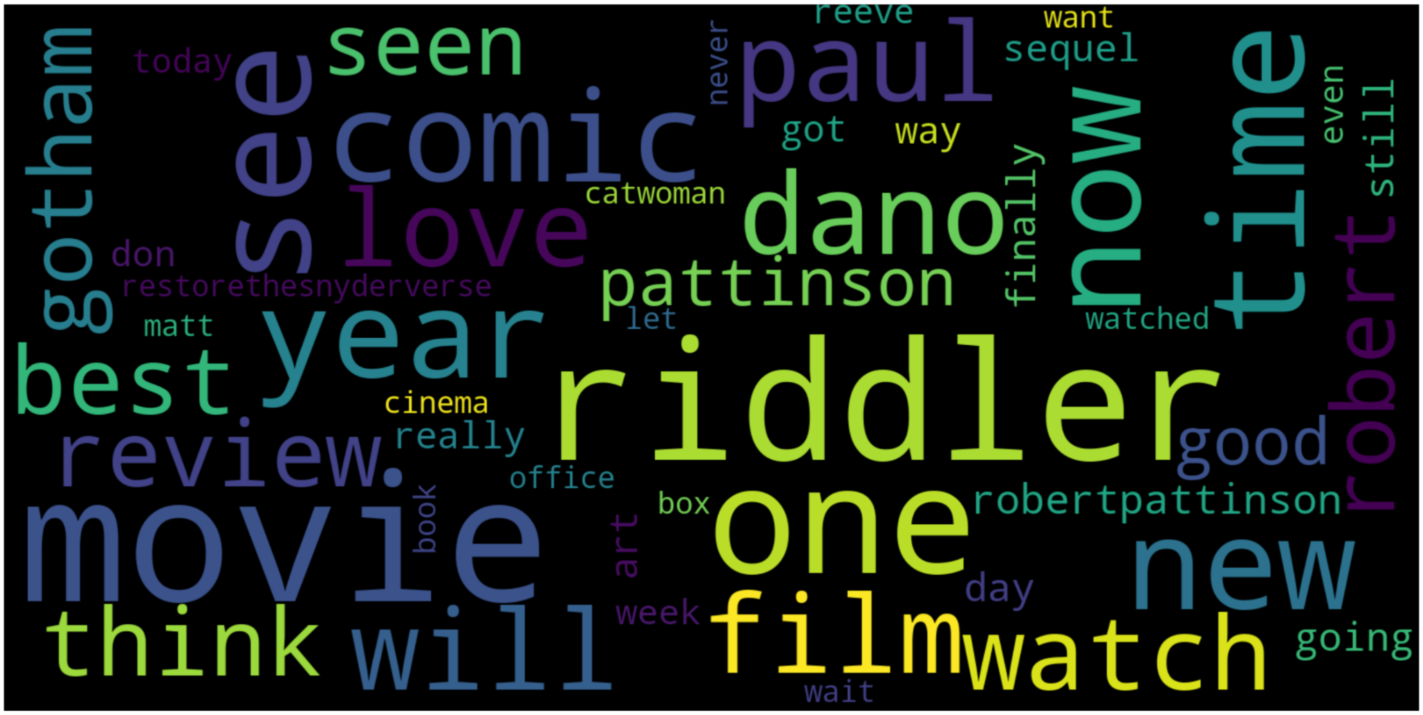


	Name	Publication	Date	Review Text	Rating
0	Liz Shannon Miller	Consequence	Mar 19, 2022	What this film does achieve, however, is telli...	fresh
1	Marie Asner	Phantom Tollbooth	Mar 18, 2022	As for special effects, they are very good, an...	fresh
2	Jana Monji	Age of the Geek	Mar 18, 2022	How you like this iteration of Batman may depe...	fresh
3	Michael A. Smith	MediaMikes	Mar 18, 2022	The cast is fine, with Pattinson adding his ow...	fresh
4	Rich Cline	Shadows on the Wall	Mar 17, 2022	A beefy, stylised approach and committed perfo...	fresh

Analysis from tweets:

Frequently used words:

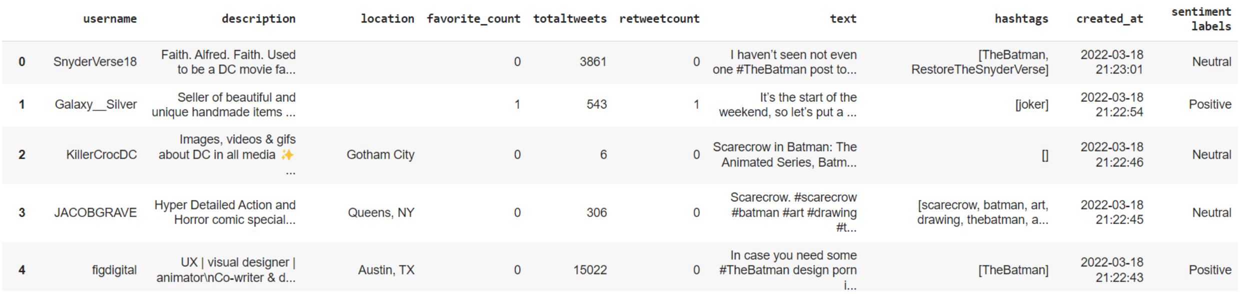
Firstly I have found out the most common 50 words which are being used in tweets regarding the movie. By looking at these words we can get an idea of what is being said about the movie. I have visualized these words in a word cloud.



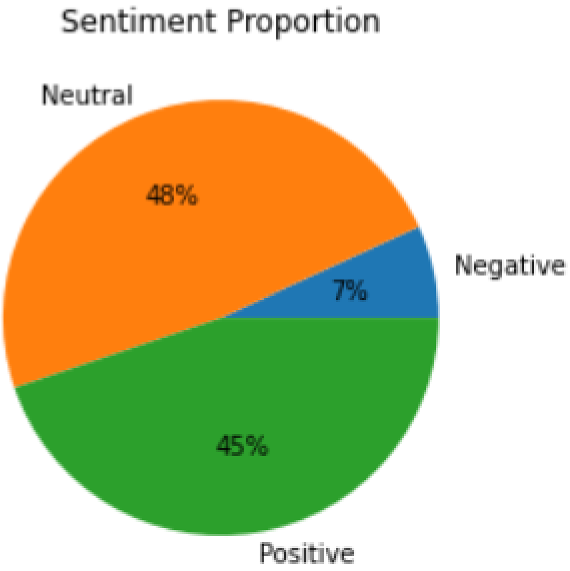
So from the word cloud, we can see that people have used words like watch, best, good, watched, seen, see, love, sequel, etc. These words indicate that people are liking the movie. Cannot find any negative word in this word cloud. People are also talking about the characters, places, actors, and director from the movie.

Sentiment Analysis:

I have done sentiment analysis of the tweets using the model described in https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis. By analyzing the tweet text, each tweet is labeled as either positive, negative, or neutral.

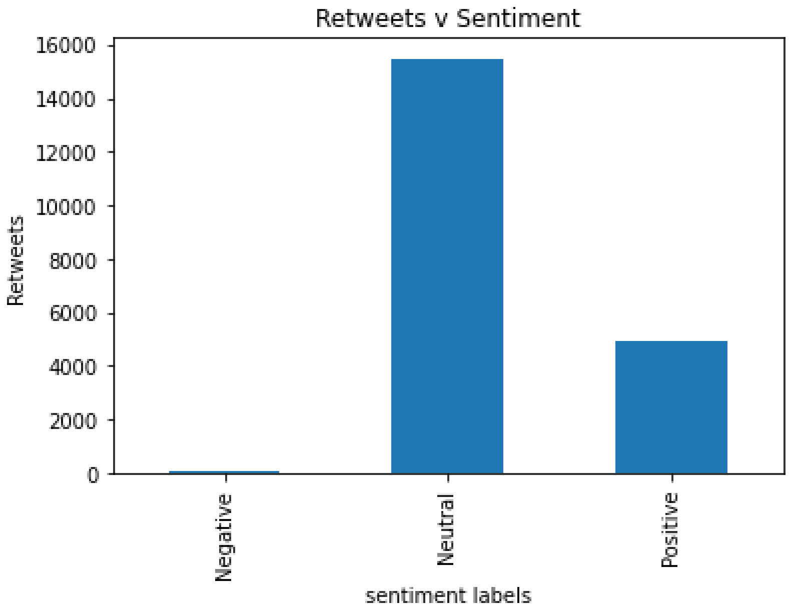


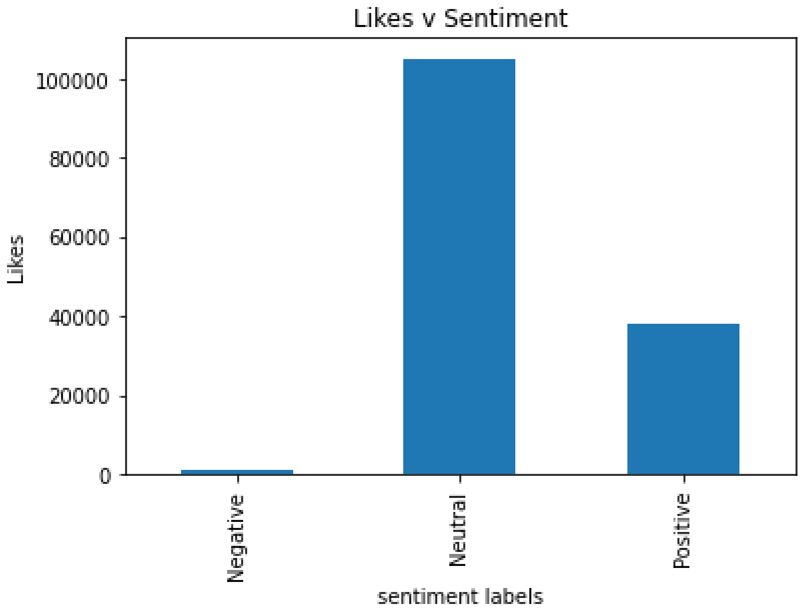
	username	description	location	favorite_count	totaltweets	retweetcount	text	hashtags	created_at	sentiment labels
0	SnyderVerse18	Faith. Alfred. Faith. Used to be a DC movie fa...		0	3861	0	I haven't seen not even one #TheBatman post to...	[TheBatman, RestoreTheSnyderVerse]	2022-03-18 21:23:01	Neutral
1	Galaxy__Silver	Seller of beautiful and unique handmade items ...		1	543	1	It's the start of the weekend, so let's put a ...	[joker]	2022-03-18 21:22:54	Positive
2	KillerCrocDC	Images, videos & gifs about DC in all media ✨ ...	Gotham City	0	6	0	Scarecrow in Batman: The Animated Series, Batm...	[]	2022-03-18 21:22:46	Neutral
3	JACOBGRAVE	Hyper Detailed Action and Horror comic special...	Queens, NY	0	306	0	Scarecrow. #scarecrow #batman #art #drawing #t...	[scarecrow, batman, art, drawing, thebatman, a...	2022-03-18 21:22:45	Neutral
4	figdigital	UX | visual designer | animator\nCo-writer & d...	Austin, TX	0	15022	0	In case you need some #TheBatman design porn i...	[TheBatman]	2022-03-18 21:22:43	Positive



From the above pie chart, we can see the proportion of sentiments. Neutral sentiment was classified for 48% of the tweets, positive sentiment for 45% of the tweets, and only 7% for negative sentiment. So from sentiment analysis also it seems that the movie mostly has a positive impression on the audience.

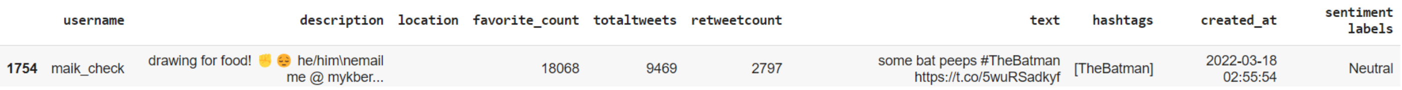
I have done an analysis to find which sentiment received the most retweets and likes.





Both the plots show similar patterns with neutral tweets getting the most retweets and likes.

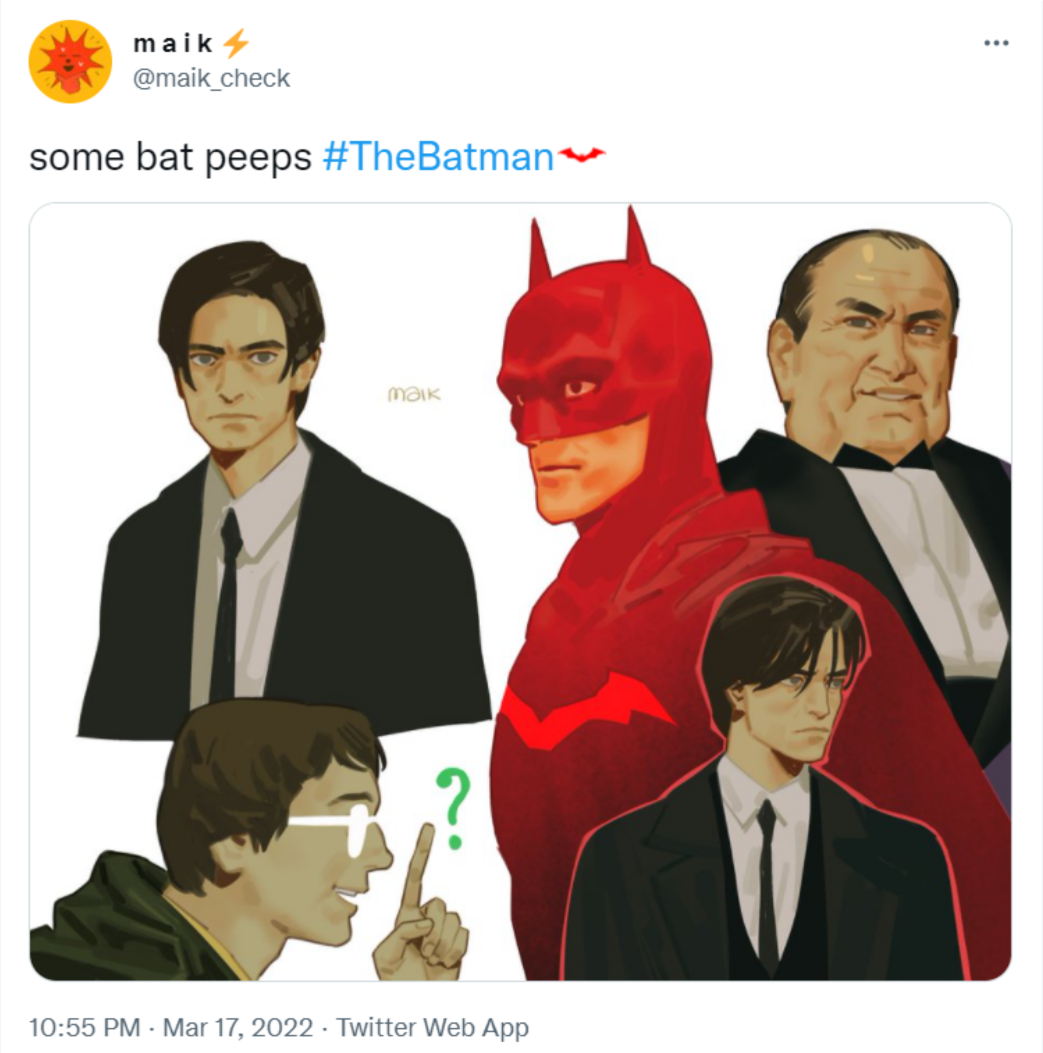
Now, let's check the tweet with the most likes and retweets from the 2000 tweets.



	username	description	location	favorite_count	totaltweets	retweetcount	text	hashtags	created_at	sentiment labels
1754	maik_check	drawing for food! ✊ 😔 he/him\nemail me @ mykber...		18068	9469	2797	some bat peeps #TheBatman https://t.co/5wuRSadkyf	[TheBatman]	2022-03-18 02:55:54	Neutral

This tweet has the highest likes and retweets. The tweet is classified as neutral.

The original tweet is:



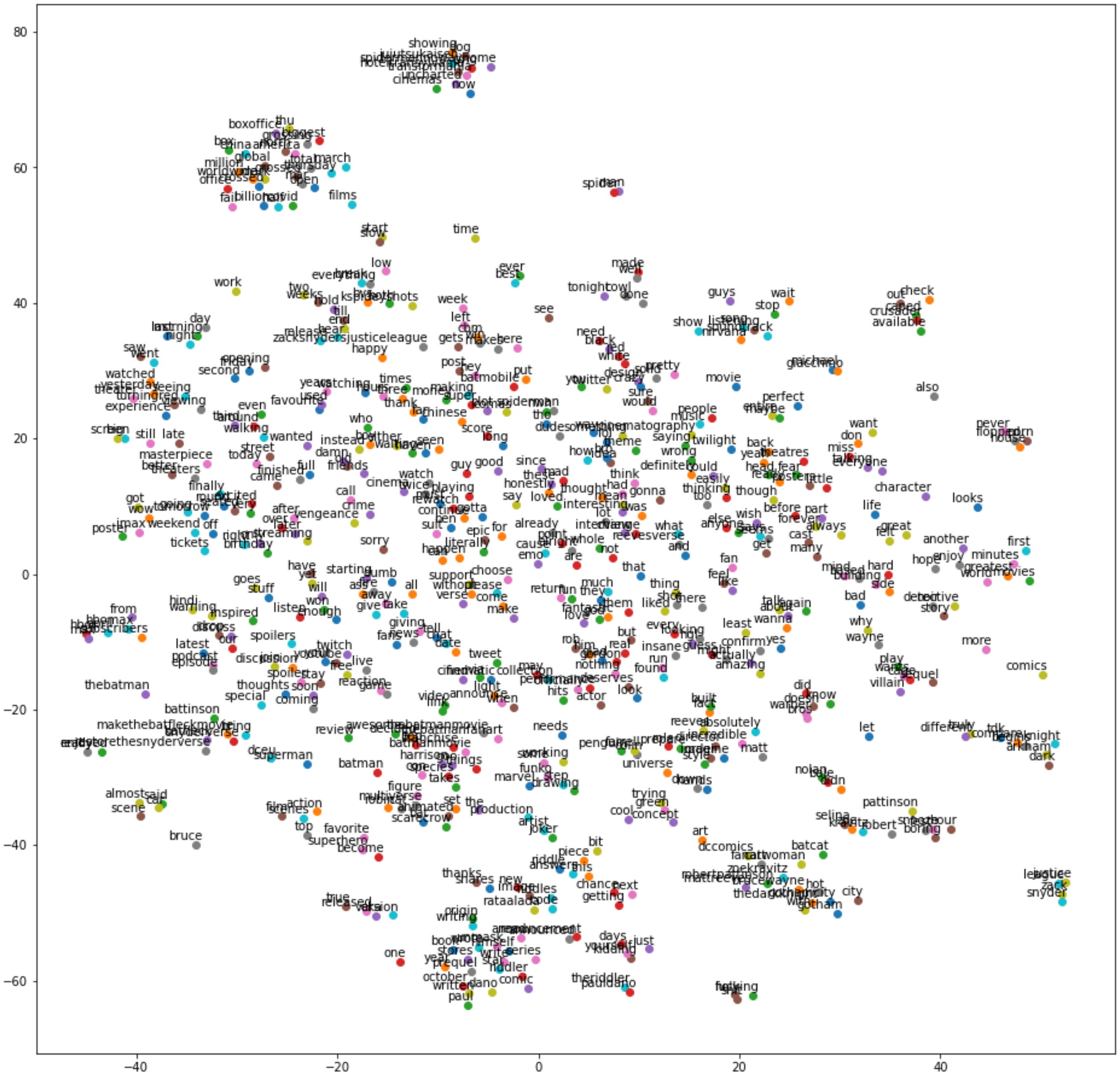
maik⚡
@maik_check

some bat peeps #TheBatman🦇

10:55 PM · Mar 17, 2022 · Twitter Web App

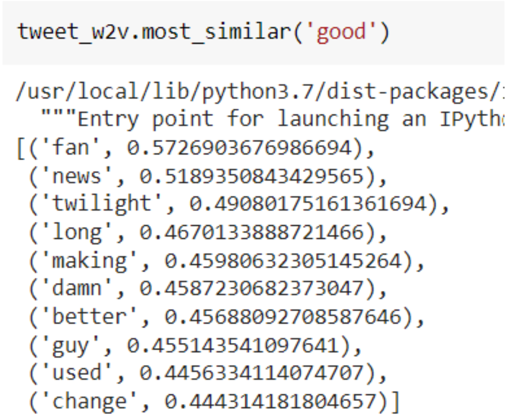
Word2Vec analysis on tweets:

Word2vec is a two-layer neural net that processes text by “vectorizing” words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep neural networks can understand. The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention. Given enough data, usage, and contexts, Word2vec can make highly accurate guesses about a word’s meaning based on past appearances. Those guesses can be used to establish a word's association with other words (e.g. "man" is to "boy" what "woman" is to "girl"), or cluster documents and classify them by topic. Those clusters can form the basis of search, sentiment analysis, and recommendations in such diverse fields as scientific research, legal discovery, e-commerce, and customer relationship management (https://wiki.pathmind.com/word2vec).



Using word2vec I am going to find the words people are using to describe the movie, characters, makers, and cast.

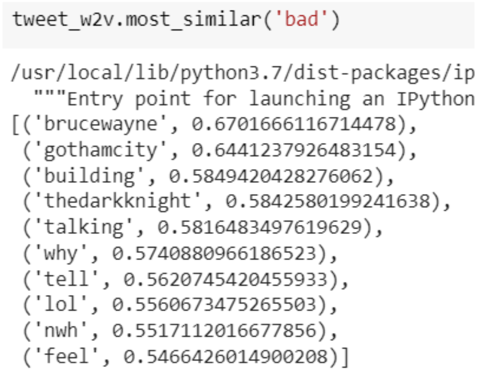
Words used similar to good:



```
tweet_w2v.most_similar('good')
```

```
/usr/local/lib/python3.7/dist-packages/
  """Entry point for launching an IPyth
[('fan', 0.5726903676986694),
 ('news', 0.5189350843429565),
 ('twilight', 0.49080175161361694),
 ('long', 0.4670133888721466),
 ('making', 0.45980632305145264),
 ('damn', 0.4587230682373047),
 ('better', 0.45688092708587646),
 ('guy', 0.455143541097641),
 ('used', 0.4456334114074707),
 ('change', 0.444314181804657)]
```

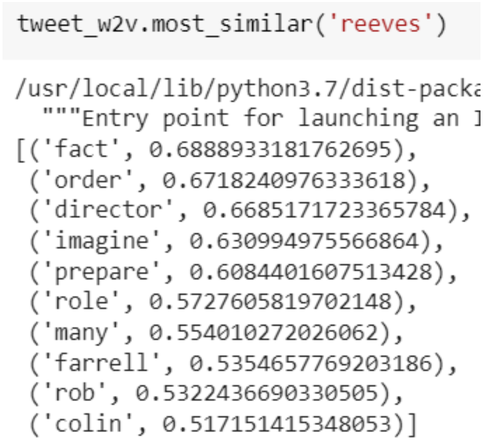
Words used similar to bad:



```
tweet_w2v.most_similar('bad')
```

```
/usr/local/lib/python3.7/dist-packages/ip
  """Entry point for launching an IPython
[('brucewayne', 0.6701666116714478),
 ('gothamcity', 0.6441237926483154),
 ('building', 0.5849420428276062),
 ('thedarkknight', 0.5842580199241638),
 ('talking', 0.5816483497619629),
 ('why', 0.5740880966186523),
 ('tell', 0.5620745420455933),
 ('lol', 0.5560673475265503),
 ('nwh', 0.5517112016677856),
 ('feel', 0.546642601490020 8)]
```

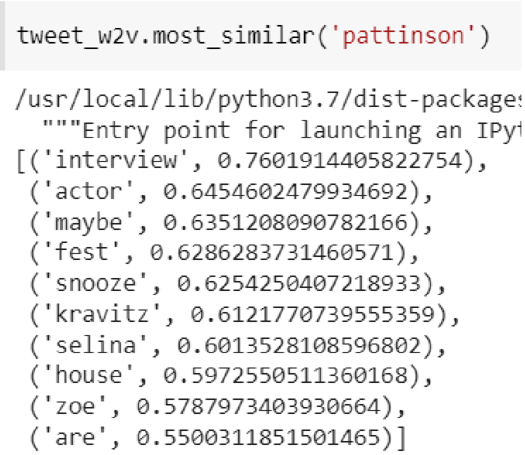
What words are associated with the director Matt Reeves:



```
tweet_w2v.most_similar('reeves')
```

```
/usr/local/lib/python3.7/dist-packa
  """Entry point for launching an I
[('fact', 0.6888933181762695),
 ('order', 0.6718240976333618),
 ('director', 0.6685171723365784),
 ('imagine', 0.630994975566864),
 ('prepare', 0.6084401607513428),
 ('role', 0.5727605819702148),
 ('many', 0.554010272026062),
 ('farrell', 0.5354657769203186),
 ('rob', 0.5322436690330505),
 ('colin', 0.517151415348053)]
```

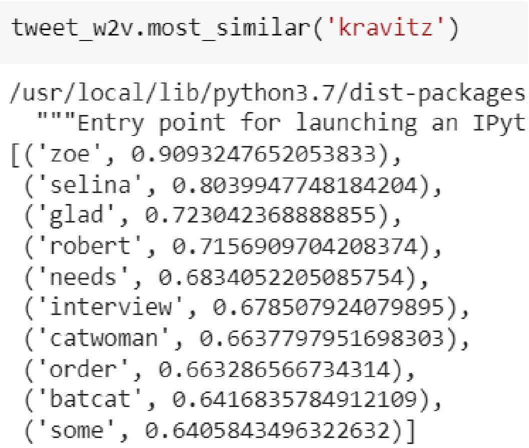
Words associated with the lead actor Robert Pattinson:



```
tweet_w2v.most_similar('pattinson')
```

```
/usr/local/lib/python3.7/dist-package
  """Entry point for launching an IPy
[('interview', 0.7601914405822754),
 ('actor', 0.6454602479934692),
 ('maybe', 0.6351208090782166),
 ('fest', 0.6286283731460571),
 ('snooze', 0.6254250407218933),
 ('kravitz', 0.6121770739555359),
 ('selina', 0.6013528108596802),
 ('house', 0.5972550511360168),
 ('zoe', 0.5787973403930664),
 ('are', 0.5500311851501465)]
```

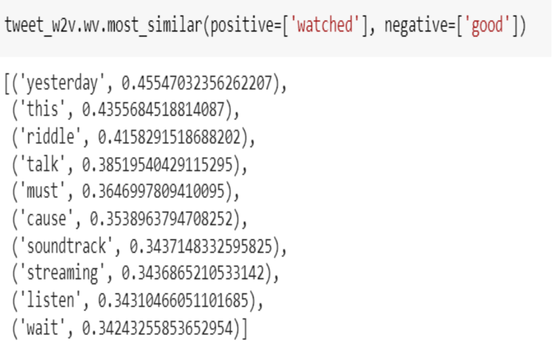
Words associated with the lead actress Zoë Kravitz:



```
tweet_w2v.most_similar('kravitz')
```

```
/usr/local/lib/python3.7/dist-packages
  """Entry point for launching an IPyt
[('zoe', 0.9093247652053833),
 ('selina', 0.8039947748184204),
 ('glad', 0.723042368888855),
 ('robert', 0.7156909704208374),
 ('needs', 0.6834052205085754),
 ('interview', 0.678507924079895),
 ('catwoman', 0.6637797951698303),
 ('order', 0.663286566734314),
 ('batcat', 0.6416835784912109),
 ('some', 0.6405843496322632)]
```

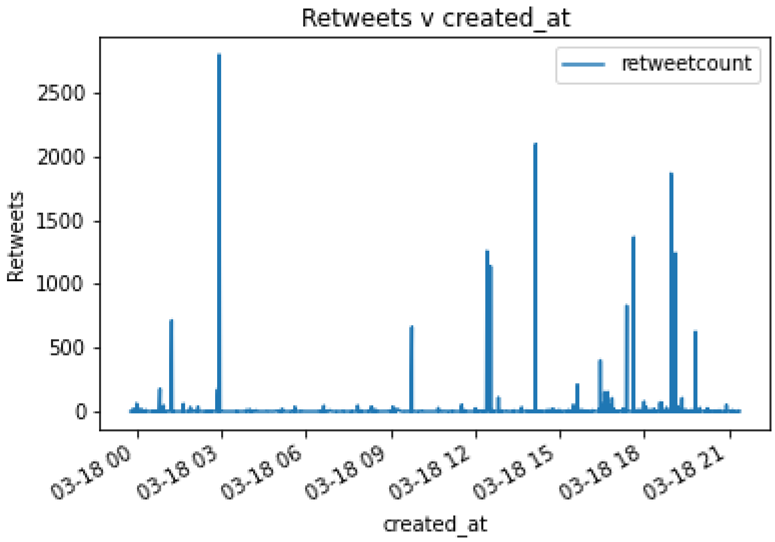
Words used by people who watched the movie but said it’s not good:

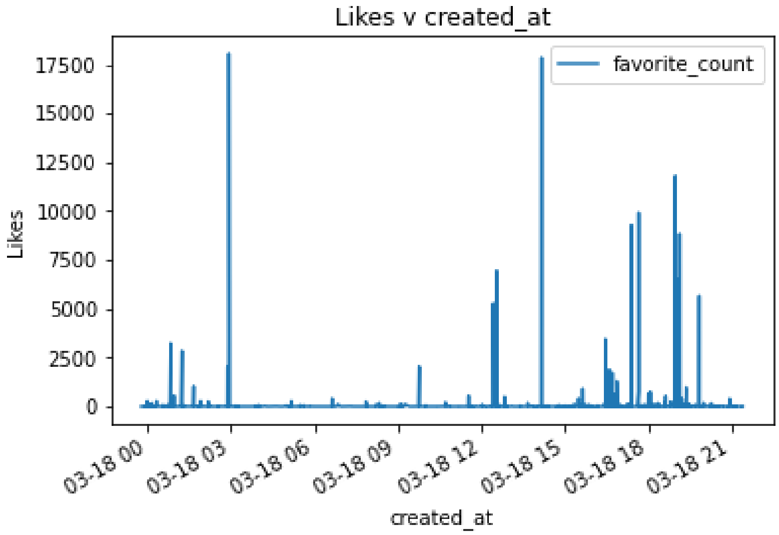


```
tweet_w2v.wv.most_similar(positive=['watched'], negative=['good'])
```

```
[('yesterday', 0.45547032356262207),
 ('this', 0.4355684518814087),
 ('riddle', 0.4158291518688202),
 ('talk', 0.38519540429115295),
 ('must', 0.3646997809410095),
 ('cause', 0.3538963794708252),
 ('soundtrack', 0.3437148332595825),
 ('streaming', 0.3436865210533142),
 ('listen', 0.3431046605110168 5),
 ('wait', 0.34243255853652954)]
```

Likes and retweets pattern:

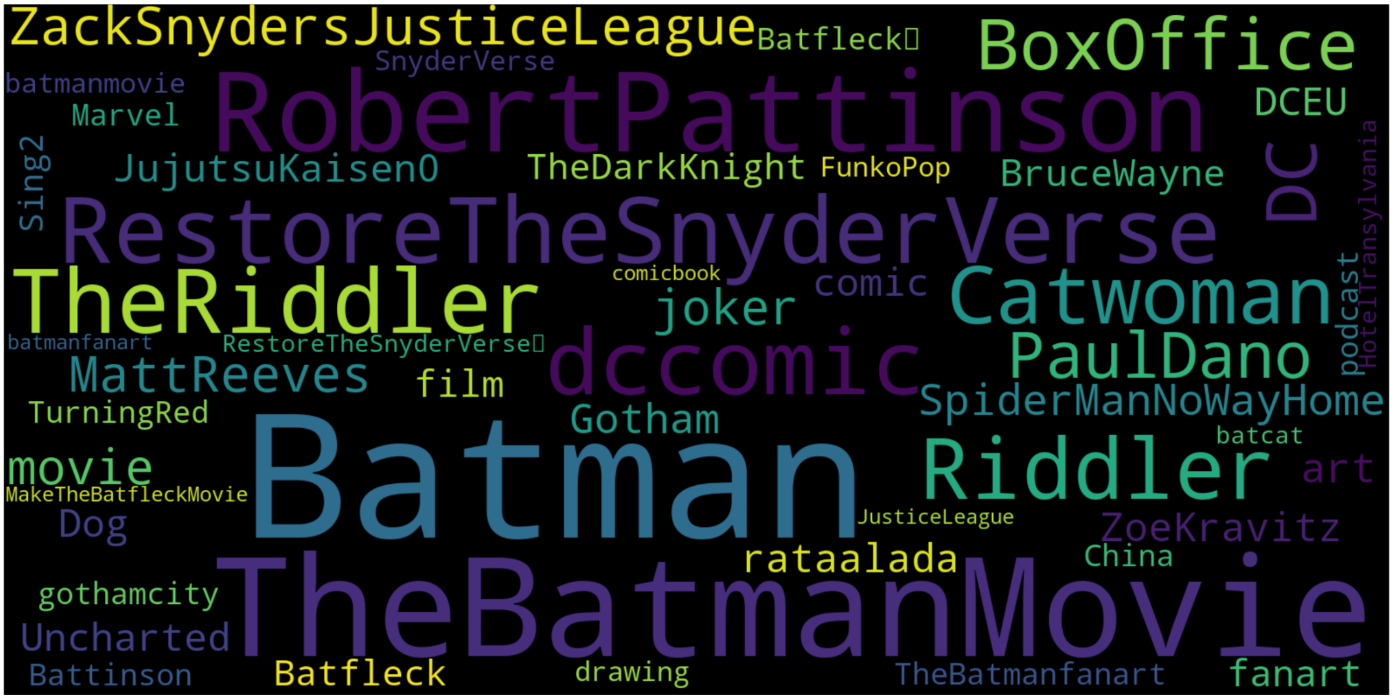




Similar patterns can be observed for likes and retweets with respect to time.

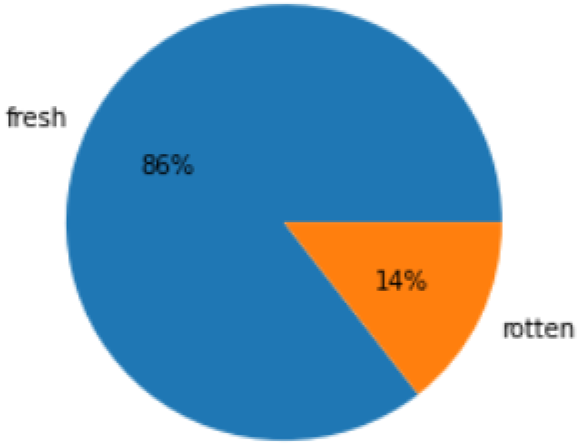
Popular hashtags used:

I have visualized the most popular hashtags used in tweets except #*TheBatman* with help of the word cloud.



Analysis from Rotten Tomatoes:

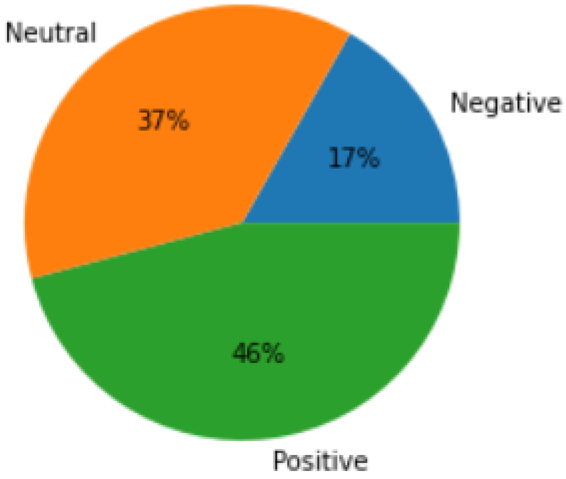
Analysis from Rotten Tomatoes ratings:



86% of the critics gave a fresh (good) rating. Only 14% of the critics gave a rotten (bad) rating. So just like audience tweets, it seems that most critics also loved the movie.

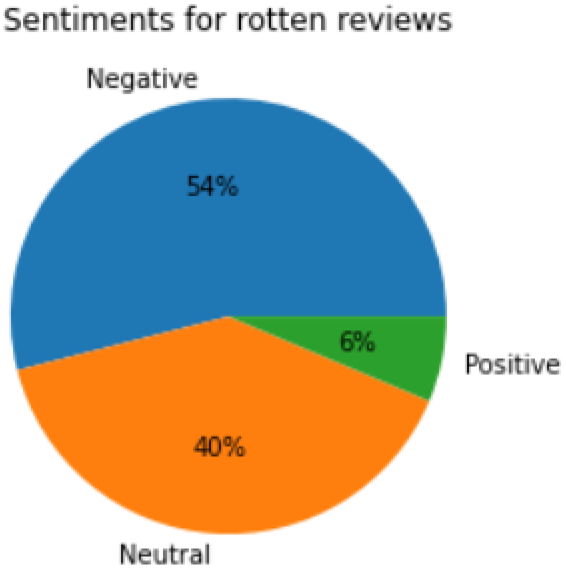
Sentiment analysis of the review text:

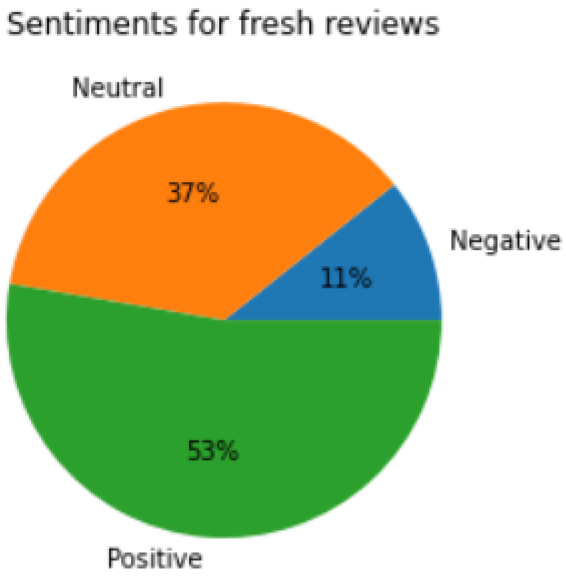
To perform sentiment analysis I have used the same model used for sentiment analysis of tweets (https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis).



46% of the reviews were classified as positive and 37% as neutral. 17% of the reviews were classified as negative which is close to the rotten rating at 14%.

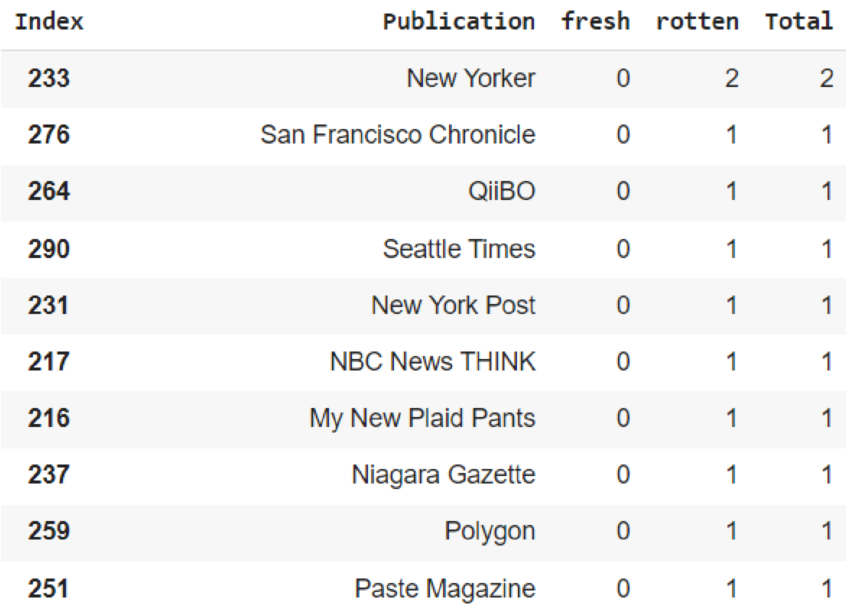
Further, I have checked how sentiments are classified for only rotten reviews and fresh reviews.





Fresh ratings have mostly positive review sentiments and rotten ratings have mostly negative review sentiments. Neutral sentiment has a significant chunk in both.

I have also found publications that gave all rotten ratings for the movie. There are 58 publications that gave all rotten ratings. Some of them are:



Index	Publication	fresh	rotten	Total
233	New Yorker	0	2	2
276	San Francisco Chronicle	0	1	1
264	QiiBO	0	1	1
290	Seattle Times	0	1	1
231	New York Post	0	1	1
217	NBC News THINK	0	1	1
216	My New Plaid Pants	0	1	1
237	Niagara Gazette	0	1	1
259	Polygon	0	1	1
251	Paste Magazine	0	1	1

Conclusion:

So from tweets analysis and rotten tomatoes scraping analysis, we can conclude that the movie is doing good and both critics and audience are liking it. The purpose of the project was to capture data and perform NLP. I gathered data from the Twitter API and scraped Rotten Tomatoes reviews. Further, I did sentiment analysis using the hugging face model and word2vec. I could find insights and visualized them to show the patterns.

Future Scope:

More analysis can be done by collecting more data, especially more tweets. More data will also give better results for the word2vec model.