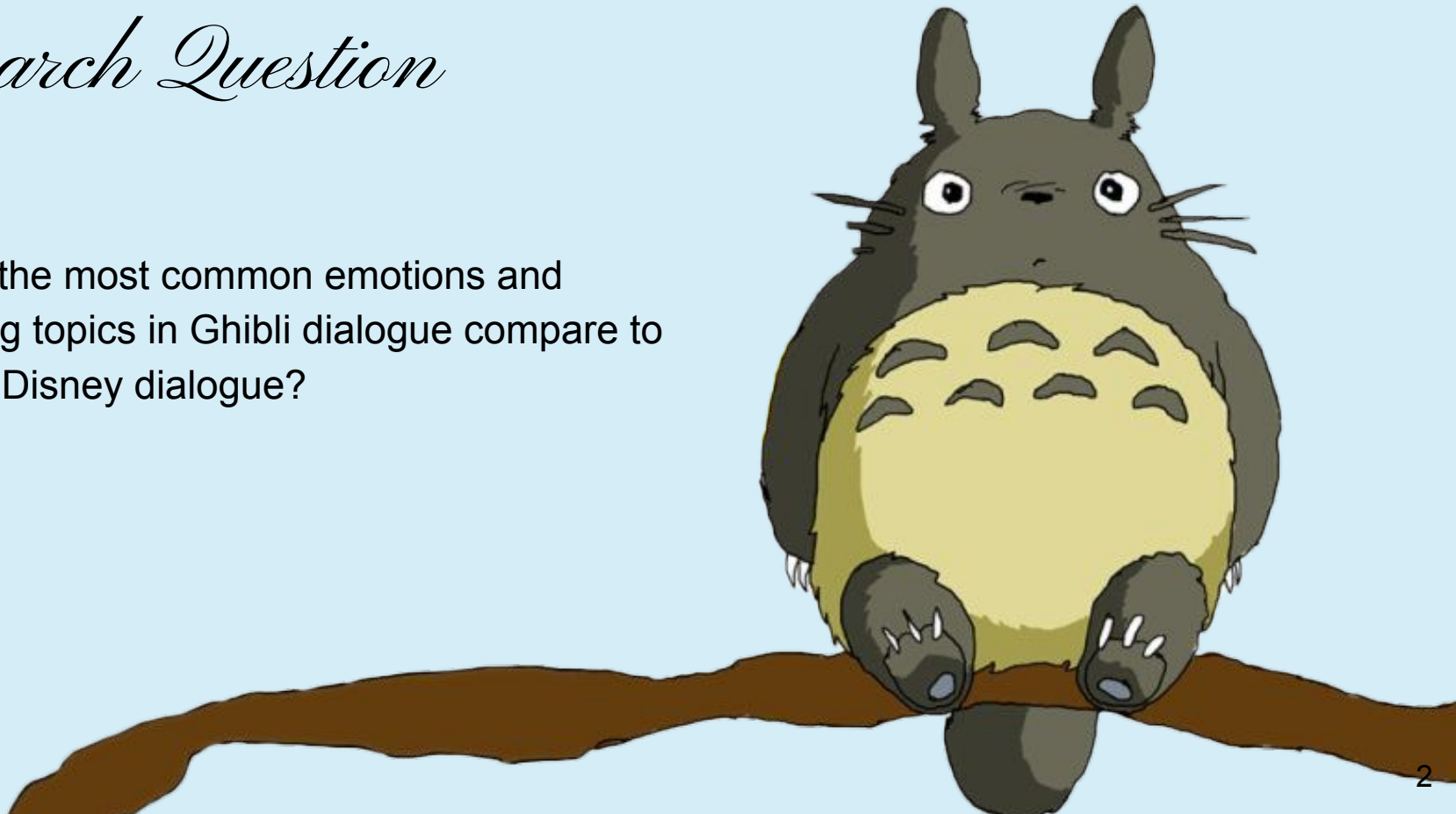
A Comparative Analysis Using Text Mining Techniques

Francesca Albio, Violeta Duró & Ana Sistac

Research Question

How do the most common emotions and prevailing topics in Ghibli dialogue compare to those in Disney dialogue?



Motivatio

DATA

data2 > disney > Moana.srt

```
1
00:00:54,304 --> 00:00:55,848
In the beginning...

2
00:00:56,014 --> 00:00:58,851
there was only ocean...

3
00:00:59,017 --> 00:01:02,980
until the mother island emerged.

4
00:01:03,146 --> 00:01:04,982
Te Fiti.

5
00:01:05,607 --> 00:01:09,361
Her heart held the greatest power
ever known.
```

Subtitles

- 17 movies from Disney and 17 movies from Studio Ghibli
- Datasource: Just Subtitles

https://www.justsubtitles.com/

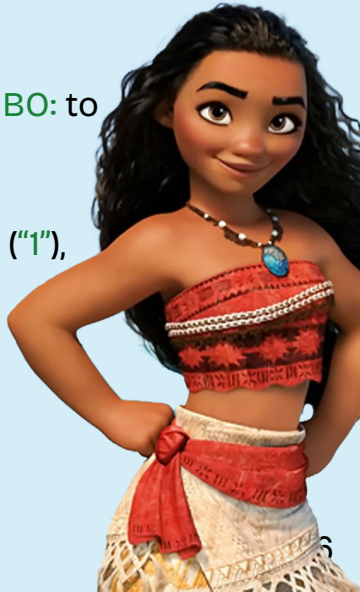
- Why subtitles?

data2 > disney > Moana.srt

```
1
00:00:54,304 --> 00:00:55,848
In the beginning...

2
00:00:56,014 --> 00:00:58,851
there was only ocean...

3
00:00:59,017 --> 00:01:02,980
until the mother island emerged.

4
00:01:03,146 --> 00:01:04,982
Te Fiti.

5
00:01:05,607 --> 00:01:09,361
Her heart held the greatest power
ever known.
```

Preprocessing steps

- Leading dashes: dialogue
- HTML tags: strip ...
- Non-speech cues: all-caps like (LAUGHING) or [MUSIC]
- Speaker labels: deletes e.g. DUMBO: to avoid them in word counts
- Skip logic: filter out pure indices ("1"), timestamps, and blank lines
- Invisible chars



data2 > disney > Moana.srt

```
1
00:00:54,304 --> 00:00:55,848
In the beginning...

2
00:00:56,014 --> 00:00:58,851
there was only ocean...

3
00:00:59,017 --> 00:01:02,980
until the mother island emerged.

4
00:01:03,146 --> 00:01:04,982
Te Fiti.

5
00:01:05,607 --> 00:01:09,361
Her heart held the greatest power
ever known.
```

data_cleaned > disney > Moana_cleaned.txt

```
In the beginning...
there was only ocean...
until the mother island emerged.
Te Fiti.
Her heart held the greatest power
ever known.
It could create life itself.
And Te Fiti shared it with the world.
But in time...
some began to seek Te Fiti's heart.
They believed
if they could possess it...
the great power of creation
would be theirs.
And one day...
the most daring of them all...
voyaged across the vast ocean
to take it.
He was a demigod of the wind and sea.
He was a warrior.
A trickster.
```

Sentences

SENTIMENT ANALYSIS

data_in_sentences > disney > Wreck-It.Ralph_cleaned_sentences.txt

```
1   My name's Ralph, and I'm a bad guy.
2   Let's see.
3   I'm nine feet tall.
4   I weigh 643 pounds.
5   Got a little bit of a temper on me.
6   My passion bubbles very near the surface, I guess, not gonna lie.
7   Anyhoo, what else?
8   I'm a wrecker.
9   I wreck things.
10  Professionally.
11  I'm very good at what I do.
12  Probably the best I know.
13  The thing is, fixing is the name of the game.
14  Literally, Fix-It Felix, Jr.
15  So, yeah, naturally, the guy with the name Fix-It Felix is the good guy.
16  He's nice enough as good guys go.
17  Definitely fixes stuff really well.
18  But if you've got a magic hammer from your father, how hard can it be?
```

- (?<=[.!?])\s+ → period, exclamation, question mark followed by space = split
- Keeps punctuation
- Result: splits into sentences → writes one sentence per line

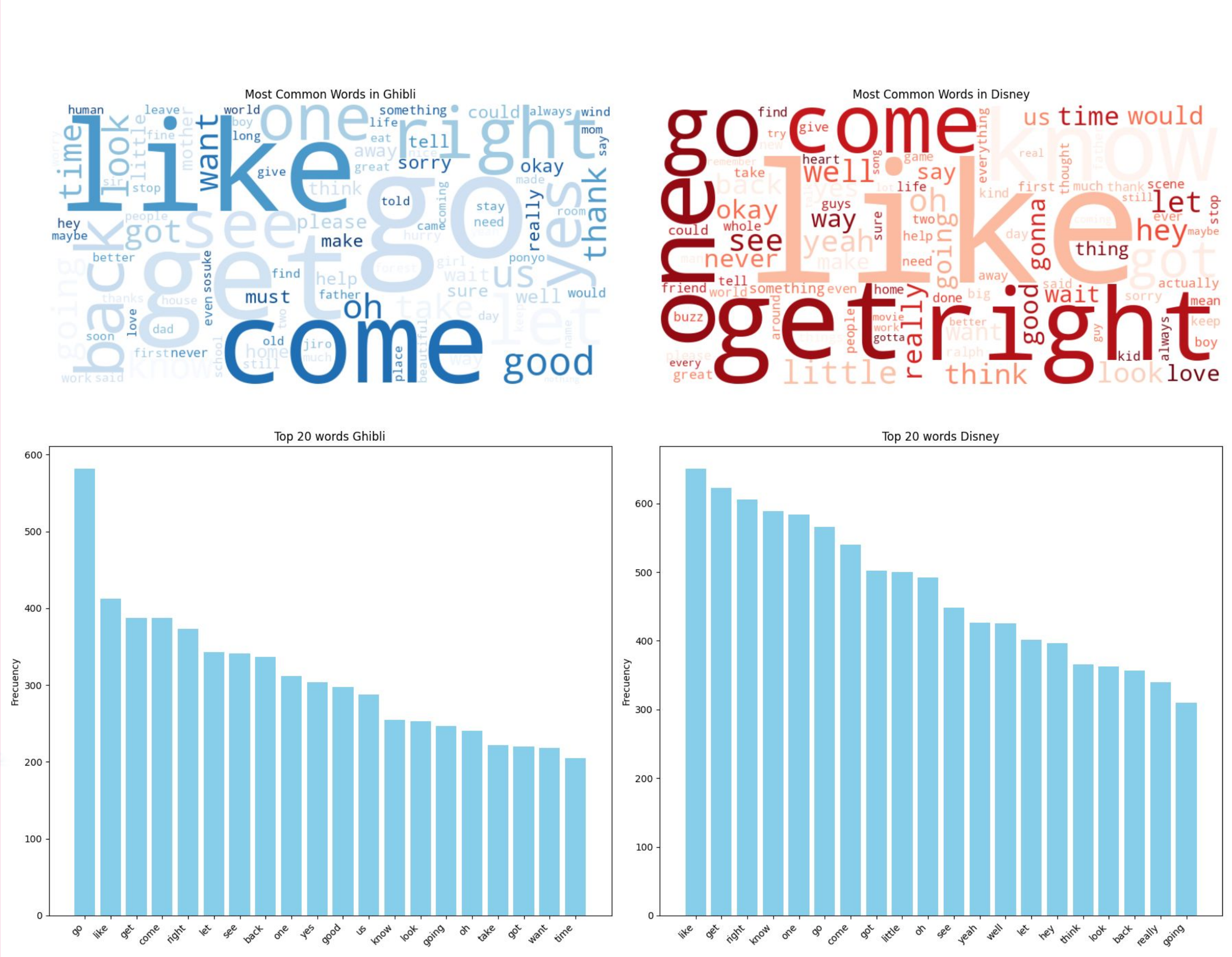
Tokenization

TOPIC MODELING

data_preprocessed > disney > Moana_cleaned_tokens.txt

```
1   beginning ocean mother island emerged te fiti heart held greatest power ever known could create life
    te fiti shared world time began seek te fiti heart believed could possess great power creation would
    one day daring voyaged across vast ocean take demigod wind sea warrior trickster shapeshifter could
    change form power magical fish hook name maui without heart te fiti began crumble giving birth
    terrible darkness maui tried escape confronted another sought heart te kā demon earth fire maui
    struck sky never seen magical fish hook heart te fiti lost sea even 1 000 years later te kā demons
    deep still hunt heart hiding darkness continue spread chasing away fish draining life island island
    every one us devoured bloodthirsty jaws inescapable death one day heart found someone journey beyond
    reef find maui deliver across great ocean restore te fiti heart save us thank mother enough papa one
    goes outside reef safe darkness monsters monsters monsters monsters darkness nothing beyond reef
    storms rough seas gonna throw long stay safe island fine legends true someone go mother motunui
    paradise would want go anywhere else shoo shoo moana moana scared wanna go back know know go
    dangerous moana come let go back village next great chief people wondrous things little minnow oh yes
    first must learn meant moana make way make way moana time knew village motunui need dancers
    practicing dance ancient song needs new song old one need tradit
    need share everything make joke weave baskets fishermen come bac
    ground people need chief comes day gonna look around realize hap
    use part coconut need make nets fibers water sweet inside use le
    consider coconuts trunks leaves island gives us need one leaves
    future okay time learn must find happiness right like dance wate
    like misbehaves village may think crazy say drift far know like well father daughter stubbornness
    pride mind says remember may hear voice inside voice starts whisper follow farthest star moana voice
```

- Remove punctuation
- Lowercase
- Stop words


Most Common Words in Ghibli
Most Common Words in Disney
Top 20 words Ghibli
Frecuency
0
100
200
300
400
500
600
go like get come right let see back one yes good us know look going oh take got want time
Top 20 words Disney
Frecuency
0
100
200
300
400
500
600
like get right know one go come got little oh see yeah well let hey think look back really going

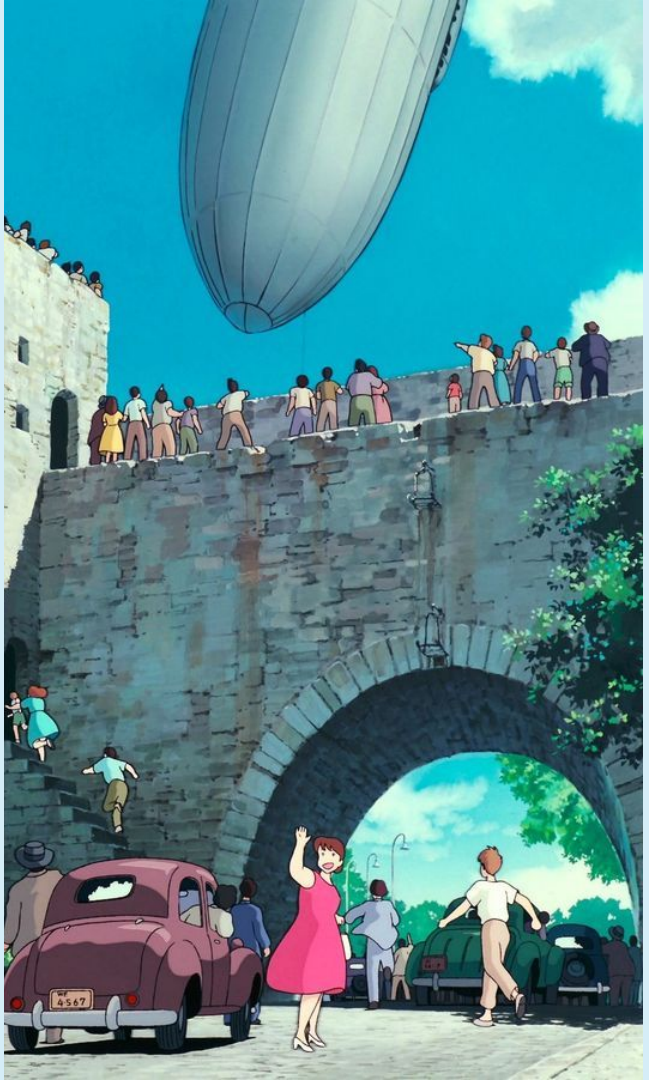

MODELLIN
G

Topic Modelling

We used **Latent Dirichlet Allocation (LDA)** and used to **Coherences Scores** to evaluate the quality of the model

Preprocessing:

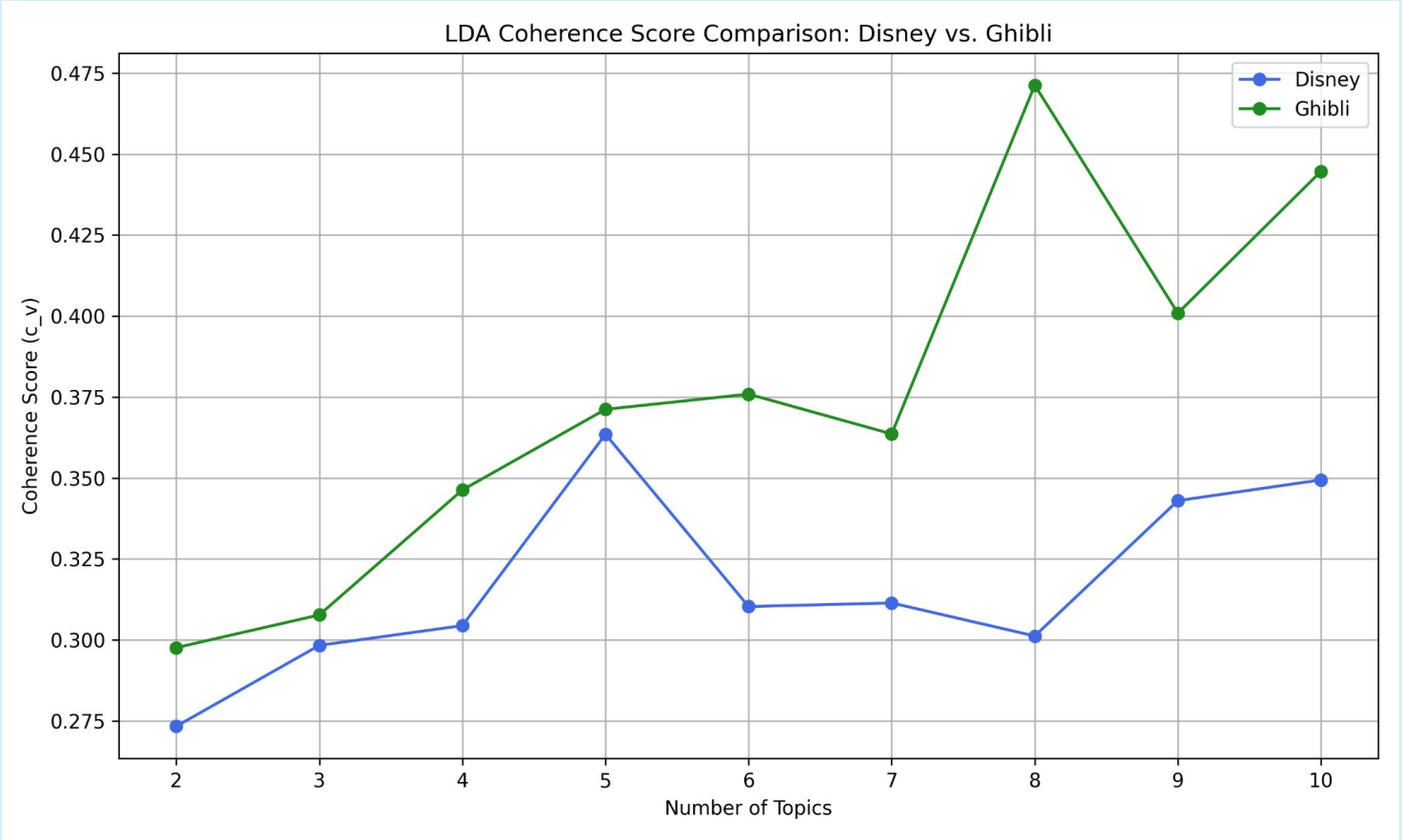
- Divided each movie script (17 per genre) in **6 equal part**
  → total document number: **170 for each genre**
- **Part-of-speech tagging** using NLTK's Universal Tagset
- **Lemmatized words** using WordNet Lemmatizer
- Kept only **nouns** to focus on meaningful content words



Topic Modelling

- Filtered out **stopwords**: Standard English stopwords (e.g., "the", "is"), Custom stopwords (e.g., character names, studio-specific terms)
- Removed **short tokens** (less than 3 characters)
- Adjusted **hyperparameter**: chunksize and passes
- Filtered tokens by frequency: Removed words that appear in fewer than 3 documents & Removed words that appear in more than 80% of documents

Coherence Scores



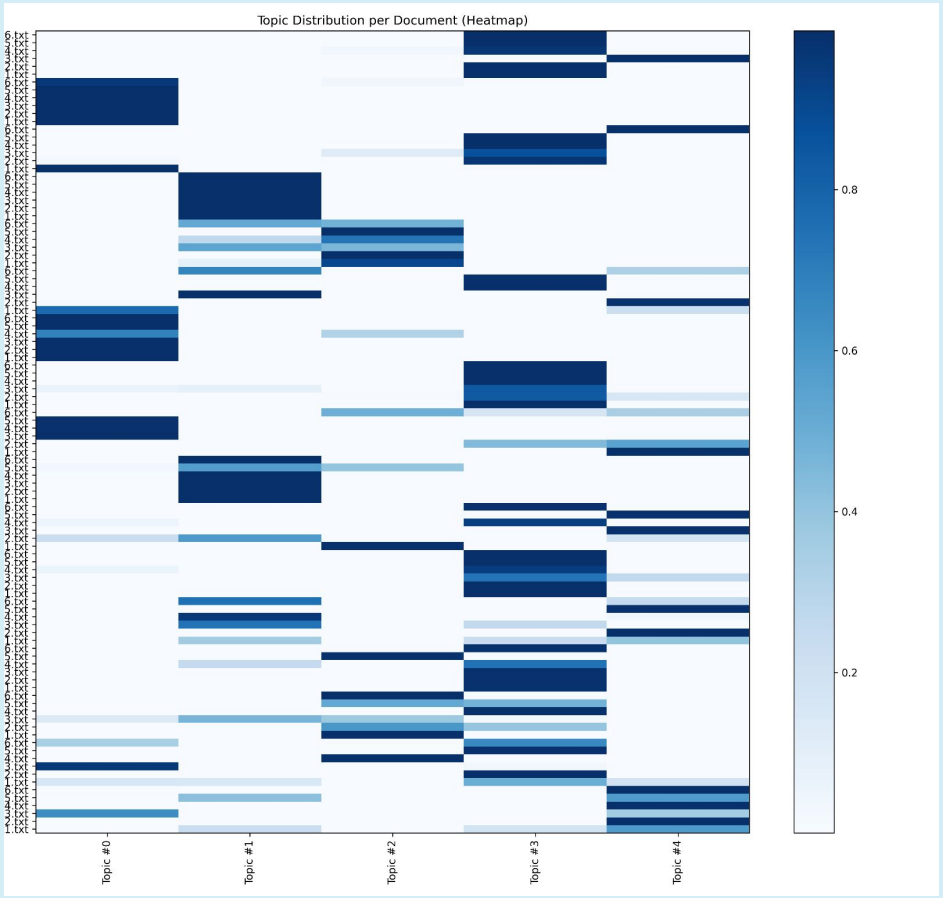
Disney:

1) **Digital Adventure:** game, kid, friend, man, race, home, heart, life
2) **The wild world:** heart, night, savage, predator, world, son, night, rock
3) **Dreamed life:** man, world, dream, life, heart, talk, home, gold
4) **Ocean:** heart, dream, people, island, hook, sea, boat, love
5) **Unclear Topic:** hero, planet, baby, year, god, place, guest, prince

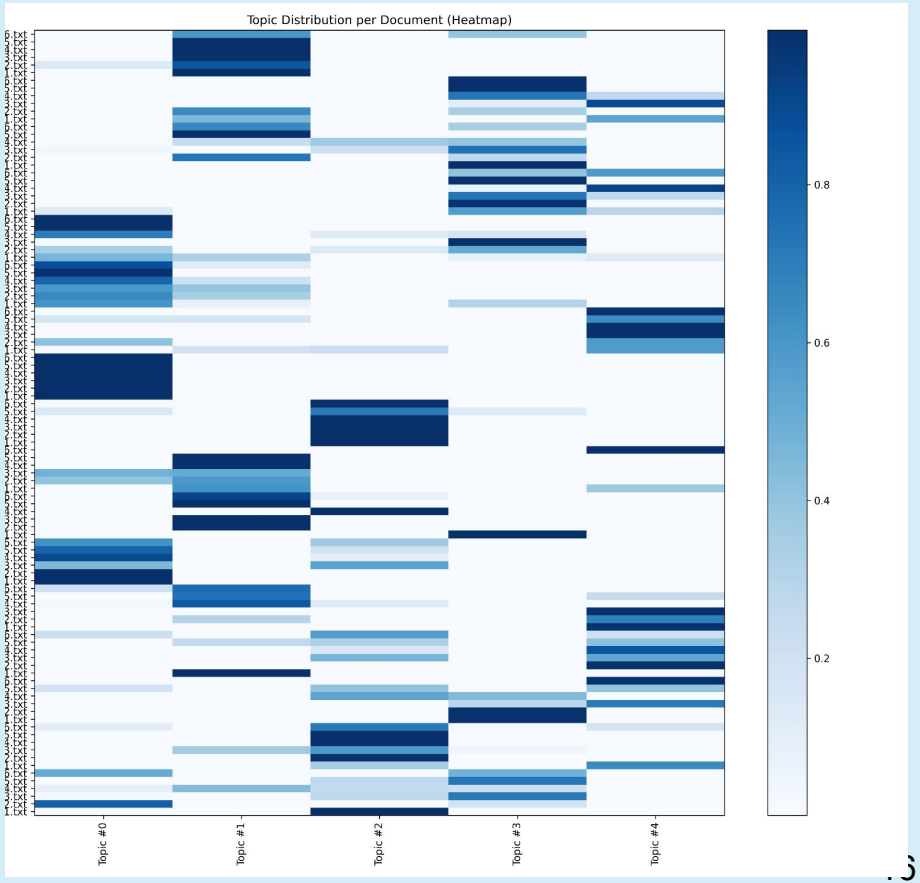
Ghibli:

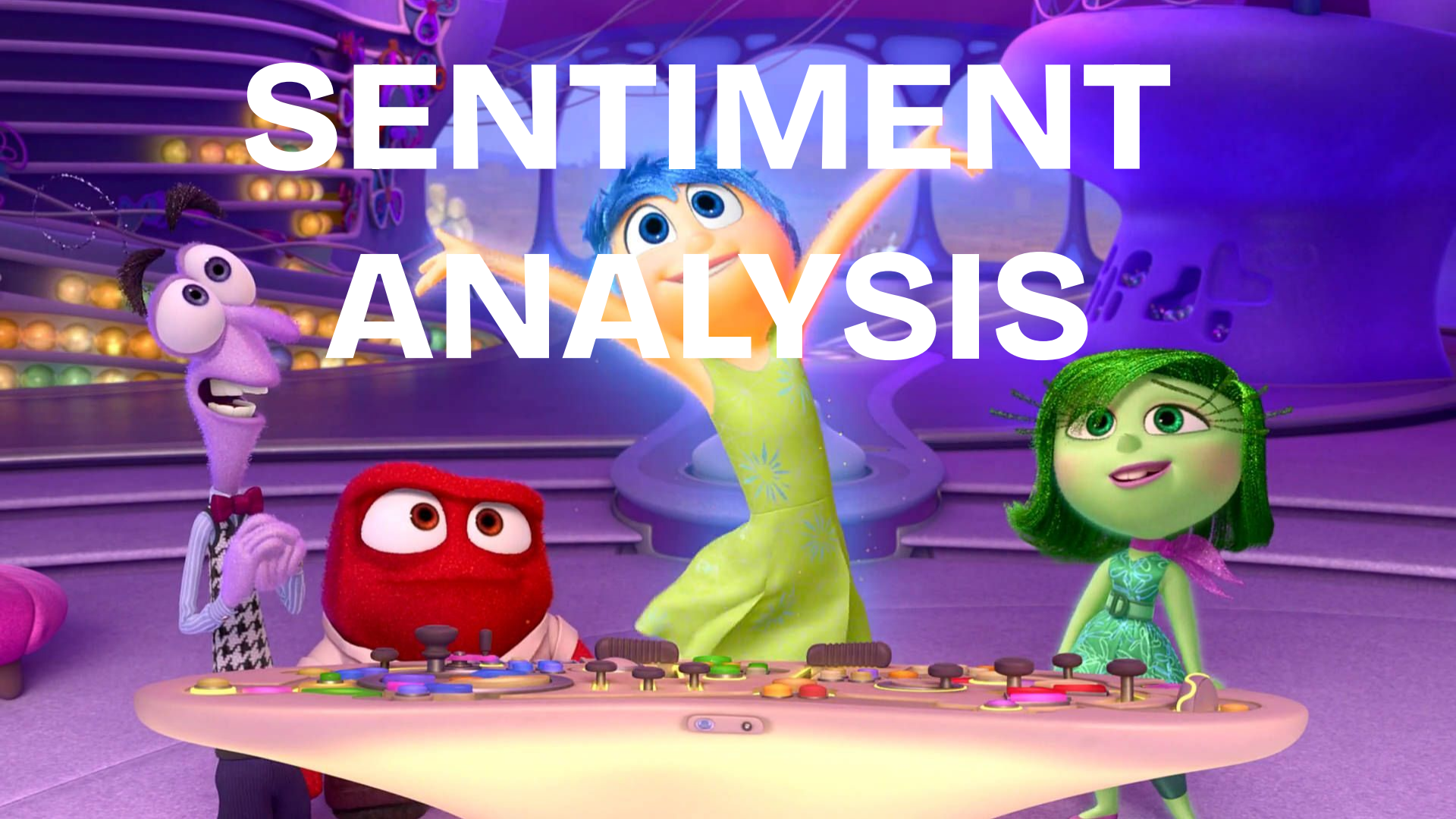
1) **Feminine Power and Nature:** life, spirit, world, lord, woman, sea, human, princess
2) **City life:** home, school, road, hand, people, night, work, town
3) **In the air:** plane, engine, air, quarter, club, hand, work, fly
4) **Family**: home, mother, year, mama, house, family, dream, country
5) **Magic:** wizard, work, eat, witch, morning, dad, home, water

Disney



Studio Ghibli




SENTIMENT
ANALYSIS

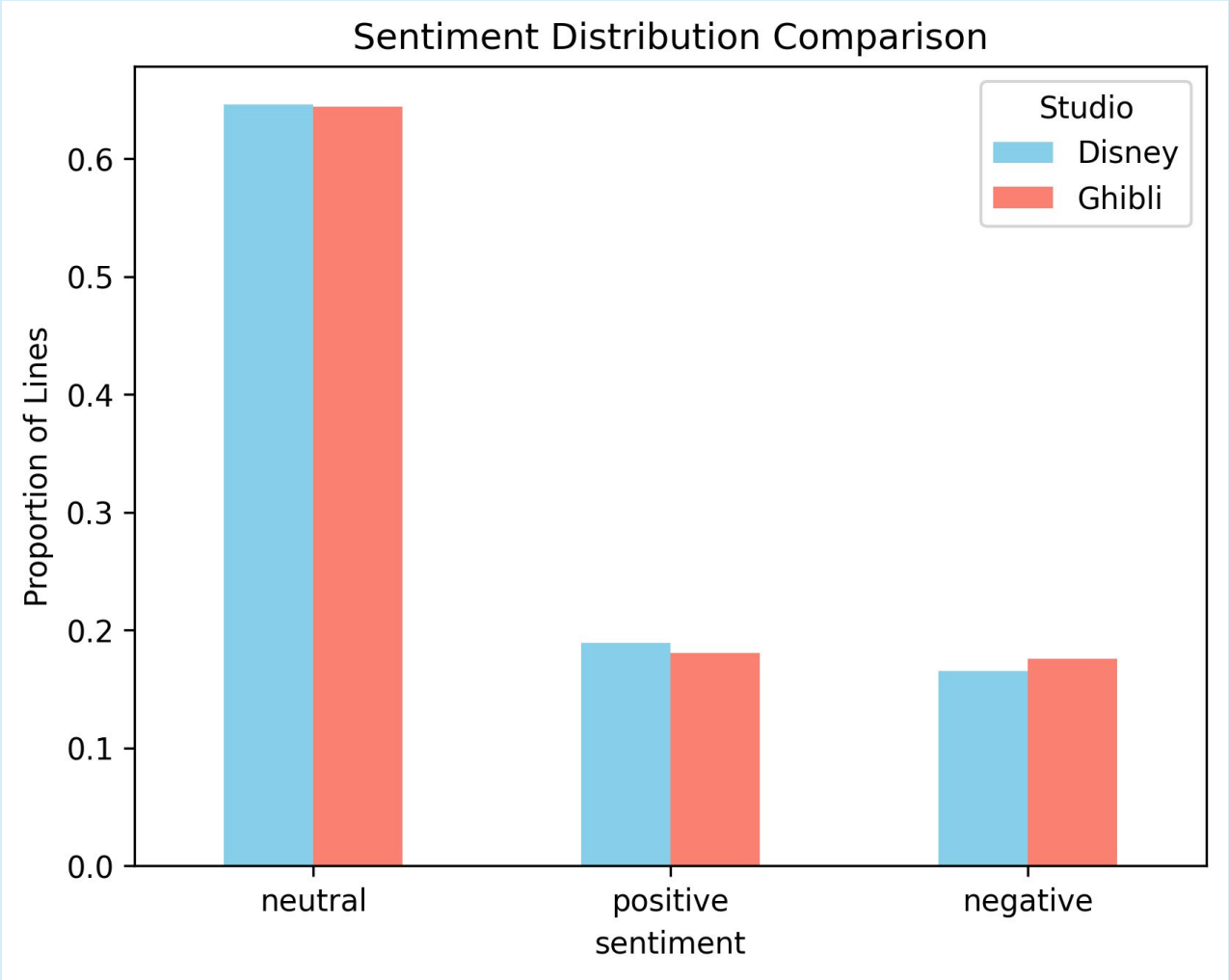
Sentiment Analysis Methods 1

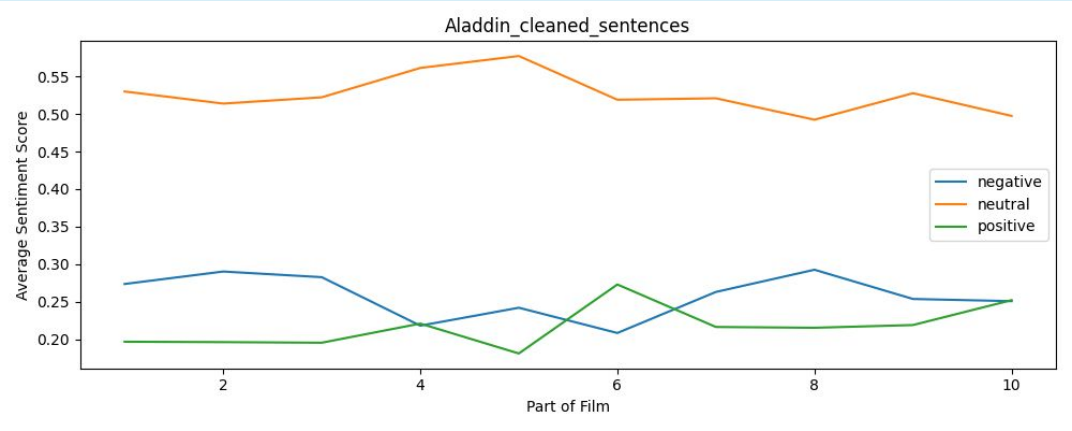
cardiffnlp/twitter-roberta-base-sentiment: fine tuned on millions of labeled tweets

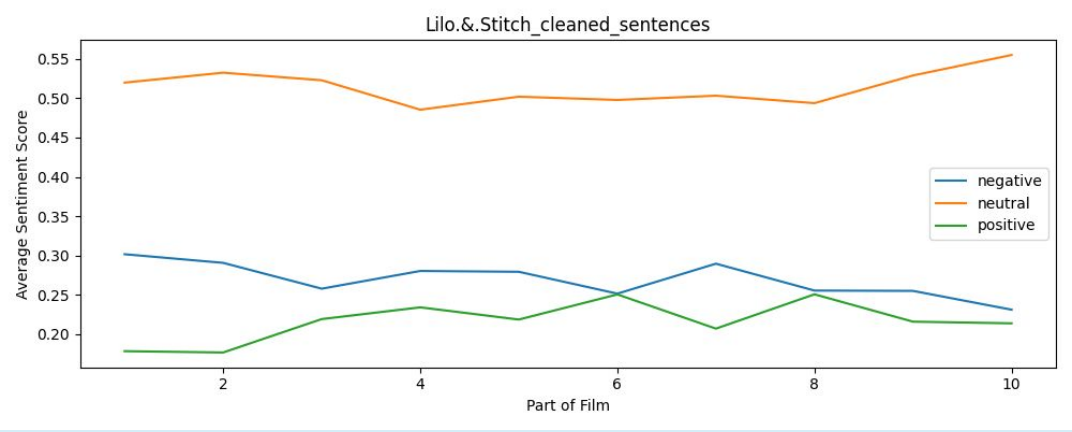
- Tweets are short, informal, and often punctuated like subtitles

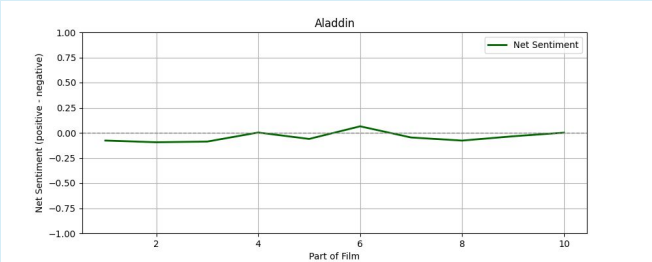
sarahai/movie-sentiment-analysis: positive, neutral, negative, uncertain

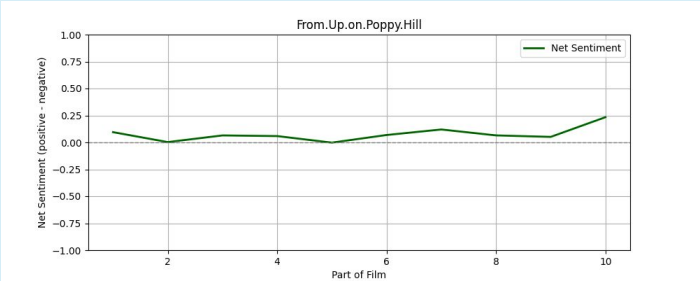
tabularisai/multilingual-sentiment-analysis: very negative, negative, neutral, positive, very positive

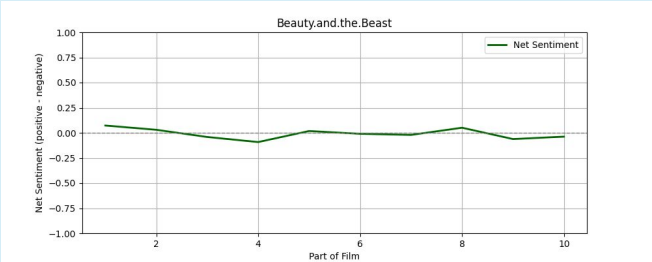

Sentiment Distribution Comparison
Studio
Disney
Ghibli
Proportion of Lines
0.0
0.1
0.2
0.3
0.4
0.5
0.6
neutral
positive
negative
sentiment

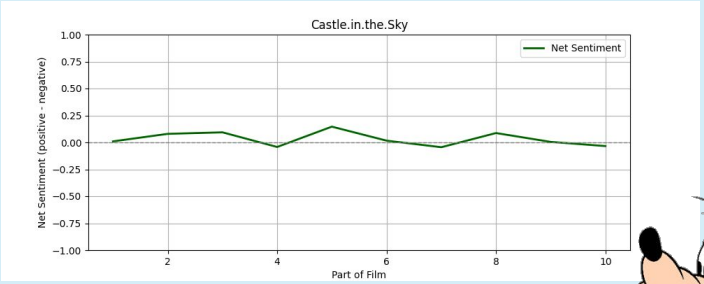

Aladdin_cleaned_sentences
Average Sentiment Score
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
2
4
6
8
10
Part of Film
negative
neutral
positive

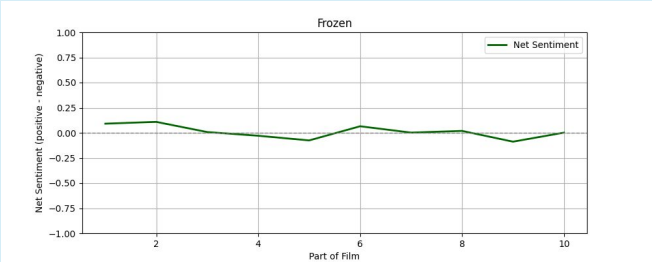

Lilo.&.Stitch_cleaned_sentences
Average Sentiment Score
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
2
4
6
8
10
Part of Film
negative
neutral
positive

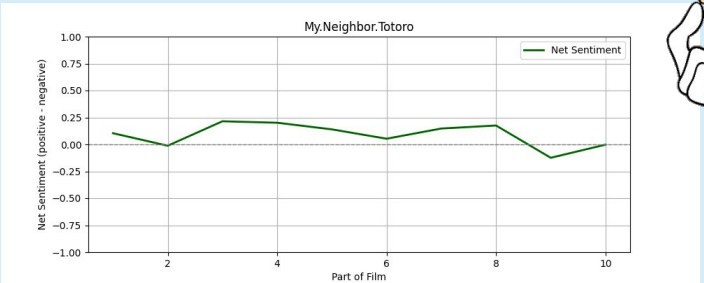

Aladdin
Net Sentiment
Net Sentiment (positive - negative)
Part of Film

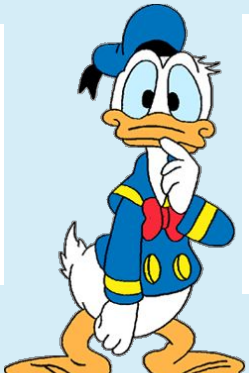

From.Up.on.Poppy.Hill
Net Sentiment
Net Sentiment (positive - negative)
Part of Film


Beauty.and.the.Beast
Net Sentiment
Net Sentiment (positive - negative)
Part of Film


Castle.in.the.Sky
Net Sentiment
Net Sentiment (positive - negative)
Part of Film


Frozen
Net Sentiment
Net Sentiment (positive - negative)
Part of Film


My.Neighbor.Totoro
Net Sentiment
Net Sentiment (positive - negative)
Part of Film



Sentiment Analysis Method 2

Theme	Ghibli	Disney
Resilience	0.449	0.453
Grief	0.733	0.500
Trauma	0.518	0.510
Personal Growth	0.562	0.528
Empathy	0.585	0.505
Romantic Love	0.612	0.497
Friendship	0.850	0.668
Family	0.758	0.691
Loneliness	0.792	0.578
Identity	0.501	0.571

Sentiment Model:
`tabularisai/multilingual-sentiment-analysis`
https://huggingface.co/tabularisai/multilingual-sentiment-analysis

- Classifies text into 5 categories: very negative, negative, neutral, positive, very positive
- Pre-trained multilingual transformer-based model

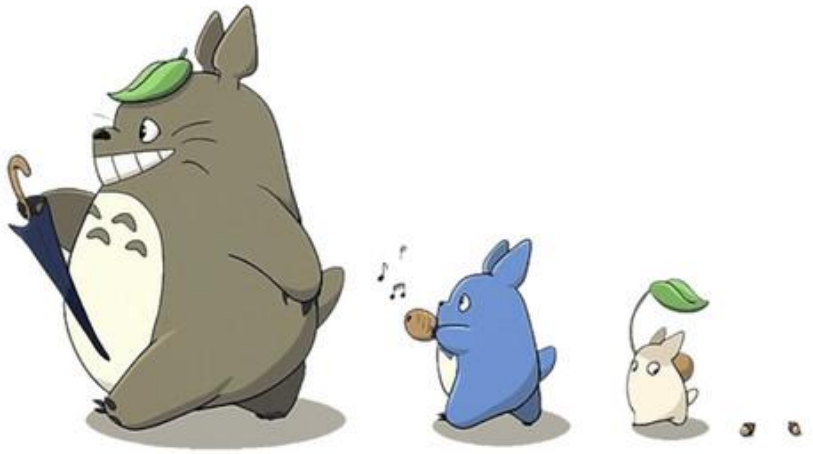
Theme Classifier:
`facebook/bart-large-mnli`

- Zero-shot classification to identify narrative themes
- No specific training required for the defined themes

Limitations

- Limited number of documents
- Difficulty with the models, since they are not trained for movie scripts
- Maybe future labeled data?

Conclusions




WALT DISNEY'S
MICKEY MOUSE
THE END

References

Chao, B., & Sirmorya, A. (2016). Automated Movie Genre Classification with LDA-based Topic Modeling. *International Journal of Computer Applications*, *145*(13), 1–5. https://doi.org/10.5120/ijca2016910822

Lee, S.-H., Yu, H.-Y., & Cheong, Y.-G. (2017). Analyzing Movie Scripts as Unstructured Text. *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*. https://doi.org/10.1109/bigdataservice.2017.43

nfasano. (2023). GitHub - nfasano/movie_recsys: Enhanced Movie Recommendations with Collaborative Topic Modeling. Retrieved May 23, 2025, from GitHub website: https://github.com/nfasano/movie_recsys