DHBW Ravensburg

Burp NLP… An multidimensional Text analysis

Intelligent Text Analysis

18.03.2025

Anton Geiger, Cedric Fechner

Data Understanding

Datasources - transcript

episode no.	speaker	dialouge
1	Rick	stumbles in drunkenly, and turns on the lights. … Morty
1	Morty	rubs his eyes. What, Rick? What's going on?

~ 9000 datapoints

until season 5

Data Understanding

Datasources - episode description

ID	Episode name	text
0	Pilot	middle night obviously drunk rick bursts morty ’s room tells …
1	Lawnmower Dog	jerry complains family dog snuffles stupid….

~ 80 datapoints

until season 7

Data Understanding

Datasources - IMDB rating

ID	Season	Episode	Episode name	rating
o	S1	E1	Pilot	7.9
1	S1	E2	Lawnmower Dog	8.6

~ 80 datapoints

until season 7

Data Prep

Text preparation spacy pipeline

1. Remove HTML
2. set text to lowercase
3. Remove Abbreviations
4. Spacy Pipeline
   a. Remove Lemmatizer
   b. Add Custom Porterstemmer
   c. Add Stopword remover

stumbles in drunkenly, and turns on the lights. Morty! You gotta come on. Jus'... you gotta come with me.

stumbl drunkenli turn light morty got tocom ju got come

Data Exploration

Wordcloud - Top 100 words (transcript)

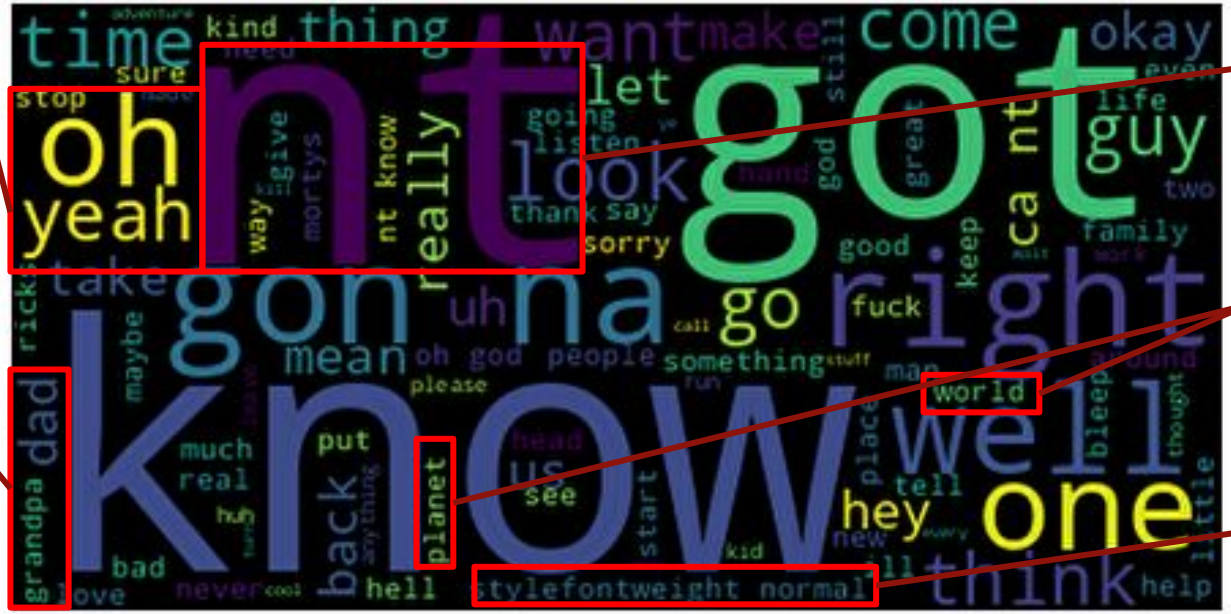
colloquial language

abbreviations

solar system

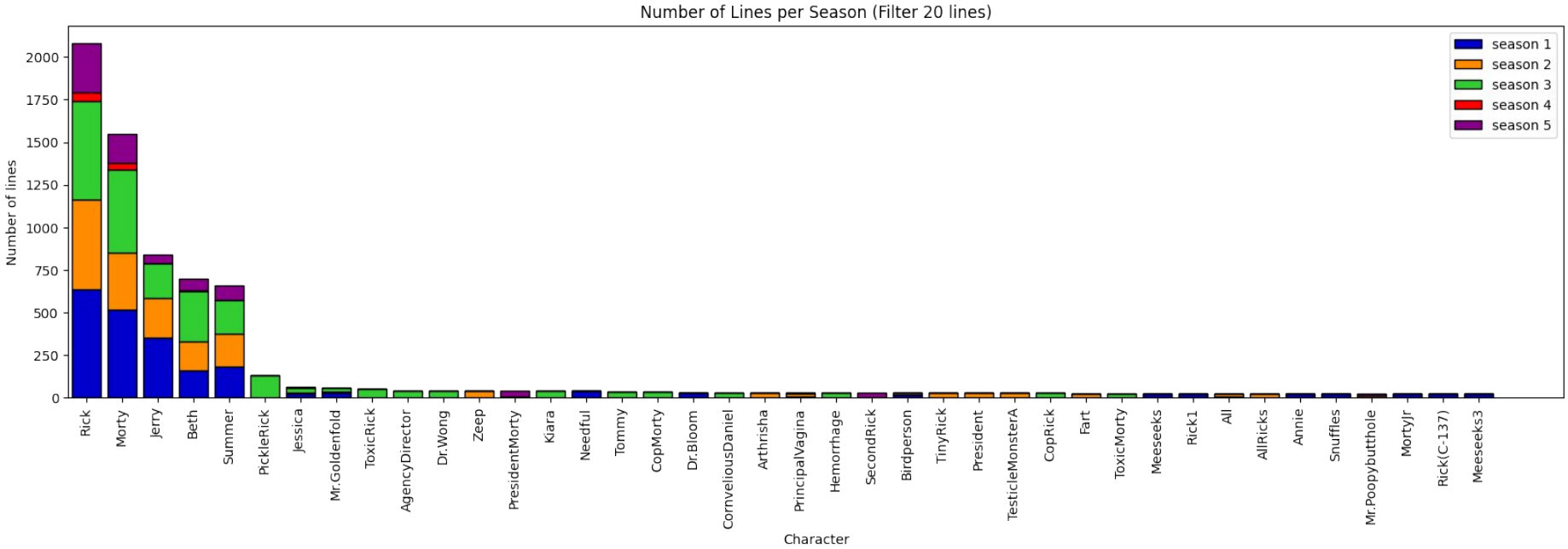
family terms

html code



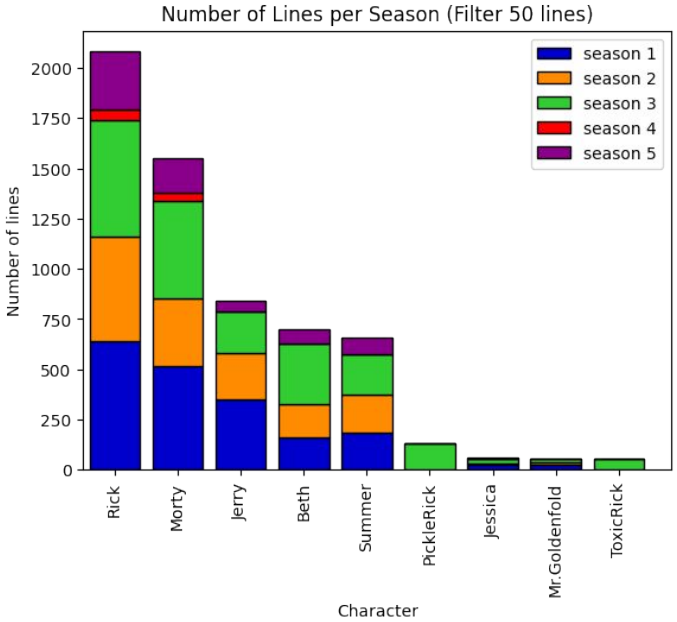
Data Exploration

Speaker distribution (transcript)



Data Exploration

Speaker distribution (transcript)



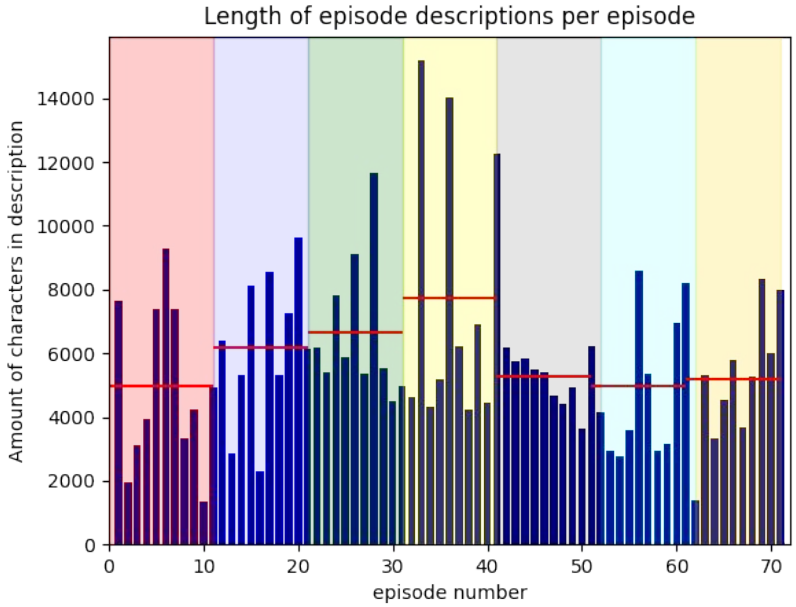
Family Sanchez shares main part

Characters turn into side characters (Pickle Rick)

Rick and Morty most frequently (esp. late seasons)

Data Exploration

Length of episode description



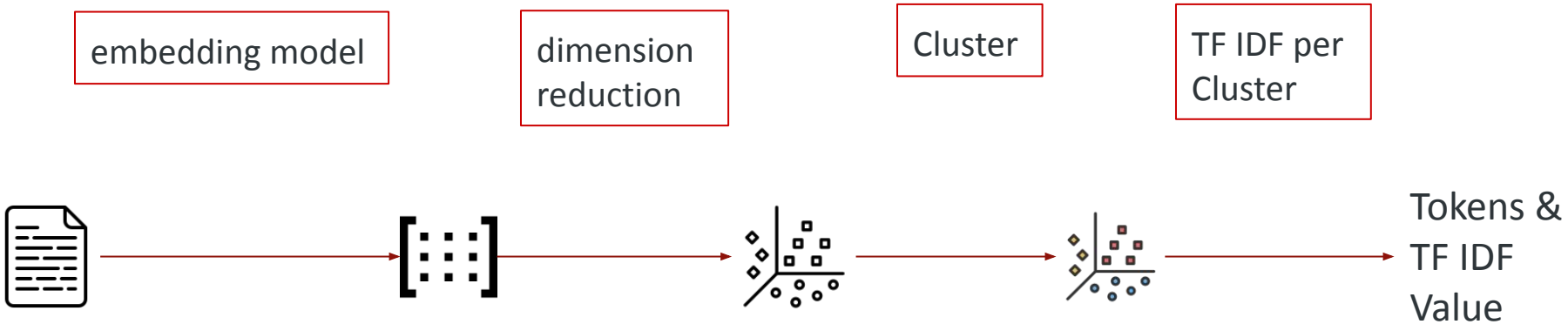
episode description varies (e.g season 4)

avrg description increases until season 4

episode description varies

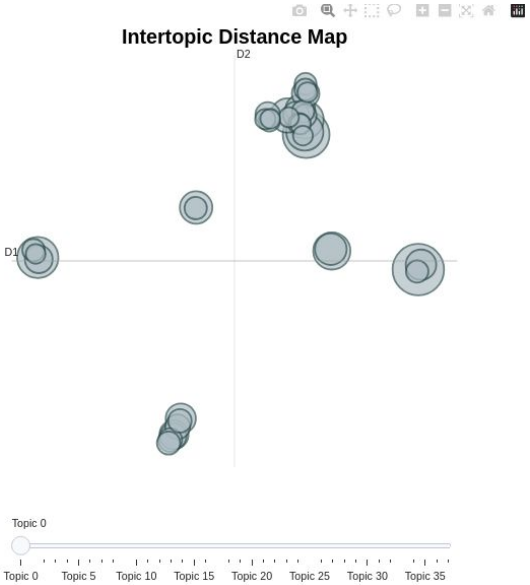
Topic Modeling

Transcript (BERTopic)



Topic Modeling

Transcript (BERTopic)



intertopic distance map with UMAP

Min Cluster size (50)

37 Topics ⟶ overlap

left: family
bottom: adventures
top: colloquial language
top: planetary system

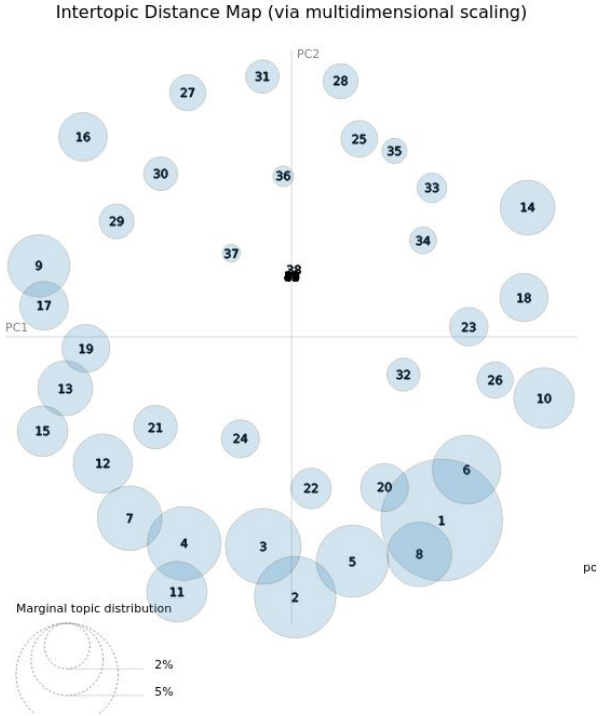
Topic Modeling

Transcript Examples (BERTopic)

Topic Modeling

episode descriptions (LDA)

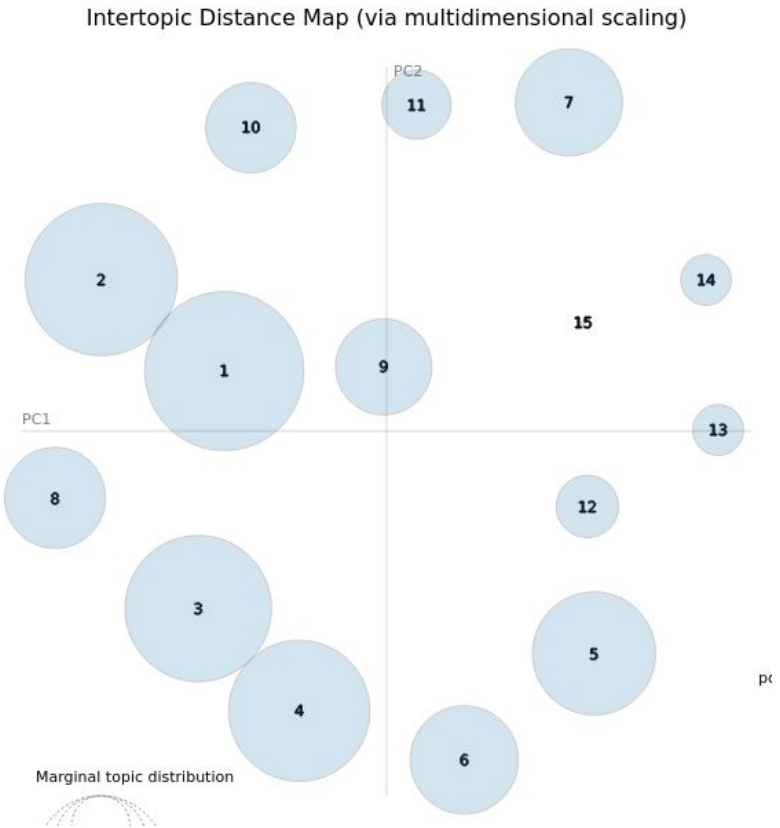
hardly overlapping clusters (40)



Topic Modeling

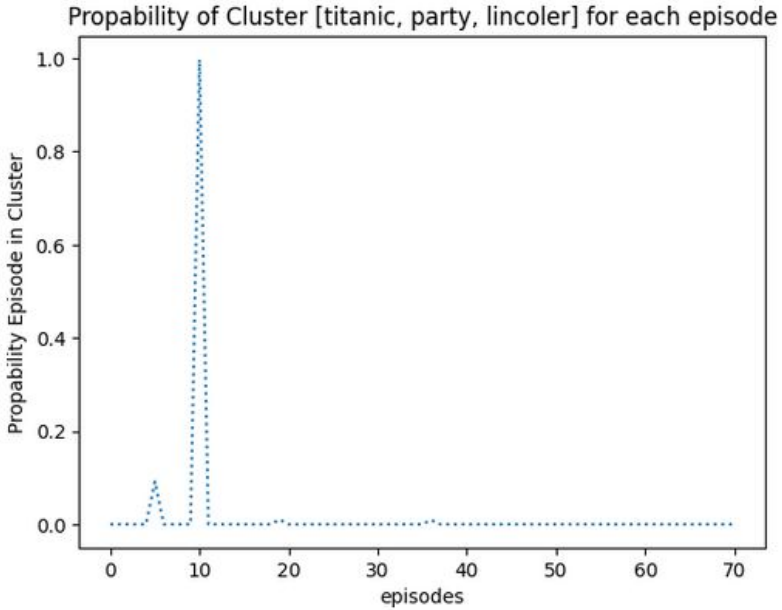
episode descriptions (LDA)

hardly overlapping clusters (15)



Topic Modeling

episode descriptions (LDA)



P(z|d) -> Dirchlat Prop. topic by given document

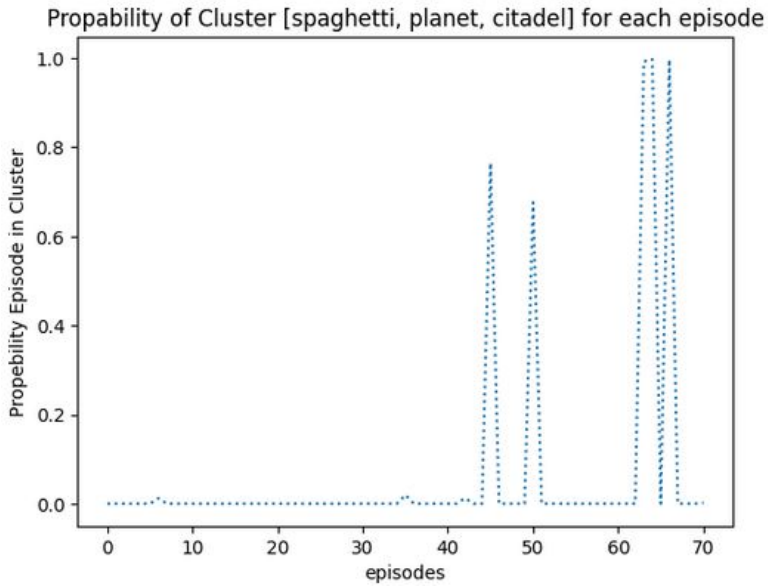
topic terms:

- titanic
- party
- lincoler

→ Topic has high prop in a few episodes (titanic party) and appearance of lincoler

Topic Modeling

episode descriptions (LDA)



P(z|d) -> Dirchlat Prop. topic by given document

topic terms:

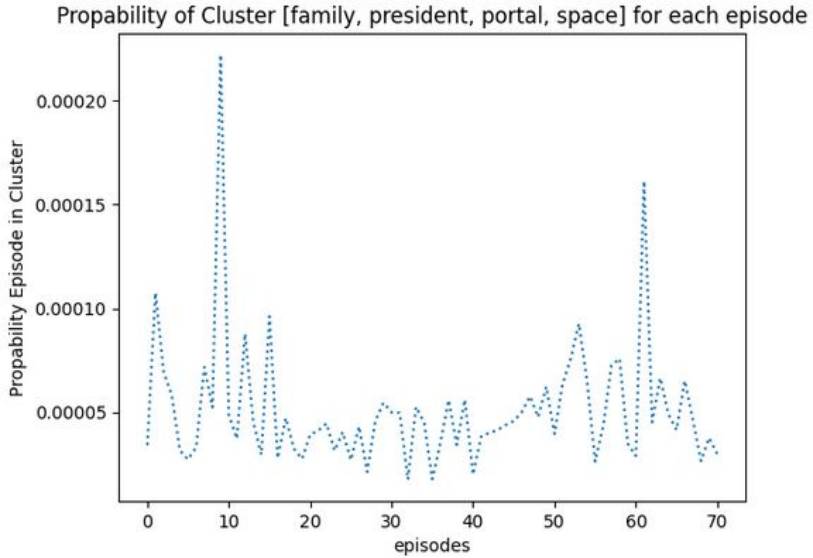
- spaghetti
- planet
- citadel

Topic Modeling

episode descriptions (LDA)

topic terms:
- family
- president
- space

→ topic present in every episode



Sentiment Analysis - Aim & Goals

Aim:

- Analyze emotional trends
- Explore how episodes and events shape character emotions

Goals:

- sentiment over the course of the series
- Character Specific Analyses
- Linking emotional extremes to specific episodes and their narrative

Sentiment Analysis - Methodology

Model:

- pre-trained DistilRoBERTa-based transformer model

Methodology:

- Focus on six basic emotions:
  - anger, disgust, fear, joy, sadness and surprise
- Rolling average applied to reduce noise and highlight trends

Sentiment Analysis - Restrictions

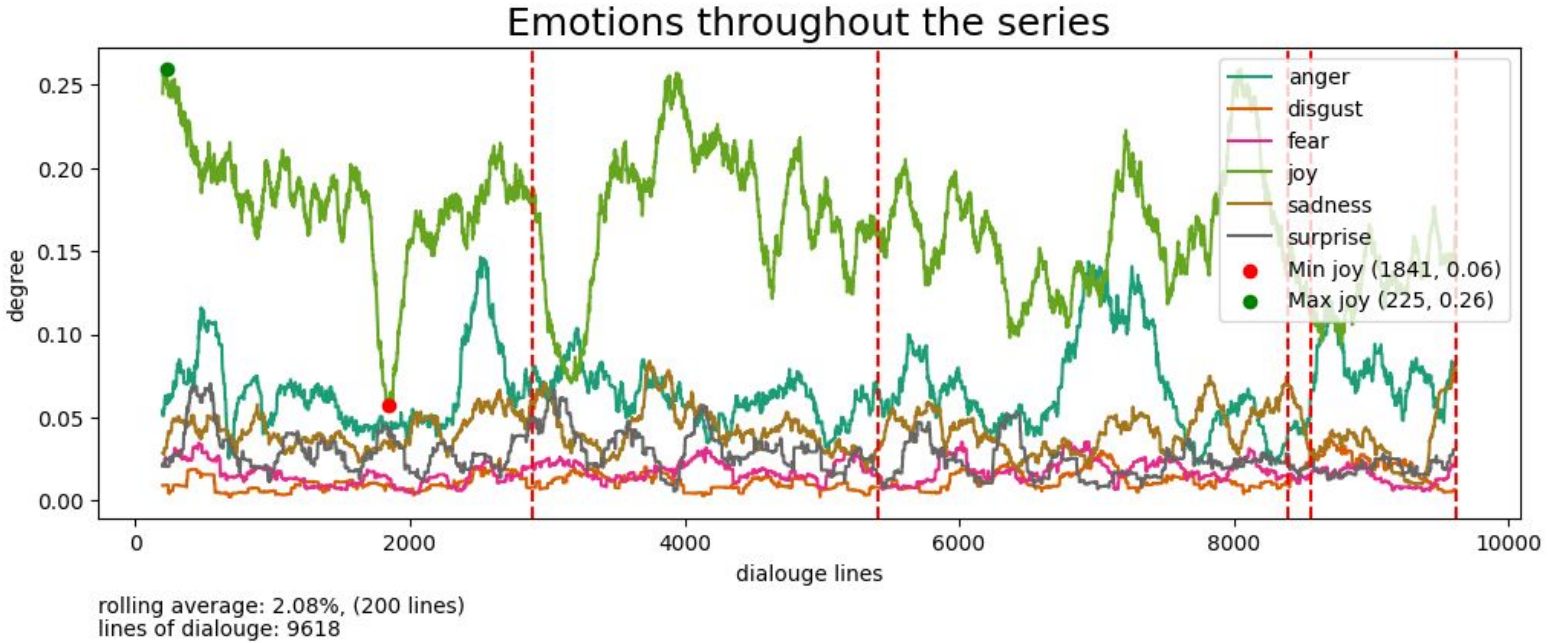
- Lack of a method to evaluate our findings

=> subjective reasoning based on domain knowledge

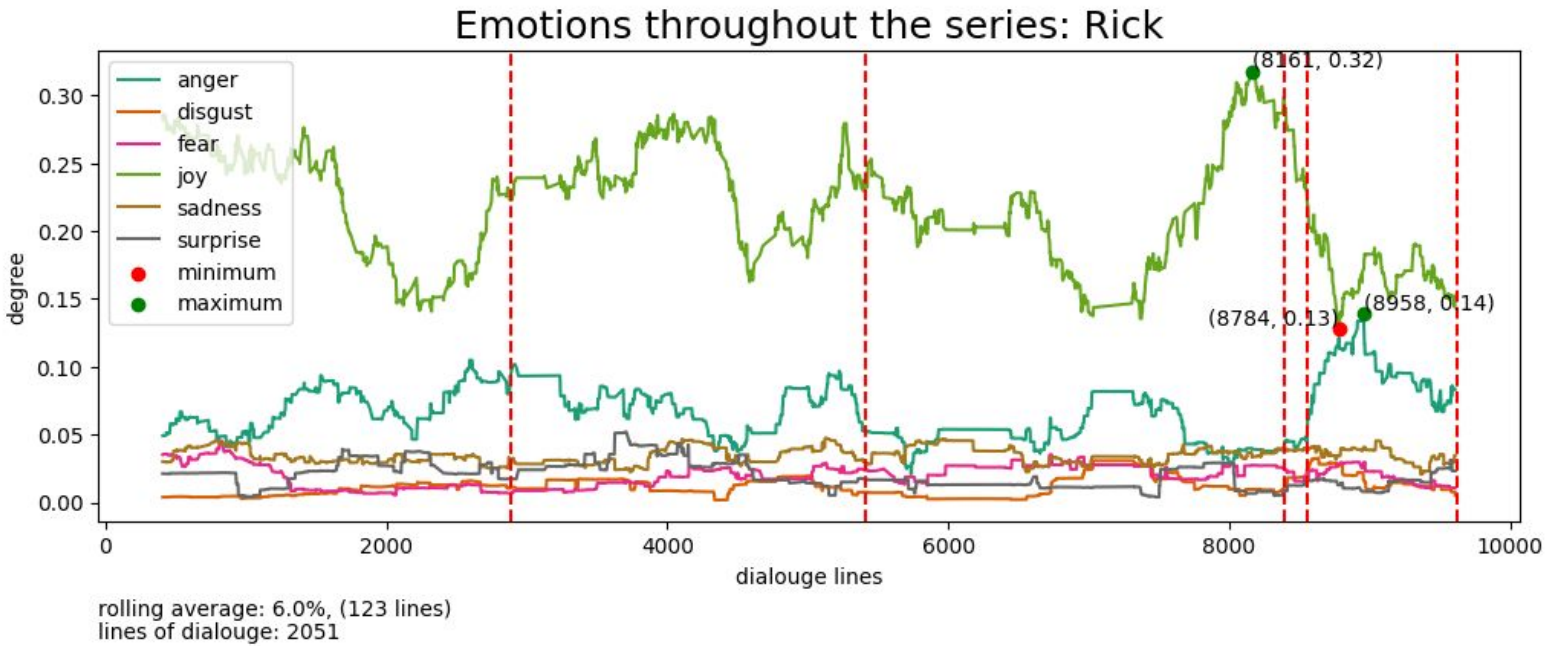
=> subject to hindsight bias

thus, to minimize the effect, we'll only be looking at the most prominent extremes

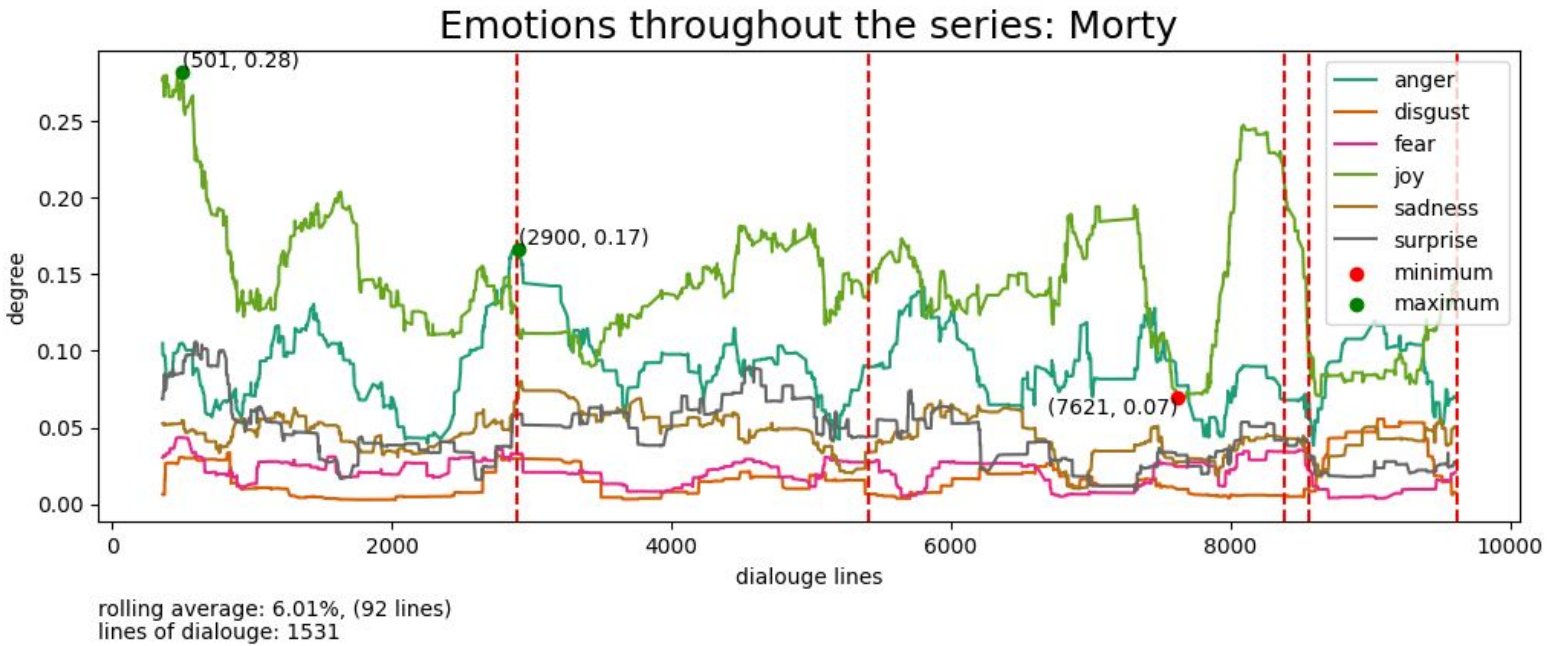
Sentiment Analysis - Series



Sentiment Analysis - Rick



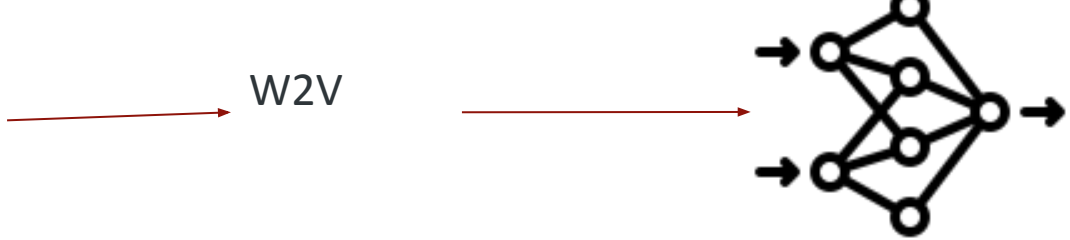
Sentiment Analysis - Morty



Modeling IMDB Prediction

LSTM Regression

Lorem ipsum dolor
sit amet, consetetur
sadipscing elitr



Modeling IMDB Prediction

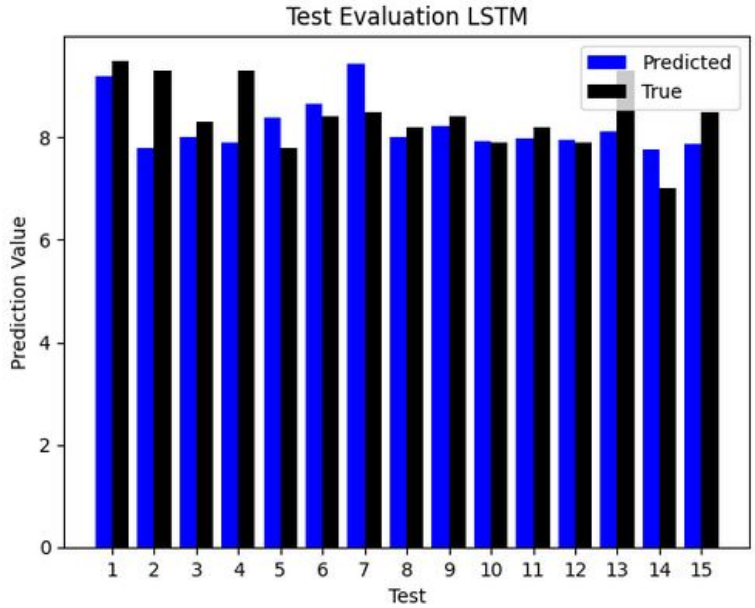
LSTM Regression - Results

Details:

- epochs: 45
- MAE (5CV) : 0.49
- Dropout 2x: (0,2)

Test Results:

“rick is on an adventure” ⟶ 7.2
“rick and morty are on an adventure” ⟶ 10.0
“rick and birdperson are on an adventure” ⟶ 6.6
“rick and morty” ⟶ 9.6
“morty and rick” ⟶ 8.3



Modeling Author recognition

DistillbertSequenceClassification - Results

Details:

- epochs : 5
- F1: 0.61

Test Results:

“earth is a planet” ⟶ jerry

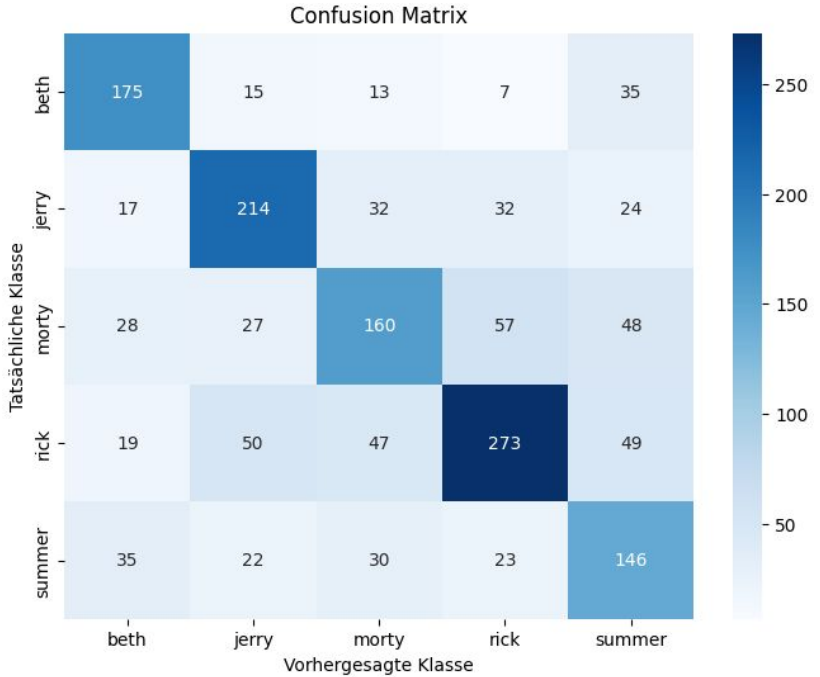
“I love jessica” ⟶ morty

“Thats weird” ⟶ jerry

“Thats weird burp” ⟶ rick

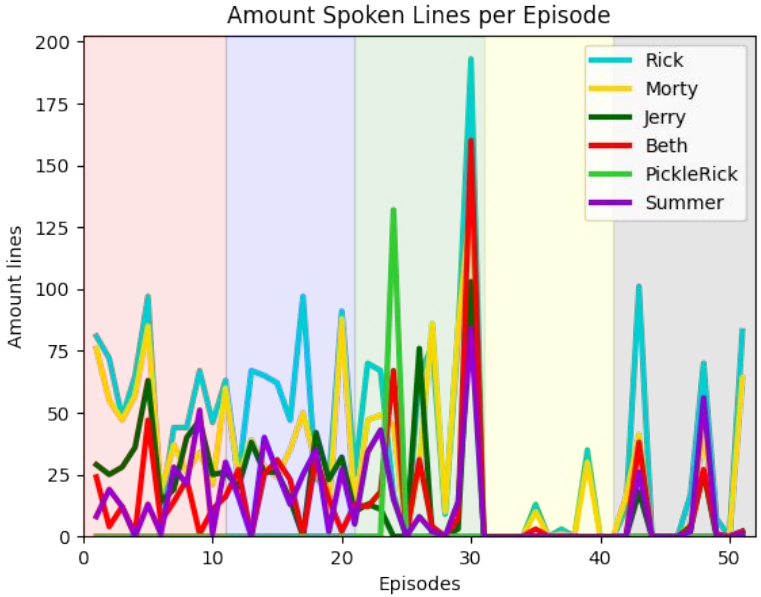
season 6:

“Come on Rick, w-we're almost there.” ⟶ morty



Topic Modeling (Appendix)

Speaker per season (transcript)



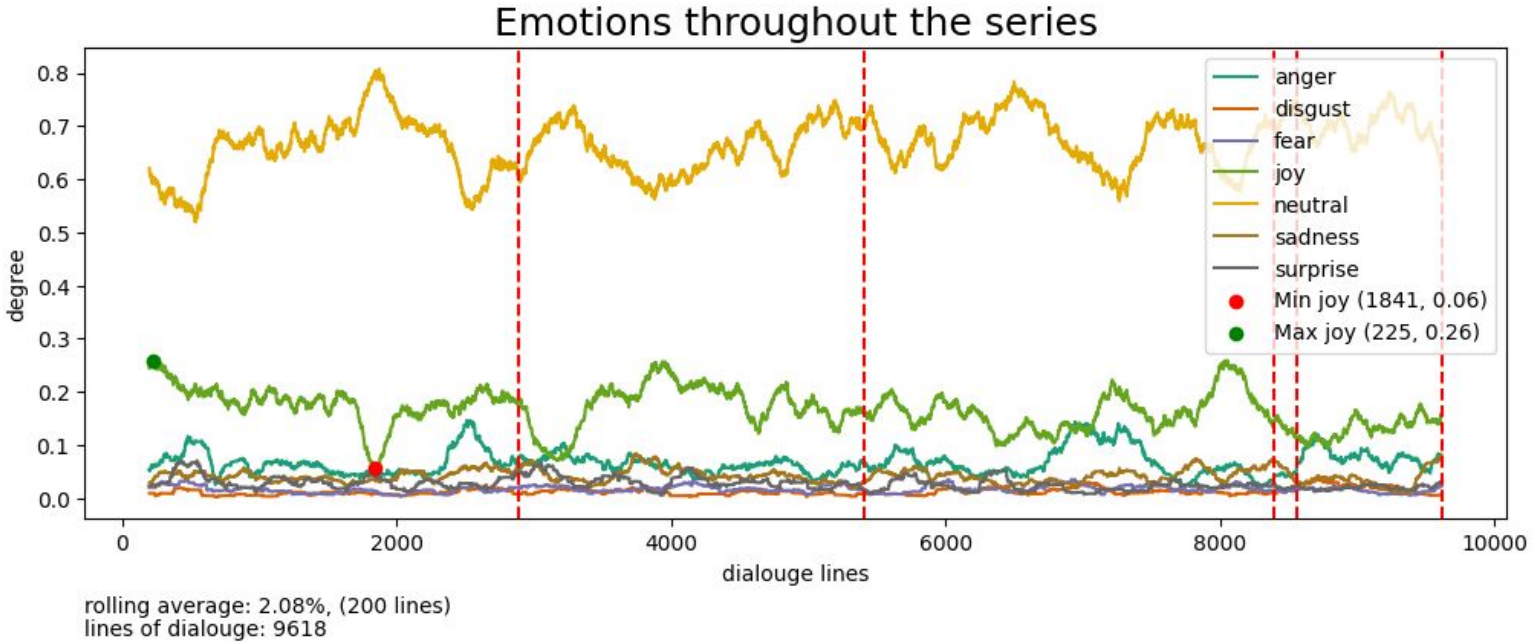
other characters appear in single episodes

lack of data starting from episode 4

Rick and Morty talk more than rest of the family members

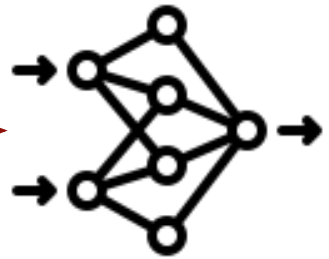
distribution of jerry, beth and summer varies per episode

Sentiment Analysis - Series (Appendix)



Modeling IMDB Prediction (Appendix)

Neural Net Classification

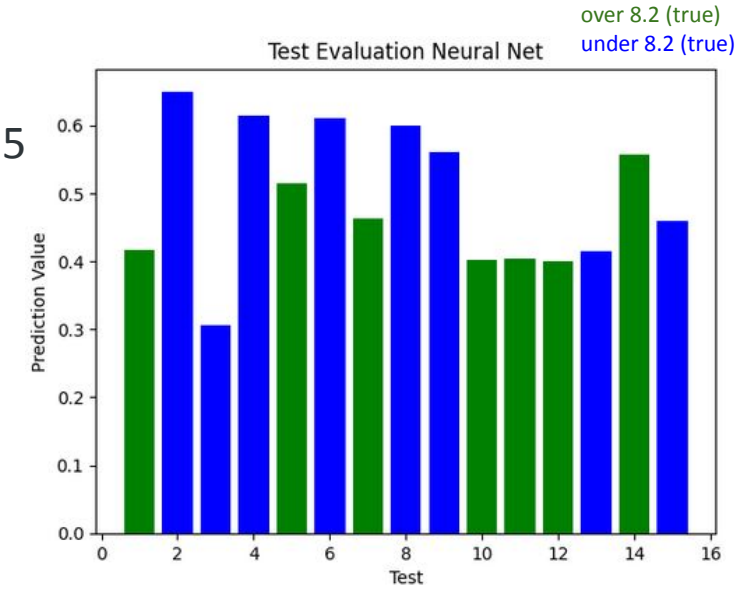
Lorem ipsum dolor sit amet, consetetur sadipscing elitr → Avrg W2V →  → Sigmoid

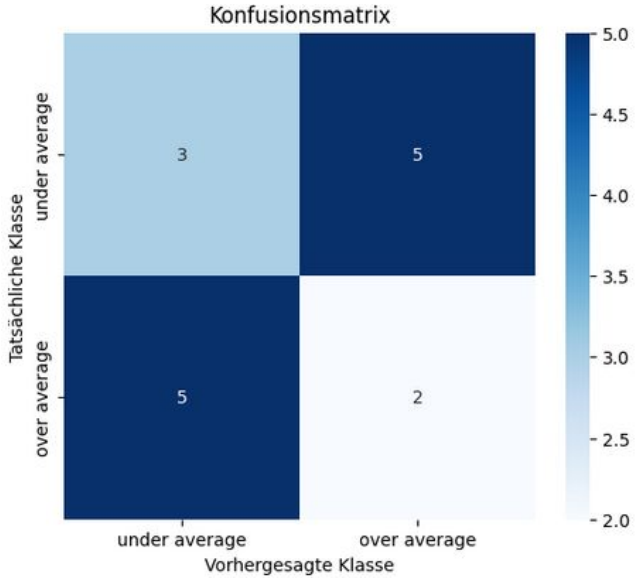
Modeling IMDB Prediction (Appendix)

Neural Net Classification - Results

Details:

- Threshold : 0.5
- epochs: 45
- F1: 0.375



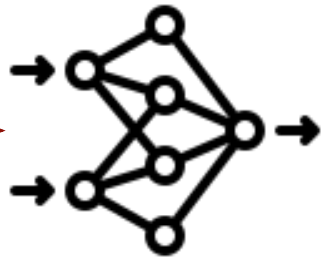


Modeling IMDB Prediction (Appendix)

LSTM Classification

Lorem ipsum dolor
sit amet, consetetur
sadipscing elitr

→ W2V
LSTM →



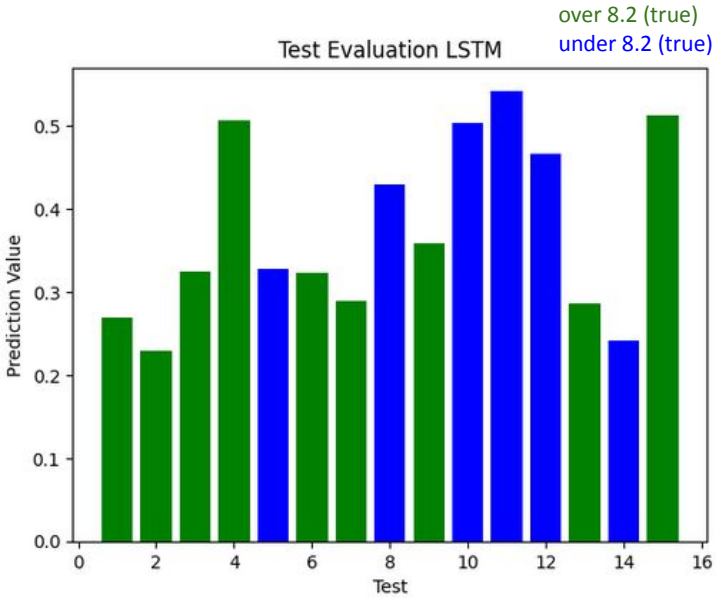
→ Sigmoid

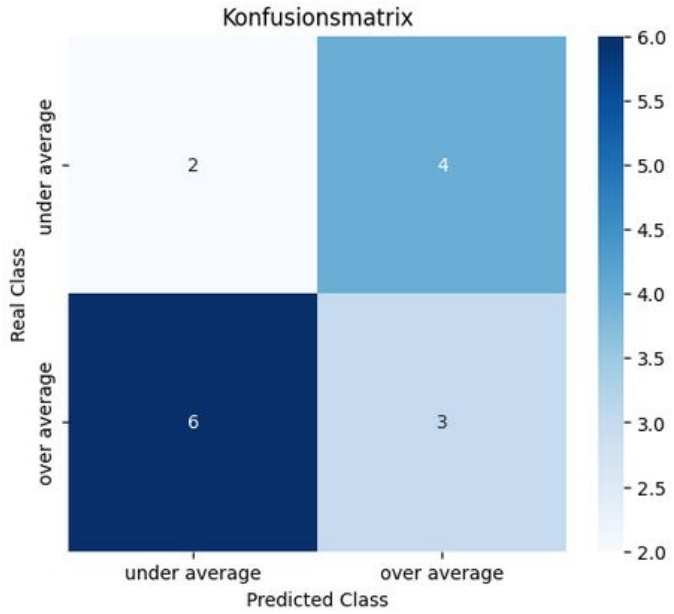
Modeling IMDB Prediction (Appendix)

LSTM Classification - Results

Details:

- Threshold : 0.375
- epochs: 40
- F1: 0.3





Modeling IMDB Prediction (Appendix)

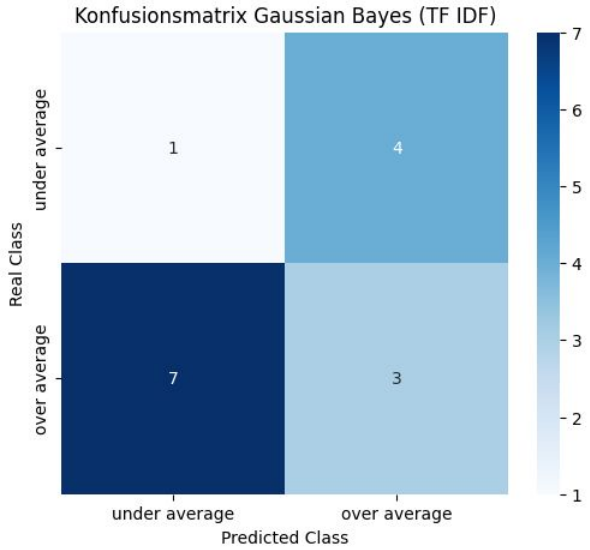
Naive Bayes Classification

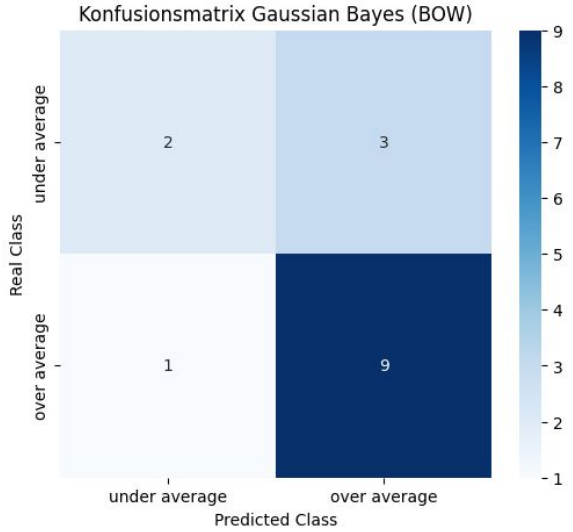
Modeling IMDB Prediction (Appendix)

Naive Bayes - Results

Details:

- F1(TF-IDF): 0.15
- F1(BOW): 0.5





Modeling IMDB Prediction (Appendix)

Linear Regression TF-IDF

Modeling IMDB Prediction (Appendix)

Linear Regression (TF-IDF) - Results

IMDB rating
amount term in description

extreme Coefficients:

- president: 8.82
- mortytown: 3.57
- family: 3.38
- destruction: -2.16