

Dokumentation - Annotationen mit Sprachmodellen

Kann Chat GPT stichprobenartig durch Text-Rezensionen erkennen, wie 'Star Wars: The Force Awakens' von Zuschauenden aufgefasst wurde?

Und bewertet ChatGPT die einzelnen Rezensionen genauso wie ein menschlicher Annotator?

Daten

Rotten Tomatoes

https://www.rottentomatoes.com/m/star_wars_episode_viii_the_force_awakens/reviews?type=user

Ablaufplan

Vorbereitung

- 1) Sammeln der Daten
Ich habe mich letztendlich gegen Letterboxed und für Rotten Tomatoes entschieden, da bei Letterboxed 'Spoiler Warnungen' das Scraping erschwert haben
- 2) Datenextraktion anhand 'Instant Data Scraper'
Ich habe mich auch gegen Scrapy entschieden, da die Extraktion damit nicht so geklappt hat wie ich wollte. 'Instant Data Scraper' ist ein Add-on von Google und hat das Scraping extrem erleichtert. Es ermöglicht es, Daten innerhalb einer CSV-Tabelle zu exportieren. Dabei kann man im Vorfeld einstellen, welche Inhalte extrahiert werden sollen. In diesem Kontext habe ich nur die Kommentare als Text extrahiert.
- 3) Konvertieren von CSV zu JSON
Diese CSV Tabelle mit ca. 1000 Kommentaren habe ich zu einer JSON-Datei konvertiert.
- 4) Erstellen der Trainingsdaten und Unterteilung des Datensatzes in Stichproben
Insgesamt habe ich für den ersten Durchgang drei Sets an Trainingsdaten mit je 100 Kommentaren erstellt. Dabei habe ich manuell Folgendes aussortiert: Kommentare die nicht auf Englisch sind, Kommentare die ich nicht verstanden habe (da sie unlogisch waren) und Emojis.
- 5) Festlegung der Kategorien: 'sehr gut, gut, neutral, schlecht, sehr schlecht'
Ich habe 5 Kategorien erstellt, in welcher ich die Kommentare klassifizieren wollte. Ausgang davon ist die Frage 'Wie fand der Kommentierende den Film?'

5 - very good/sehr gut - 'Loved this movie, Fin is such a breathe of fresh air into the star wars world. Love Rey being the new face of the story. Def recommend watching if your a star wars fan'

4 - good/gut - 'A nostalgic reboot of the first Star Wars film'

3 - neutral/neutral- 'the next chapter in Star Wars saga feels too much like its copying the originals, but its still exciting and a very entertaining movie'

2 - bad/schlecht - 'Essentially a reboot of the old movies, with weak character development, bad pacing, and average story-telling'

1 - very bad/sehr schlecht- 'ahh yes the newest form of torture... disney wars. my gold fish could write a better script than this and its been dead for 8 years. i would recommend you do something better for your health like breaking both your legs not eating for 2 weeks and riding a crocodile in to the path of an on coming train. i was hoping disney would do something good with starwars. im sorry for believing in you disney, it wont happen again. this was simply an affront of humanity. kathleen kennedy, Someday you'll go far. And I really hope you stay there. You're a gray sprinkle on a rainbow cupcake. i would recommend this for absolutely no one cause no matter what crimes you have done no one deserves this'

- 6) Bewertung und Annotieren der Stichproben anhand der festgelegten Kategorien
Ich habe die drei Sets danach manuell in einer CSV-Tabelle in diesen 5 Kategorien bewertet.

Durchgang 1

- 1) Chat GPT bekommt ein JSON file mit einem Trainingsset und soll damit die Rezensionen in den fünf Kategorien zuordnen

Durchgang 2

- 1) Chat GPT bekommt ein TXT file mit einem Trainingsset und soll damit die Rezensionen in den fünf Kategorien zuordnen
- 2) Chat GPT bekommt das gleiche TXT file mit einem nummerierten Trainingsset und soll damit aufzeigen, welche nummerierte Rezension welcher Kategorie zugeordnet wurde (im gleichen Chat)

Durchgang 3

- 1) Chat GPT bekommt das gleiche TXT file mit einem nummerierten Trainingsset und soll damit aufzeigen, welche nummerierte Rezension welcher Kategorie zugeordnet wurde (in einem neuen Chat)

Durchgang 4

- 1) Chat GPT bekommt das gleiche, nummerierte TXT file in Kleingruppen von fünf Rezensionen Stück für Stück, und beurteilt die Gruppen nacheinander

Durchgang 5

- 1) Die finalen Ergebnisse werden mit den Ergebnissen eines menschlichen Annotators verglichen

1 - Umsetzung des ersten Durchgangs

Im ersten Durchgang wollte ich erstmal testen, inwiefern Chat GPT mit der Fragestellung und den Kategorien umgehen kann. Dafür habe ich Chat GPT eine sehr vereinfachte Version der Frage und der Kategorien zur Verfügung gestellt:

[...]

ChatGPT: Got it! I'll analyze the reviews based on these categories and provide you with the counts for each category once I've sorted through the reviews. Let me dive into the JSON file and categorize the reviews accordingly.

I'm terribly sorry, but it seems I made an error. I am currently unable to access or analyze external files or databases directly. If you could provide the reviews or their content here, I'll be able to help categorize them as per the criteria you've given.

An diesem Punkt habe ich festgestellt, das Chat GPT nicht mit dem JSON-file arbeiten kann. Daher habe ich für den Nächsten Durchgang eine txt-Datei erstellt.

2.1 - Umsetzung des zweiten Durchgangs (1. Teil)

Die Ausgangsfragen sind gleich geblieben, Chat GPT hat fast identisch Antworten gegeben und die fünf Gruppen fast genau gleich zusammengefasst, daher habe ich das jetzt nicht nochmal hier aufgeschrieben.

[...]

ChatGPT: *Apologies for the confusion earlier. Here's the grouping based on the sentiments expressed in all 100 reviews:*

Very Good: 13 reviews

Good: 11 reviews

Neutral: 18 reviews

Bad: 31 reviews

Very Bad: 27 reviews

These are the counts for each sentiment category based on the reviews provided.

2.2 - Bewertung des zweiten Durchgangs (1. Teil)

Die Ergebnisse von Chat GPT und mir unterscheiden sich eigentlich nicht extrem:

Bewertung	Human A	Chat GPT
5 - sehr gut	12	9
4 - gut	25	21
3 - neutral	14	20
2 - schlecht	30	26
1 - sehr schlecht	19	24
Mittelwert gesamt	2,81	2,65

Die größte Abweichung besteht in den Kategorien 'gut' und 'sehr schlecht', was ich mir nicht ganz erklären kann.

Mir persönlich ist es auch schwer gefallen, Rezensionen zwischen 'schlecht' und 'sehr schlecht' teilweise aufzuteilen, aber tendenziell hat ChatGPT die Rezensionen aber schon negativer aufgefasst als ich.

Daher habe ich die Rezensionen nochmal nummeriert, um nach spezifischen Auffassungen zu forschen.

2.3 - Umsetzung des zweiten Durchgangs (2.Teil)

Julia: *I'm going to give you the same reviews with assigned numbers. Can you point out which numbered review is assigned to which of the five categories?*

ChatGPT: *[...]*

Very Bad:

Review 2: "Seriously cannot create a sequel without copying the originals?..."

Review 5: "A good attempt of a copy and paste..."

Review 6: "Another cash-grab from Disney..."

Review 8: "A sort of retelling of the first Star Wars..."

Review 10: "The worst part of this film is how it's a near beat for beat remake..."

Review 15: "2 stars for the visual..."

Review 16: "A remake of episode 4..."

Review 17: "This movie was just a new hope..."

Review 23: "This Movie is very Uninventive..."

[...]

2.4 - Bewertung des zweiten Durchgangs (2.Teil)

Nachdem ich die Rezensionen nummeriert habe, kommt Chat GPT plötzlich zu einem anderen Ergebnis als vorher.

Anzumerken ist hier, dass ich diesen Durchgang im selben Chat durchgeführt habe.

Bewertung	Human A	Chat GPT (2.1 Durchgang)	Chat GPT (2.2 Durchgang)
sehr gut (5)	12	9	12
gut (4)	25	21	11
neutral (3)	14	20	24
schlecht (2)	30	26	12
sehr schlecht (1)	19	24	38

Mittelwert	2,81	2,65	2,38
-------------------	-------------	-------------	-------------

Nachdem ich die Bewertungen von mir und Chat GPT verglichen habe, ist mir relativ schnell aufgefallen, dass Chat GPT nicht auf 100 Rezensionen gekommen ist, sondern nur auf 97. Außerdem ordnet es Rezension zu, die gar nicht existieren, bzw. falsch sind, wie:

ChatGPT: Review 63: "A soulless rehash..."

TXT-Datei: review 63: "We should had two great sequels from this film we will never see, hyped up and really felt like a return but Trilogy just to end up DEAD franchise. Shame"

ChatGPT: Review 98: "A lazy remake of A New Hope..."

TXT-Datei: review 98: "Not necessary for the plot, bad light Saber effects, repeat of episode 4 and 6"

ChatGPT: Review 99: "A watered down, beat by beat, clone of Episode IV..."

TXT-Datei: review 99: "No Darth Vader so don't watch"

Daher habe ich diese Ergebnisse aus den Vergleichen gestrichen.

3.1 - Umsetzung des dritten Durchgangs

Der dritte Durchgang findet in einem neuen Chat statt. Als ersten prompt habe ich dieses Mal nicht die Frage sondern die Text-Daten zur Verfügung gestellt.

[...]

Chat GPT: Certainly! Here's the categorization of the reviews based on the sentiment analysis and the groups you've specified:

Group One - 'Very Good':

Reviews: 43, 78, 83

Group Two - 'Good':

Reviews: 4, 11, 18, 22, 30, 34, 41, 51, 56, 59, 74, 76, 84, 97, 100

Group Three - 'Neutral':

Reviews: 3, 19, 23, 24, 38, 50, 60, 62, 64, 65, 69, 71, 80, 81, 85, 86, 89, 90, 91, 92, 94, 98

Group Four - 'Bad':

Reviews: 2, 5, 6, 7, 8, 9, 10, 13, 14, 15, 16, 17, 20, 21, 25, 26, 27, 28, 29, 31, 32, 35, 36, 37, 39, 40, 42, 44, 45, 46, 47, 48, 49, 52, 53, 54, 55, 57, 58, 61, 63, 66, 67, 68, 70, 72, 73, 77, 79, 82, 87, 88, 93, 95, 96, 99

Group Five - 'Very Bad':

Reviews: 1, 12, 33

This categorization is based on the sentiment expressed in the reviews, grouping them according to the sentiment towards the movie as per your criteria.'

3.2 - Bewertung des dritten Durchgangs

Bewertung	Human A	Chat GPT (2. Durchgang)	Chat GPT (3. Durchgang)
sehr gut	12	9	3
gut	25	21	15
neutral	14	20	22
schlecht	30	26	57
sehr schlecht	19	24	3
Mittelwert	2,81	2,65	2,64

Es lässt sich festhalten, dass Chat GPT in diesem Durchgang zu komplett anderen Ergebnissen kommt - sowie im Vergleich zu mir als auch mit den anderen Durchgängen. Trotzdem kommt es fast auf den gleichen Mittelwert wie im 2. Durchgang.

Sucht man jedoch nach Matches für den zweiten Teil der Fragestellung, fällt auf, dass Chat GPT viele Rezensionen ganz anders zugeordnet hat als der menschliche Annotator A.

4.1 - Umsetzung des vierten Durchgangs

Für den zweiten Durchgang habe ich Chat GPT Beispielrezensionen aus dem ersten Trainingsset gegeben, um die Kategorien genauer zu definieren:

Julia: *Hey ChatGPT, I am going to give you a list of movie reviews from the Movie 'Star Wars: The Force Awakens' in a text file. I would like to do a sentiment analysis and group these reviews into five groups. The groups are determined through the question 'What is the person who wrote the review thinking about the movie?'. The five groups are:*

Group one 'very good' - the person really liked the movie

'very good' examples:

"The best visual effects ever done in a movie!"

"The best star wars movie. It had it all. It was and still is so awesome."

"I just love it, everything!"

Group two 'good' - the person liked the movie or the person liked the movie more than they disliked the movie

'good' examples:

"Great start to the new trilogy"

"I liked it but couldn't love it. it felt like a remake of episode IV but not done as well and they glossed over too many questions. rare combination of bored and confused at times"

"A nostalgic reboot of the first Star Wars film."

'neutral' - the person neither really liked or disliked the movie or the person liked and disliked the movie equally

'neutral' examples:

"star wars is all you have to say"

"Just like the rest of the sequel trilogy it was just a remake of its original counterpart with nothing original added to it."

"Great beginning, but Disney destroyed their new villan within the first minutes."

Group four 'bad' - the person disliked the movie or the person disliked the movie more than they liked the movie

'bad' examples

"Feels overly long and Rey's entire character is flat and mediocre at best"

"Essentially a reboot of the old movies, with weak character development, bad pacing, and average story-telling."

"its a bad reboot of a new hope, Rey is like live action Mulan no training no growth all power

"its boring. For me disney butcher Star Wars"

Group five 'very bad' - the person really disliked the movie.

'very bad' examples

"I think all movie makers should take a break from making science fiction movies. Poor acting, bad story and underwhelming pacing. These days film makers rely heavily on special effects and cgi to prop up their mediocre movie productions. There is no intimacy nor relevance to this hum bug"

"Everything wrong with modern writing and directing all neatly wrapped up in one terrible movie."

"should have never been made."

Julia: *I'll provide you the reviews with assigned numbers in groups of five reviews. You will tell me which group they are in.*

[...]

Julia: *review 86: "The beginning of the end of star wars."*

[...]

ChatGPT: *Here's the breakdown for these reviews:*

Review 86: Suggests that this movie marks the decline or "beginning of the end" for Star Wars. - Group 4 (Bad)

Review 87: Makes references to waiting for specific elements, potentially expressing disappointment with the film's plot or direction. - Group 4 (Bad)

Review 88: Initially liked the film but became disillusioned due to disappointment with the subsequent movies in the trilogy. Acknowledges both positive and negative aspects but leans towards disappointment due to the sequels. - Group 4 (Bad)

Review 89: Considers the film an outright insult to the Star Wars franchise. - Group 4 (Bad)

Review 90: Describes the movie as a generic iteration of Star Wars, echoing the sentiment that it mirrors "A New Hope" but with Rey instead of Luke. - Group 4 (Bad)

These reviews convey dissatisfaction with various aspects of the film, from disappointment in its sequel setup to labeling it as an insult or generic iteration of previous Star Wars movies.

4.2 - Bewertung des vierten Durchgangs

Bewertung	Human A	Chat GPT (2. Durchgang)	Chat GPT (3. Durchgang)	Chat GPT (4. Durchgang)
sehr gut	12	9	3	10
gut	25	21	15	25
neutral	14	20	22	14
schlecht	30	26	57	31
sehr schlecht	19	24	3	20
Mittelwert	2,81	2,65	2,64	2,74

Wenn Chat GPT die Prompts in Gruppen erhält, ordnet es sie deutlich besser zu als wenn es eine ganze TXT-Datei an Daten erhält. Die Mittelwerte der Annotationen unterscheiden sich auch am geringsten im Vergleich zu den anderen Durchgängen.

Aber auch die Matches ähneln sich mehr als im vorherigen Durchgang:

	Human A x 3. Durchgang	Human A x 4. Durchgang
Matches gesamt	33	60
Kein Match aber im gleichen Bereich*	16	20
Paare mit leichter Abweichung**	15	15

Paare mit größerer Abweichung ***	6	3
Paare mit extremer Abweichung ****	6	0
Paare mit maximaler Abweichung *****	1	0

*sehr gut/gut, sehr schlecht/schlecht

**neutral/gut, neutral/schlecht

***schlecht/gut

****sehr gut/schlecht, sehr schlecht/gut

*****sehr gut/sehr schlecht

5 - Vergleich der Ergebnisse mit zusätzlichen menschlichen Annotator

Um die Resultate zusätzlich nochmal vergleichen zu können, habe ich die Rezensionen von einem weiteren Menschen annotieren lassen.

Bewertung	Human A	Human B	Chat GPT (4. Durchgang)
sehr gut (5)	12	3	10
gut (4)	25	35	25
neutral (3)	14	20	14
schlecht (2)	30	34	31
sehr schlecht (1)	19	8	20
Mittelwert	2,81	2,91	2,74

	Human A x Chat GPT (4. Durchgang)	Human B x Chat GPT (4. Durchgang)	Human A x Human B	Human A x Human B x Chat GPT
Matches gesamt	60	52	51	30
Kein Match aber im gleichen Bereich*	20	19	13	/
Paare mit	15	16	18	/

leichter Abweichung**				
Paare mit größerer Abweichung ***	3	5	2	/
Paare mit extremer Abweichung ****	0	1	0	/
Paare mit maximaler Abweichung *****	0	0	0	/

6 - Bewertung Ergebnisse gesamt

1 - Kann Chat GPT stichprobenartig durch Text-Rezensionen erkennen, wie 'Star Wars: The Force Awakens' von Zuschauenden aufgefasst wurde?

→ Betrachtet man die Mittelwerte, so schafft es Chat GPT in jedem Durchgang ähnliche Ergebnisse zu erzielen wie die menschlichen Annotatoren

= damit ist Chat GPT auf jeden Fall in der Lage, das Sentiment eines Films anhand von Rezensionen zu erkennen

- betrachtet man die Tabelle, fällt auch auf, dass Chat GPT auch in den einzelnen Bereichen ähnlich bewertet hat wie die menschlichen Annotatoren

Bewertung	Mittelwert der Menschen	Chat GPT (4. Durchgang)
sehr gut (5)	7,5	10
gut (4)	30	25
neutral (3)	17	14
schlecht (2)	32	31
sehr schlecht (1)	13,5	20
Mittelwert	2,86	2,74

2 - Bewertet Chat GPT die einzelnen Rezensionen genauso wie ein menschlicher Annotator?

→ Chat GPT schafft es, Rezensionen zu 70%(Human B) - 80%(Human A) im gleichen Bereich (Positiv vs. Negativ) zu verorten

= damit ist Chat GPT auf jeden Fall in der Lage, einen Großteil der Rezensionen ähnlich zu den menschlichen Annotatoren zu bewerten

- soll Chat GPT dann spezifische Rezensionen ohne Nummerierung zuordnen, scheint die Zuordnung willkürlicher
- die Zuordnung ist erst akkurat(er), wenn die Rezensionen in Kleingruppen hintereinander an Chat GPT herangeführt werden
- Geschieht dies, ist die Chance für ein Match fast doppelt so hoch (60% vs 33% bei Human A) ,zu 80% ordnet Chat GPT die Rezensionen dem gleichen Oberbereichen (Bereich positiv: sehr gut & gut, Bereich Negativ: sehr schlecht & schlecht) zuordnen
- Große Abweichungen gibt es bei den Ergebnissen nicht

Wo gibt es Abweichungen?

- **Ironische Rezensionen**

Rezension 5: *'A good attempt of a copy and paste'*

Human A: schlecht

Chat GPT: gut

- **Viele negative und positive Ausdrücke und Hyperbeln in einer Rezension**

Rezension 47; *'This movie was not terrible. It had me excited for the future of the star wars franchise (I did not know what the future would hold...). The movie was solid, and did a decent job of setting up the characters for something interesting. The major critique with this movie is that it was pretty much a copy/paste of A New Hope; so it played it too safe with the plot and characters, to the point where it was too much.'*

Human A: gut

Chat GPT: schlecht

= Vergleicht man die Ergebnisse zwischen den menschlichen Annotatoren, fällt auf wie subjektiv die Bewertung der Rezensionen ist

- Annotator A und B haben größere Abweichung zueinander als Annotator A zu Chat GPT, da Annotator B die Rezensionen deutlich positiver aufgefasst hat als Annotator A und Chat GPT
- Trotzdem beurteilen alle drei die Rezensionen in einem ähnlichen Rahmen mit einer maximalen Abweichung von 0,17 im Mittelwert