Burp... NLP! A multidimensional analysis through the characters of Rick and Morty

Seminararbeit zur Vorlesung
Aktuelle Data Science Entwicklungen - Intelligent Text Analysis

für die
Prüfung zum Bachelor of Science

an der Fakultät für Wirtschaft
im Studiengang Wirtschaftsinformatik
in der Studienrichtung Data Science

an der
DHBW Ravensburg

Verfasser:	Anton Geiger, Cedric Philip Fechner
Matrikelnummer:	
Kurs:	WWIDS122
Dualer Partner:	
Anschrift:	
Wiss. Betreuer:	Dr. Oliver R. Sampson
Abgabedatum:	31.03.2025

Contents

List of Figures

1 Introduction

The animated series Rick and Morty has captivated audiences with its unique blend of absurd humor, intricate storytelling, and profound philosophical undertones. At its core, the show revolves around the adventures of Rick Sanchez, a brilliant yet morally ambiguous scientist, and his grandson Morty Smith, a kind-hearted but often overwhelmed teenager. Together, they navigate a multiverse filled with bizarre creatures and darkly comedic scenarios while exploring existential questions. Beyond its surface-level entertainment, Rick and Morty offers a rich spectrum of themes—ranging from existentialism and family dynamics to critiques of societal norms — making it a compelling subject for deeper analysis.

Natural Language Processing (NLP) provides a powerful toolkit for investigating the linguistic and thematic complexities of Rick and Morty. By applying NLP techniques to the show's dialogue and narrative structure, we can uncover patterns in character interactions, emotional expression, and thematic evolution. This paper, titled Burp NLP: A Multi-Dimensional Analysis through the Characters of Rick and Morty, seeks to explore several key questions: What recurring themes emerge across episodes? How do characters express emotions, and how do these expressions align with the show's narrative arcs? Can we distinguish characters based on their speaking style, and even predict who is likely to have spoken a given line of dialogue? And finally, can we predict episode ratings based on textual content?

To address these questions, we employ a variety of NLP methods, including topic modeling to identify dominant themes, sentiment analysis to assess emotional tone, and speaker detection to differentiate character contributions. Additionally, we develop a custom model to predict IMDB ratings based on textual features extracted from episode transcripts. Our analysis is conducted on a dataset comprising transcripts from the first five seasons of Rick and Morty, providing a solid foundation for exploring the show's linguistic and thematic patterns.

The paper is structured as follows: Chapter 2 provides an overview of the dataset, detailing its sources, collection process, and initial insights into its structure and quality. Chapter 3 outlines the necessary preprocessing steps to prepare the data for analysis. In Chapter 4, we present our findings from topic modeling and sentiment analysis, followed by the development and evaluation of our speaker detection and IMDB rating prediction models in Chapter 5. Finally, we conclude with a discussion of our findings, highlighting limitations and suggesting directions for future research.

2 Data Understanding

For our project we use three different datasets. The first one is a dataset is downloaded from Kaggle (Prarabdha 2025, n.p.) where we find the rick and morty transcripts until season 5 formatted in the following shape (**Tab. 2.1**). As the transcript also holds information about the scenery, there is not just spoken text included in there. We also find short descriptions about the sorroundings and the scenery. An example is also visible in the first row of table (**Tab. 2.1**).

episode no.	speaker	dialouge
1	Rick	stumbles in drunkenly, and turns on the lights. Morty! You gotta come on. Jus'... you gotta come with me.
1	Morty	rubs his eyes. What, Rick? What's going on?
1	Rick	I got a surprise for you, Morty.

Table 2.1: Rick and Morty Transcript Dataset

The other datasets are retrieved by self written web scraping scripts from the Rick and Morty Wikipage and the IMDB website (IMDB 2025, n.p.) (Rick and Morty wiki 2025, n.p). With the first script, we were able to create following table containing all episode descriptions provided in the Rick and Morty Fandom (**Tab. 2.2**).

id	title	text
0	Pilot	In the middle of the night, an obviously drunk Rick bursts..
1	Lawnmower Dog	Jerry complains that the family dog, Snuffles, is stupid ...
2	Anatomy Park (Episode)	It's Christmas, and Jerry tries to enforce the idea ...

Table 2.2: Rick and Morty episode descriptions dataset

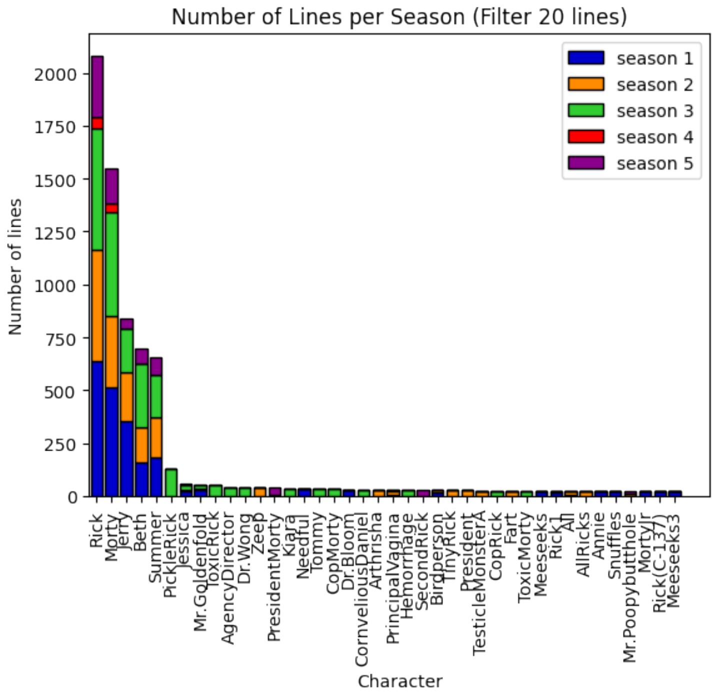
The last datasource stores information about the average IMDB rating per episode in the format provided in sample (**Tab. 2.3**) where we also find the related season and title for each episode.

id	season	episode	title	rating
0	S1	E1	Pilot	7.9
1	S1	E2	Lawnmower Dog	8.6

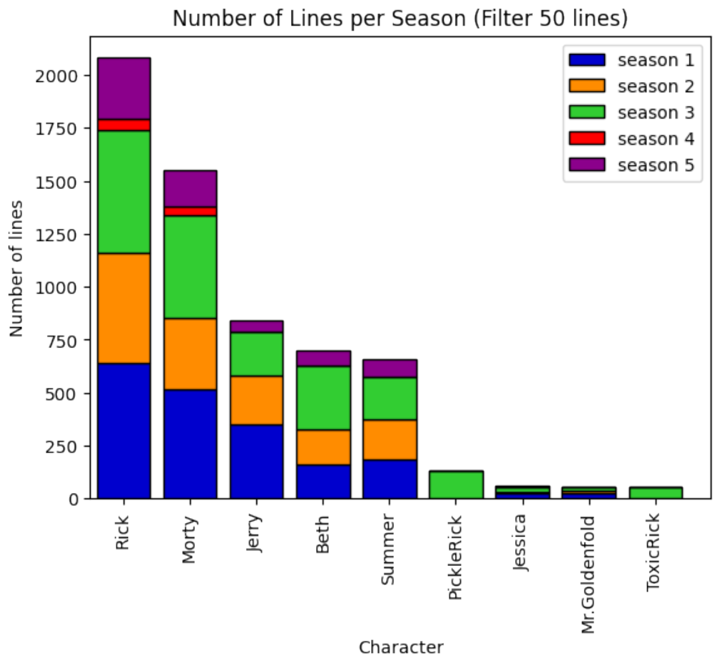
Table 2.3: IMDB ratings per episode dataset

The rating score provided in the table (**Tab. 2.3**) is the average rating score on IMDB calculated on every rating that was given to an episode. For the last two data sets we not only have content up to season 5 but also up to season 7 in the data maintained.

Further analysis show that there are in general 970 different speakers. The charts (**Fig. 2.1**) visualize the speaker distribution by counting the lines that each character speaks in each season. In general, there are 970 different characters in the series, with the Smith family (Rick, Morty, Jerry, Summer, Beth) making up the majority of the dialogue.



Speaker distribution of speakers with more than 20 lines



Speaker distribution of speakers with more than 50 lines

Figure 2.1: Speaker distribution per lines

Surprisingly, the transcript dataset has some lacks. As we see in chart (**Fig. 2.2**) that shows the number of dialogues per episode, the dataset is incomplete as there are a lot of empty episodes starting from season 4. The background colors highlight the intervals of the different seasons over the whole series.

The chart (**Fig. 2.3**) provides a clearer picture on how often which main charatcer

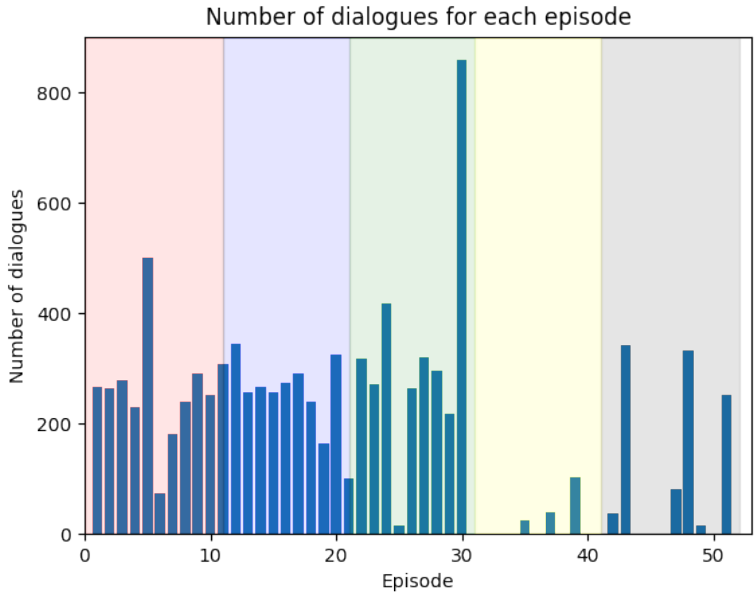


Figure 2.2: Episode distribution (transcript)

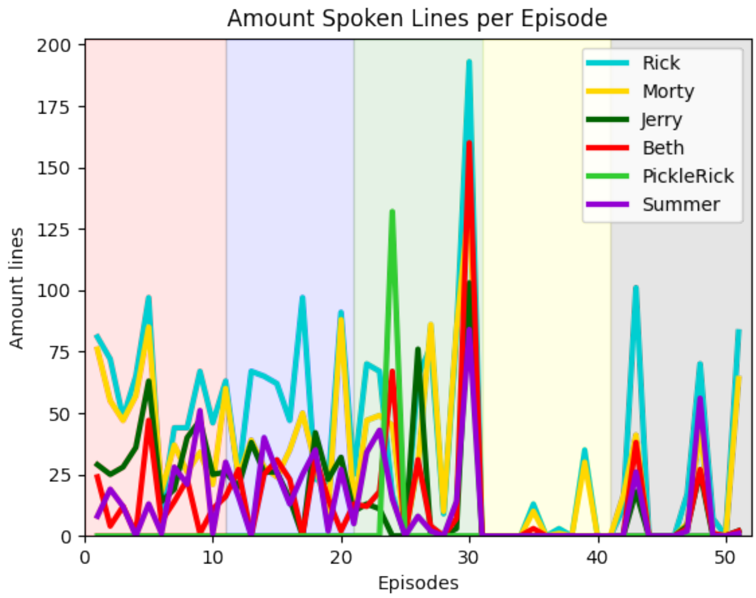
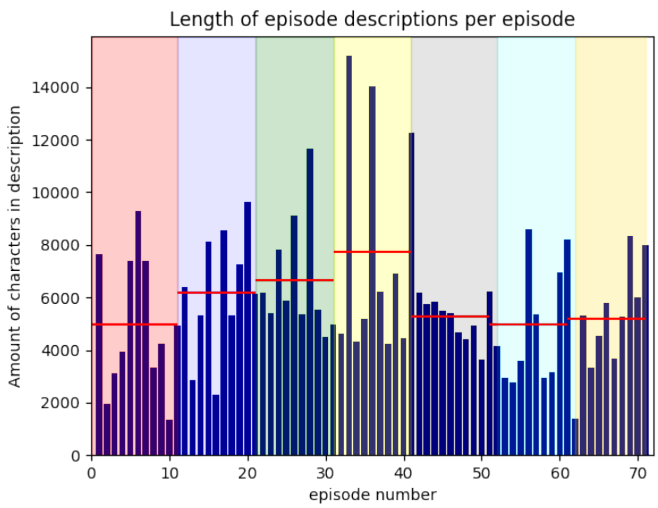


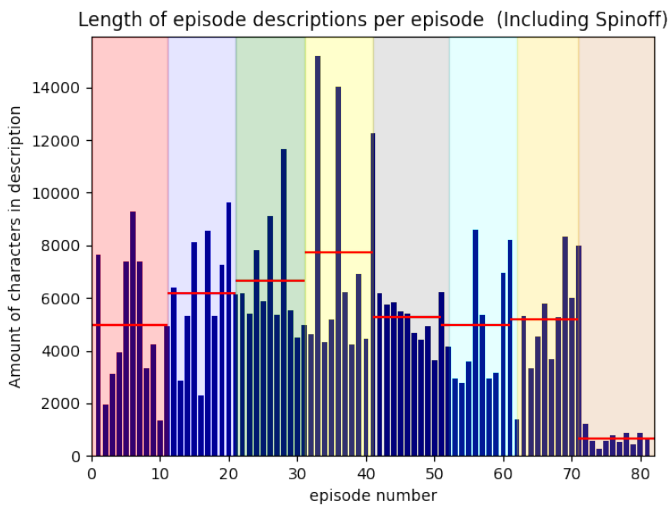
Figure 2.3: Family Sanchez spoken lines in detail

speaks in detail. In most of the episodes, Rick is the main character and holds the most shares in the speaking distribution. The other family members also speak in every episode but not as much as he does. Sidecharacters, like Pickle Rick often just have a large share in a few episodes as they often disapear after one episode.

Looking deeply at the description dataset, we find two charts (**Fig. 2.4**) which compare the length the episode description for each episode. Surprisingly, the descriptions from episode 72 until 81 are way shorter compared to other episodes. These episodes are not part of the normal Rick and Morty series, as they are spin-off episodes. We therefore exclude these episodes from the further analysis. The red lines show the average description length per each season. We can see that the length of the description is different for each episode. An example can be the comparison between episode 33 and 34, where episode 33's descriptions contains more than 3 times the amounf of characters as the description of episode 34. Another fact is that from season 5 onwards, the average number of characters in the descriptions seems to remain at a similar level. That is different compared to the first 4 seasons in which the average length of the description increased in each season.



Episode Length description distribution



Episode Length description distribution (Including Spinoffs)

Figure 2.4: Length of episode descriptions per episode

The last analysis step is an analysis regarding the top 100 most spoken words in the transcript. We used a wordcloud (Mueller n.d., p. 6) to visualize those in an appropriate way. The results can be seen in (**Fig. 2.5**). After removing the character names, we find colloquial terms in the word cloud such as 'yeah', 'gonna' but also family and space related words. We can therefore assume that the spoken English is not very advanced.

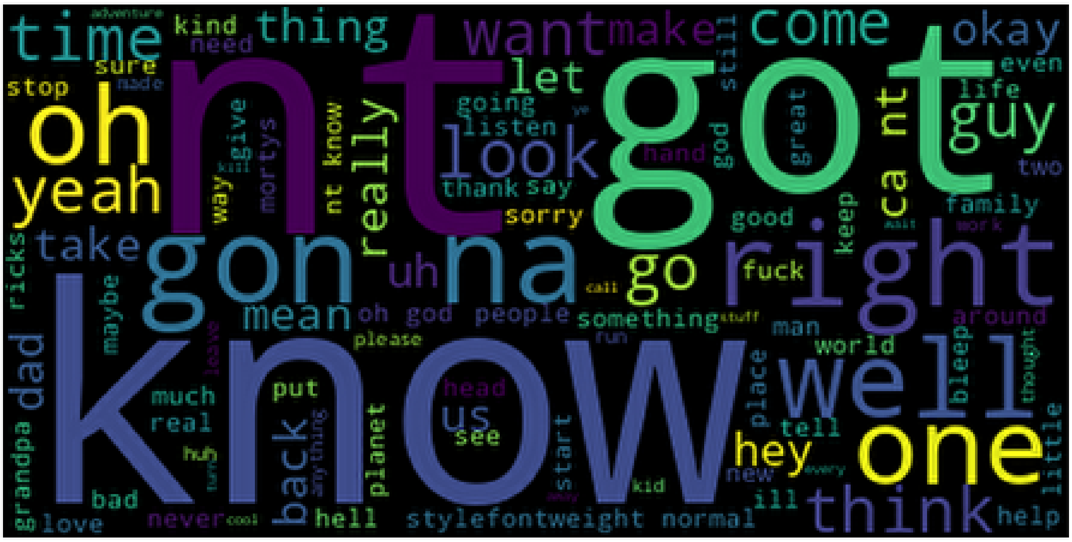


Figure 2.5: Word Cloud transcript

3 Data Preparation

Here, we describe the procedure we followed to clean and prepare the data. To clean up and standardize the data, we removed all characters without spaces and set each character to lower. Since the Rick and Morty characters speak to each other in colloquial language, we removed some abbreviations such as "ain't" with "is not". In addition, the data set contains some HTML tags that were also removed during the data preparation phase using regex.

Using the spacy and nltk library, the text preparation pipeline consists of a custom build stopwordremover and a custom build stemmer which uses the Porterstemmer to reduce every single token in a sentence to the word stem but ignores the character names for stemming. Porterstemmer looks at te suffix of the words and removes it if neccesary (Sushmita 2014, p. 3).

Before that, the sentences were tokenized using the small_en_web model spacy provides. To see the results, the first dialogue turns from

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

into

stumbl drunkenli turn light morty got tocom jus got come

4 Analyses

4.1 Topic Modeling

In this section, we want to examine the series Rick and Morty with regard to the themes addressed in the individual seasons. To do this, we will look at the transcripts and all episode descriptions, excluding the names of the main characters as they are not really content.

Topic Modeling Transcript

At first, the topics where analyzed based on the transcript of the first 5 seasons. As the dataset contains roughly 9.000 datapoints with a small length, we use BERTopic model to receive the most relevant topics. To use the BERT Topic model, we have downloaded the all-MiniLM-L6v2 embedding model to embed the text. The embedded text was then reduced by UMAP and clustered by the hierarchical HDBscan cluster algorithm using the eom cluster selection method. After clustering, BERTopic calculates the TF-IDF Scores for each documents and reduces the topics to the size the user wants them to have (Grootendorst 2022, p. 3). The resulting topic map showcased in the image (**Fig. 4.1**).

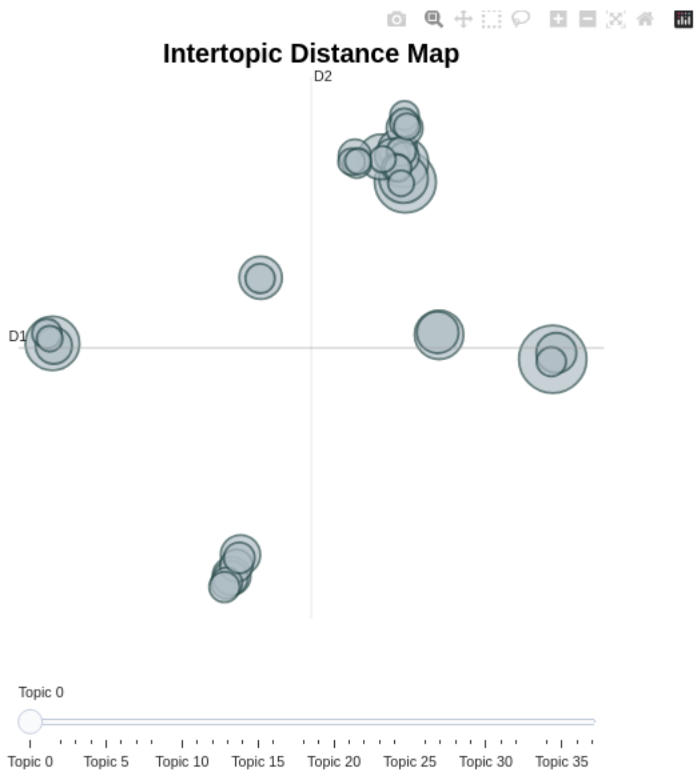


Figure 4.1: BERT Topics for Transcript (all seasons)

Setting the minimum topic size to 40 words, the BERTopic model created 37 different topics. As in the figure provided, we see that most of the bubbles overlap. Therefore there would be the possibility to decrease the number of topics consindering that this would lead to more imature topics. analyzing the content of the topics, we get on the left side topics that can be summarized with the keyword family. There we find family related terms like dad, mom, grandpa. The topics at the bottom of the screen mostly relate to general information about the adventures that Rick and Morty experience, such as adventure, portal gun and treasure. The word portal gun in particular is a very important word, as Rick's portal gun is an important part of the series. Most of the other topics can be summarized by the speech habits of the individual characters. As there are topics where there are slang words like 'wubabdubadabab', 'geez' , 'oh' or 'crap'. Surpisingly, there is one topic located next to those slang words containing information about the planetary system which contains words like pluto, planet and space.

In (**Fig. A.2**), we see a visualization of the most important topics and its words. It turns out, that there are topics containing slang words like 'oh' or 'man'. Topic 2 and topic 4 show that the series is about a human family. Topic 0, 11 and 16 contain more information about the content of the series as there are a lot of terms referring to space and murdering activities that Rick and Morty are exploring along their adventures in space.

Topic Modeling on episode description

Since the dialogues of the series are full of abbreviations and slang without much content, we wanted to take a closer look at the topics by looking at all the episode descriptions. This time, we analyzed the topics using Latent Dirichlet Allocation (LDA), as there are far fewer data points, each containing more than 512 tokens which is the limit of BERTopic. To setup our LDA model we create an Document Term Matrix (DTM) containing all words and documents (Albanese 2022, n.p). Another dictionary maps the words with a given id. We also limit the number of topics to 50 so that there are fewer topics than episodes. In the series, the episodes do not depend on each other as Rick and Morty often discover new planets and characters each episode. The following topic analysis can lead to the same conclusion as a lot of the created topics just contain information about the content of one or a few episodes. In the chart (**Fig. 4.2**) we see the Dirichlet priorities (Albanese 2022, n.p) of the topic containing words such as family and portal gun along all episodes.

This topic is an exception as it holds terms, such as "family" and "portal gun" which are frequently used in many episodes of the series and genereal topics of the series. As there are many topics that strongly belong to single episodes, the overall propabilities are very low for this topic in all episodes.

As already mentioned, most of the other topics just refer to a few or one single episode. This is the case for the topic 'titanic party' where jerry and beth attend a party that recreates the titanic drama. As the chart (**Fig. 4.3(b)**) shows, this topic holds information that is is mainly present in one episode.

The other spikes that are showcased in the chart can be explained as the sidecharacter

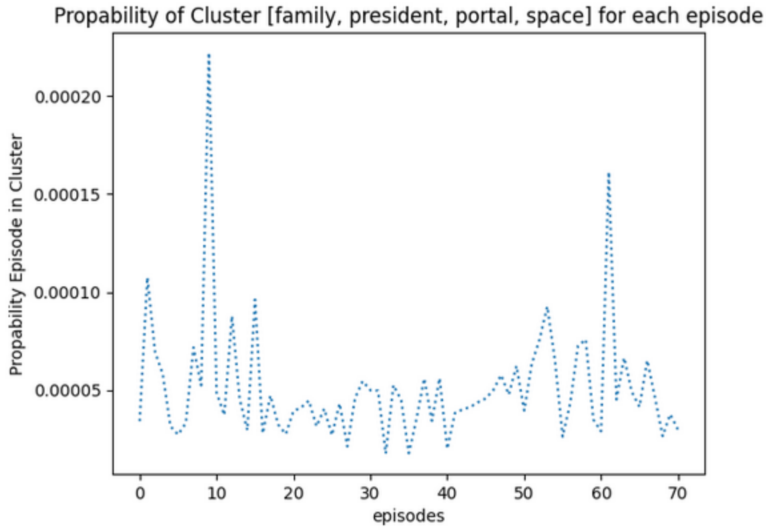
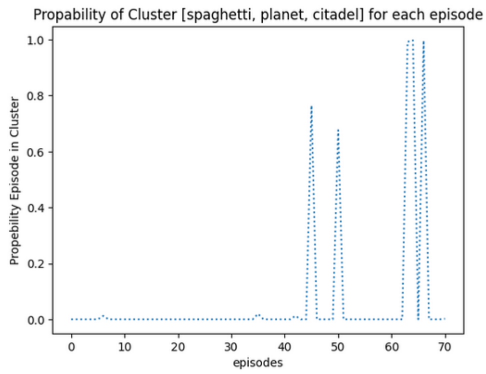


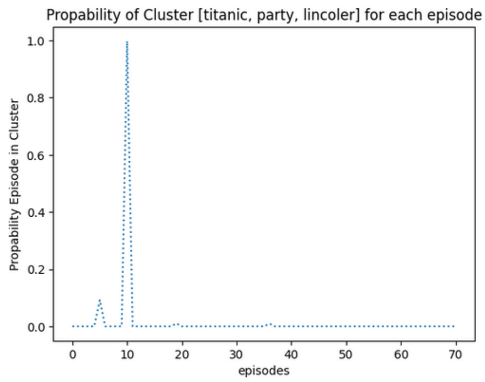
Figure 4.2: Topic relevance Family Portal Gun over time

'Abradolf' 'Lincoler' appears in other episodes as well.

Another example for a similar distribution is the topic about the spaghetti planet and its citadel, which is analysed in chart (**Fig. 4.3(a)**).



Topic 'Spaghett Citadel' in all episodes



Topic 'titanic party lincoler' in all episodes

Figure 4.3: episode related topics over time

The fact that most of the topics differ a lot from each other, can also be seen here in the bubble chart (**Fig. A.4**). It is remarkable that a vast amount of the topics are distributed across the chart while just a handful topics seem to overlap. In addition, the size of the bubbles indicates the number of documents in which the topic appears. Since many bubbles appear to be small, this could mean that the topics mainly refer to a few episode descriptions.

4.2 Sentiment Analysis

In this section, we want to highlight how different emotions developed throughout the first five seasons of the show and which events caused the rise or fall of certain emotions. For this task, we used a pre-trained DistilRoBERTa-based transformer model, which predicts 6 dimensions of basic emotions, as well as a neutral class. The emotions include anger, disgust, fear, joy, sadness, and surprise.
To emphasize trends in the data, we applied a rolling average to reduce short-term fluctuations. This improves the readability of the resulting visualisation since there is less noise and it helps to identify key points in the series which may be characterised by a sudden drastic change in emotion or a turning point in the long-term development. The latter is not to be expected, as successive episodes of the series are generally hardly connected to each other and storylines are told over the span of several episodes to leave space for intergalactic shenanigans as well as other side stories.
However, there is a downside to using this rolling average, as short, emotionally intense moments in the series can be lost. As one might already deduce from this, it is highly important to select the size of the rolling average carefully and according to one's goals. The same principle also applies to the scaling of the axes because it is also a determining factor in our perception of the graph, which is important since there is no way for us to evaluate our results apart from subjective reasoning based on domain knowledge. Of course, it would have been possible to develop theses beforehand, but it would have required an immense effort that was not feasible in our case. Hence why our interpretations are subject to the hindsight bias.
To counteract this effect and not fall into the trap of overanalyzing, we will only be looking at the global extreme points, which gives us a more objective basis for looking at our results. For the upcoming analyses we decided on omitting the neutral class as it did not provide much useful information and was largely uninterpretable. A graph including the neutral class can be found in the appendix **Fig. A.1**. Removing the neutral classs also had the added benefit of reducing the y-scale-space as the neutral emotion was the dominant one. The remaining emotions are kept within the same context as their relative positioning remains to be important. The size of the rolling average has been chosen on the basis of what is most pleasing to the eye, although the more scientific way would have been to look at how many lines of dialogue are in an episode on average and choose a percentage of that as the size of the rolling average. This method would still leave some

subjectivity, of course, but for obvious reasons it is preferable. Despite this fact, we can justify our choice since we are mainly interested in when extreme points have occurred and their position remains relatively unchanged even for a large frame of reference, which is why we will be analyzing our results on an episode level. Not to mention the fact that we are unable to determine the exact moment an emotion arises anyway, due to fluctuations. Apart from that, calculating the size of the moving average as a percentage of the average number of lines of dialogue per episode would probably have proved difficult, given the varying quality of the transcripts and the lack of a common understanding of what constitutes a line of dialogue.

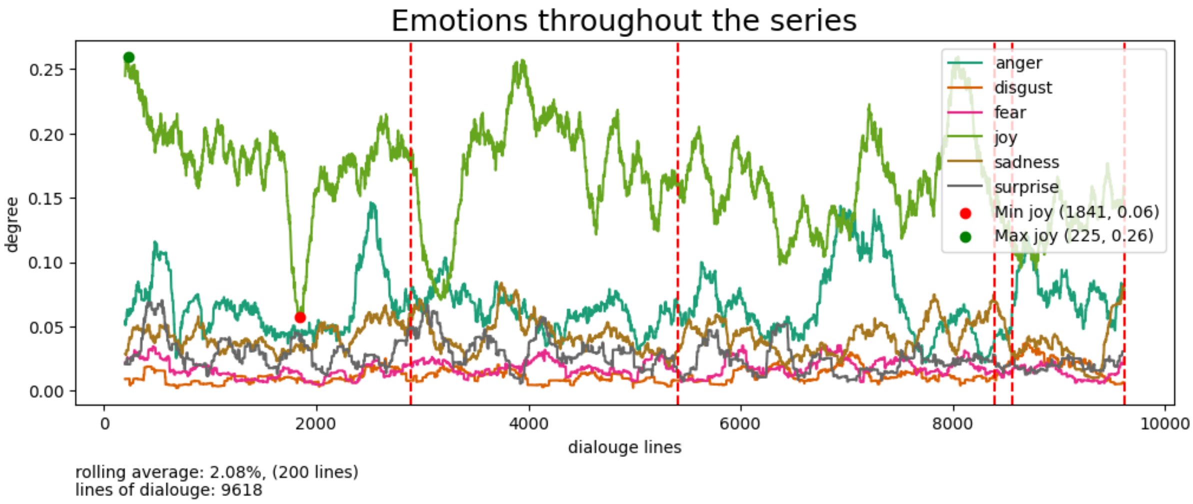


Figure 4.4: Sentiment analysis (excluding neutral sentiment)

Fig. 4.4 displays a propability distribution of the different sentiments - anger, disgust, fear, joy, sadness, and surprise - over the course of the first five seasons, with the red dashed lines marking the end of each season. The green line represents joy, which shows notable peaks and drops throughout the series.

The highest peak in joy occurs right at the start of the series, specifically in the first episode of Season 1, 'Pilot'. Since the 'Pilot' episode marks the very beginning it sets the tone for the entire series, introducing Rick's complex, chaotic and hyperactive personality and his absurd yet funny adventures with his grandson Morty. The humor, absurdity, and excitement of interdimensional travel likely contribute to this high joy score. Morty's first major adventure which includes the comedic escape from an alien world, and Rick's 'over-the-top' antics establish a lighthearted and entertaining atmosphere, leading to an emotional high point.

In contrast, the lowest point for joy appears in Episode 7 of Season 1, 'Raising Gazorpazorp'. This episode takes on a much darker tone, centering around Morty accidentally fathering an alien child that rapidly matures into an aggressive warrior. The themes of violence, neglect, and societal oppression dominate the narrative. Additionally, the subplot with Rick and Summer exploring a dystopian matriarchal society does not lend itself to humor in the same way that other episodes do, reinforcing the low joy sentiment. These underlying themes seem to overshadow the comedic nuances, resulting in a low score for the joy sentiment.

The following analyses provide an individual sentiment analysis of the series' two main characters, Rick Sanchez and Morty Smith, after whom the show is named.

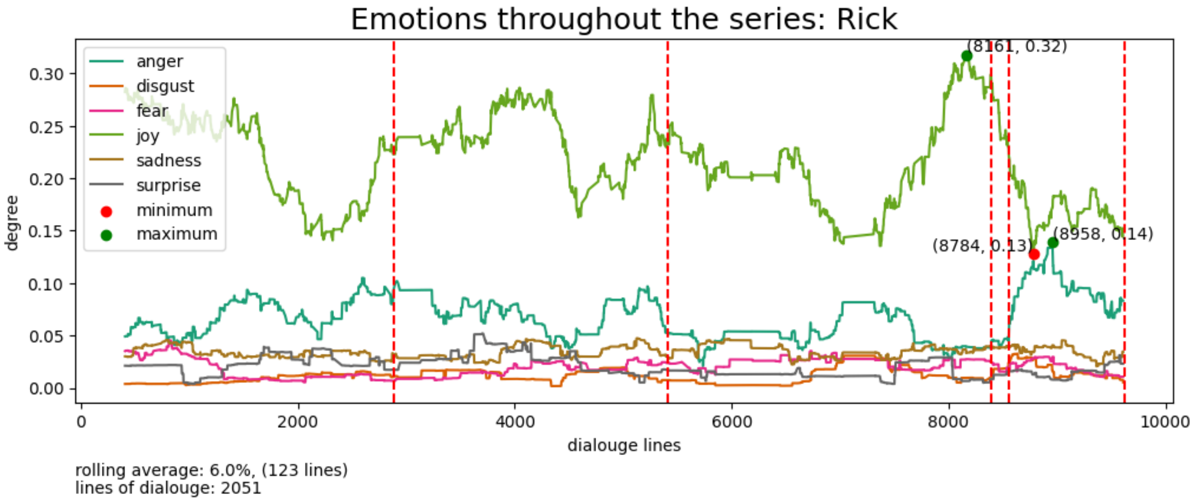


Figure 4.5: Sentiment analysis of rick sanchez

The most prominent characteristic in Rick Sanchezes Emotions (**Fig. 4.5**) is the fluctuating nature of joy, while other emotions remain relatively stable. Rick's highest recorded joy occurs in Season 3, Episode 9, 'The ABCs of Beth', where he enables Beth after some initial dislike and aversion to revisit her childhood fantasy world. However, rather than experiencing unambiguous happiness, Rick's emotional response appears more nuanced. This episode delves deeply into Beth's character, revealing a darker, more ruthless side that mirrors many of Rick's own traits. While Beth initially struggles with the implications of her past, Rick ultimately reassures her that embracing her true nature—even its more violent tendencies—is not inherently negative and they both share a father-daughter moment.

Rick's apparent satisfaction in this episode does not stem from a conventional sense of joy but rather from a complex interplay of emotions, including pride, validation, and perhaps even relief. As someone who often views himself as intellectually and morally superior to others, Rick rarely finds individuals who genuinely understand his complex worldview. In Beth, he sees not just his daughter, but a potential equal—someone who shares his pragmatism, intelligence, and capacity for detachment. This recognition may contribute to the spike in recorded joy, as he momentarily feels a deeper connection with her.

The lowest point of joy appears in Season 5, Episode 2, 'Mortyplicity', which is characterized by paranoia and chaos as Rick and his family are hunted by various decoy families. The stressful and distrustful nature of the episode aligns with the drop in joy.

The emotion anger, which remains relatively stable throughout the series, reaches a sudden peak in Season 3, Episode 5, 'A Rickconvenient Mort'. In this episode, Morty falls in love with a superhero named Planetina, leading to tensions between him and Rick. To combat this frustration, Rick engages in reckless partying with Summer and beginns a destructive relationship with an allien, over which they later have a heated arguement which results in Summer destroying the relationship leaving Rick furious, likely explaining the hightent anger level.

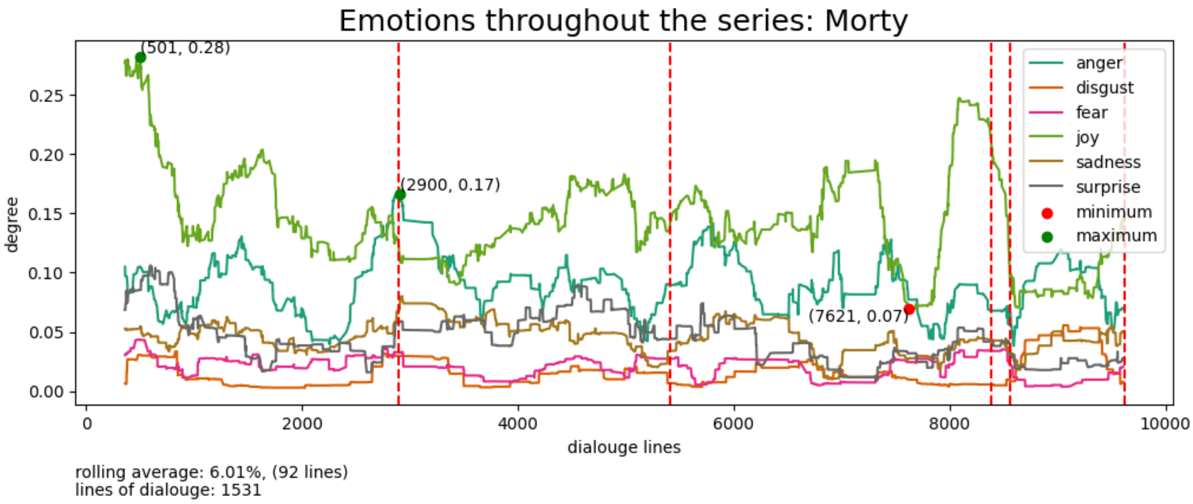


Figure 4.6: Sentiment analysis of morty smith

The sentiment anlysis of Morty (**Fig. 4.6**) also reveals joy as the predominant emotion throughout the development of the show. Two notable extremes in Morty's joy levels occur in Season 1, Episode 2, 'Lawnmower Dog', where his joy reaches its highest point, and after Season 3, Episode 8, 'Morty's Mind Blowers', where it plunges to its lowest.

In 'Lawnmower Dog', Rick manipulates Morty's math teacher, Mr. Goldenfold, into giving Morty straight A's to free up time for more adventures. Initially anxious, Morty gradually adapts to the bizarre situations he and Rick face, growing calmer and even enjoying their unpredictable escapades. Their dynamic is unusually harmonious, with effective teamwork — a rarity as the series progresses — contributing to Morty's rising joy levels.

Before the subplot unravels, Morty befriends Scary Terry, a nightmarish figure in Mr. Goldenfold's dream, by showing him kindness. This unexpected friendship boosts Morty's confidence and adds to his sense of accomplishment. Later, in the subplot, Morty is spared from a global dog uprising led by his hyper-intelligent former pet, Snuffles. He quickly

adapts to the surreal events and shares a heartfelt goodbye with Snuffles as they decided to move to another part of the multiverse, further enhancing his emotional growth and boosting his joy scores.
The combination of Morty's growing confidence, his successful partnership with Rick, the friendship with Scary Terry, and the emotional resolution with Snuffles culminates in one of Morty's highest points of joy in the series.

In contrast, Morty's joy reaches its lowest point after 'Morty's Mind Blowers.' This episode delves into the psychological toll of Morty's adventures, revealing that Rick has been systematically erasing Morty's traumatic memories to keep him functional. As Morty is forced to relive these suppressed memories, he is confronted with moments of fear, humiliation, and existential dread. The episode highlights Morty's hidden trauma and the darker side of his relationship with Rick, showing how these experiences greatly affect his emotional state. This leads to a major drop in his joy levels.

Another key extreme is the spike in anger, which builds up over season 1, Episode 11, 'Ricksy Business'. In this episode, Morty faces extreme stress due to the wild party Rick throws at their house. Morty gets increasingly frustrated over Rick's irresponsibility and lack of of care for the consequences which his parents threatened.
However the most interesting emotional shift in the graph occurs toward the end of Season 3, Episode 5, in 'A Rickconvenient Mort'. As stated in Rick Sanchez's sentiment analysis, this is the episode where Morty falls in love with Planetina, with whom he begins a relationship. This is directly in line with our previous findings, which makes it all the more interesting.

5 Modeling

5.1 IMDB Rating prediction

In this section our goal is to perform a modeling that predicts the IMDB ratings based on rick and morty episode descriptions. Therefore, we compare different strategies and architectures to solve this complex problem.
Our hypothesis is that there are some important words or characters like “birdperson” that could have an impact on the IMDB rating, as some characters or planets are more popular in the Rick and Morty universe than others.
Creating a perfectly fitting model is a challenge, as episode descriptions are written in a neutral style and the characteristics that determine whether an episode is liked or disliked can vary from person to person. In addition, there are also visual effects and musical elements such as the “Get Schwifty” or Snake Jazz song that can lead to a higher IMDB

rating. Therefore, we assume that our model will generally not be the most accurate.

Vector Embedding Approach

As a neural network cannot calculate with strings, we have to represent the given text as numeric representations. To achieve this, our intial approach uses the Word2Vec library to generate a vector embedding for all the words that are that appear at least 5 times in all descriptions. We used Word2Vec as we are able to get high value dense vector representations for low computational costs (Mikolov et al. 2013, p. 10). By that, we were able to represent most words in the Rick and Morty corpus as an vector of 100 values. These 100 values also defined the input shape of our neural network which was designed for classifying the episode descriptions. To represent whole tokenized sentences into a vector containing 100 dimensions, we tokenized all of the training data and calculated the average vector of each episode description. As a result, each episode description can be represented as one single vector of 100 dimensions. With that, we are able to train a neural network with 100 input neurons. If a token was not part of the Word2vec vector embedding list, the token will be skipped. For example, the string 'DHBW Ravensburg' would produce a NaN and the string 'Rick DHBW Ravensburg' would produce the same result as the string 'Rick', since the unkown words like DHBW are ignored.

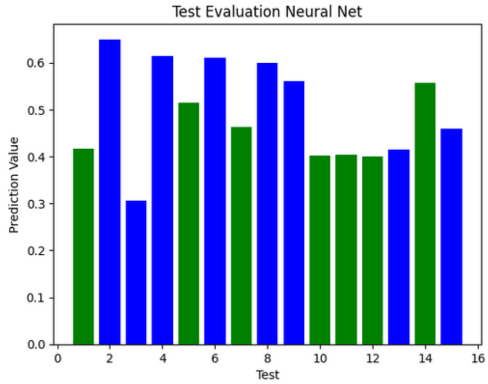
The test results are displayed in the following chart (**Fig. 5.1**). The green bars represent the episodes having an true IMDB rating above 8.2 while the blue ones are rated lower. As we did not set a thresshold yet, we can see the test result visualized in a bar chart where the heights of the bars show the predicted value by the neural net.

As expected, the neural network struggles to distinguish between those two classes. In general, it seems that the models even predicts the complete opposite, as some blue colored bars were predicted higher as the the green ones. The calculated test accuracy based on the results without an thresshold was 33%.

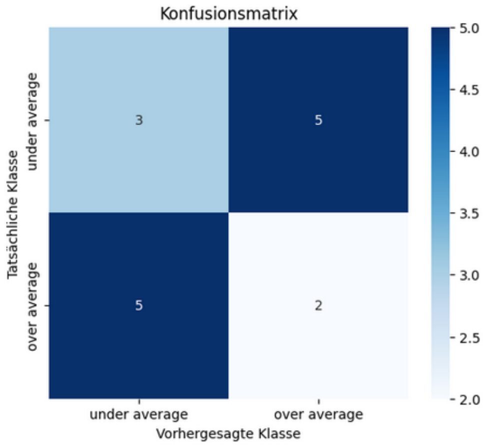
Setting the thresshold to 0.5 we receive following confusion matrix as an result Fig. (**Fig. 5.1**). The confusion matrix with a threshold of 0.5 also leads to a similar result, as there are more incorrect predictions than correct ones. This shows us that this model definitely does not solve the problem.

LSTM Approach

As the first neural network had some difficulties to examine the complex relations between the words, we also implement the Long Short Term Memory (LSTM) architecture to achieve more reliable results. The advantage that a LSTM has is that it can also process the order of the words. We used the same Word2vec embedding model as before and added our Rick and Morty corpus to it because there are a lot of Rick and Morty unique terms like 'birdperson', 'portal gun' in the descriptions which should be embedded



Neural Net Test Evaluation



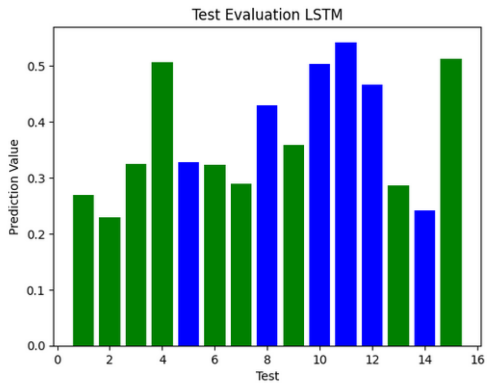
Neural Net confusion matrix (threshold = 0.5)

Figure 5.1: Neural Net Test evaluation

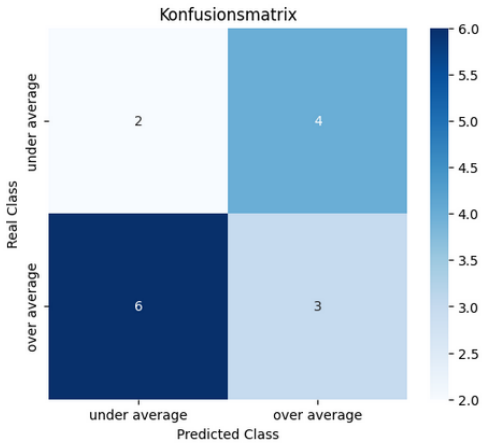
as well. Unlike the previous neural network, which computed the average vector for each sentence, we instead used the first 500 vectors in a sentence to numerically represent the entire text. The results of this approach are displayed in the chart (**Fig. 5.2**). Since the LSTM classification leads to a greater variety in text prediction, it still had some trouble figuring out which descriptions were above and below 8.2. As the confusion matrix displayed in (**Fig. 5.2**) shows, the model is also not good in predicting the classes as it produces more False predicted values as right ones.

Bag of Words Approach

As the word embedding approaches did not lead to a satifying outcome, we also tried to create a Bag of word matrix containing the the amount of the first 500 most appearing words within a text. To clean the data, we decided to remove every token in the description dataset which POS was not classified to PRPN or NOUN by spacy, in order to focus more on the character names and the nouns that are provided in the description. With that, we trained a Naive Bayes model to predict those classes. Therefore, the CountVectorizer by sklearn came into place. But unfortunally, also this approach could not return valid result as the confusion matrix in (**Fig. A.5**) shows.



LTSM Classification



LTSM Classification confusion matrix (treshhold = 0.4)

Figure 5.2: LSTM Classification results

TF-IDF Matrix Approach

Sklearn's TF-IDF vectorizer prioritizes words that occur frequently in a document, but not in all documents. With this last classification approach, we hope to prioritize terms or characters that are less common but more popular than others. Following this procedure of creating an TF-IDF matrix, and training an Gaussian Model again, it also did not lead to satisfying results (**Fig. A.6**).

LSTM Regression

We expect that the characteristics of an episode with a rating of 8.2 (class 0) are not so different from those of another episode with a rating of 8.3 (class 1). Since the correlation coefficient between the numerical IMDB rating value and the class from the classification is only 0.7, some information is definitely lost in the classification. Therefore, we aim to predict the IMDB rating not by classifying the episodes as 1 or 0 but by directly predicting it's IMDB score. With this regression approach, we hope to get better chances to identify the IMDB ratings as a lot of information is lost by transferring the numeric data into classes in the classification problems above.

To achieve this, we used the same architecture as in the classification task but changed the lossfunction as the LSTM classifier but we additionally normalised the Word2vec embedding vectors in the matrix.

In (**Tab. A.3**) we see the test accuracies by epoch size. As the epoche size of 15 leads to the lowest sqaured error, we use 45 as a default epoch size.
A Cross Validation based on models of epoch size 45 returns an average squared error of 0.47 on the 5 validation set (**Tab. A.1**), meaning that, on average, our model's predictions deviate by approximately 0.68 rating points from the true IMDB scores. As a comparison, the statistical standard deviation of all ratings is 0.97. An analysis of the test results is displayed in the chart (**Fig. 5.3**). There we, see the predicted IMDB ratings in blue and the true ones in black. In general, we see that the model predicts valid results, as none of the predicted IMDB scores were above 10 and the deviation between the actual and predicted scores is often small.

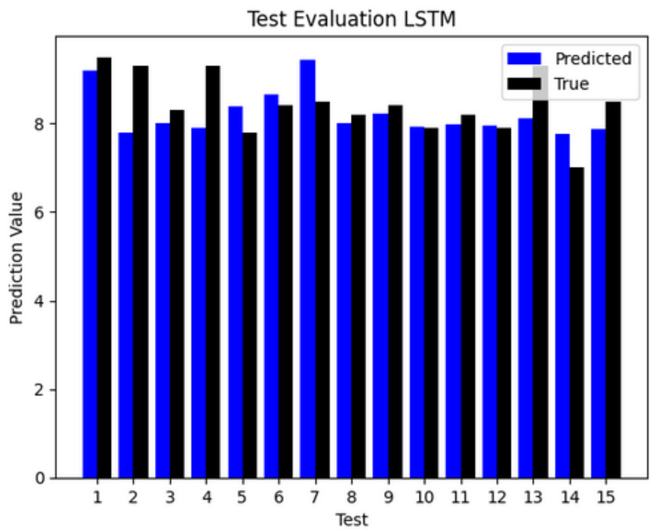


Figure 5.3: Test Results Regression

When testing the model, we noticed that in general, the combination and the appearance of characters influences the IMDB score. For example the string 'rick is on an adventure' leads to an 7.2 while 'rick and morty are on an adventure' leads to 10.1. Handing over 'rick and beth are on an adventure' leads to an output of 9.3. 'rick and bird-person are on an adventure'leads for an example to a score of 6.6. The sentence 'Morty and Rick', for example also just produces a IMDB rating of 8.3 while calculates for the sentence 'Rick and Morty' an IMDB rating of 9.5. The problem that neural networks have is that they are not explainable and is more like a black box. In the last approach we want to figure out more about the terms and its influence in other model architectures.

Regression with TF IDF
The linear regression model with TF IDF matrix as features should examine which character or nouns defnine the IMDB Prediction. To achieve this, we trained a linar OLS regression model by statsmodelapi and printed out the tokens that influence the predicted IMDB score the most. According to this model, the feature term president has

the highest coefficient along terms like saber, universe or mortytown (**Tab. A.4**). The terms with the lowest coefficient are destruction, garage and gun (**Tab. A.5**). The word president appears in the episode 16 and 18 and 27 that all are under the top 5 rated rick and morty episodes. A more impactful example is the term mortytown that just appearead in episode 27, the highest rated episode as visible in (**Fig. 5.4**), where the blue line indicates the amount of the word in an episode and the yellow line shows the IMDB rating.

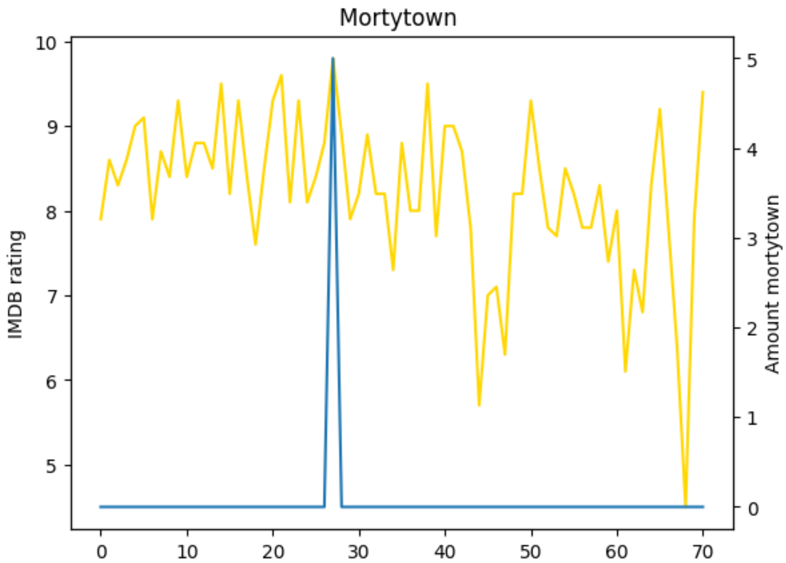


Figure 5.4: Number of Mortytown apperances with IMDB prediction

With all these different approaches we were not able to create a model which predicts the IMDB prediction correctly. Not just the vector embedding approach leads to a unsatisfying result in the classification architectures, but also the Term matrix approach with BOW and TF IDF could not distinguish between high and low IMDB prediction. One reason for this is the lack of a large data source, as there are only 70 Rick and Morty episodes. Also, most of the characters and planets only appear in a few episodes. While terms from the training set such as Mortytown correctly determine the IMDB ratings in the TF IDF approach, testing the model on new vocabulary will definitely not yield valid results.

Using the regression model, we were able to remove the information loss caused by changing the label format from numerical to qualitative. In doing so, we investigated what influence characters, their appearance and their combination could have on the IMDB prediction, and we were able to identify terms such as Mortytown or President that most strongly influence the rating according to a linear regression model.

5.2 Speaker detection

The last modeling part examines if the character‘'s speaking traits can be identified by a fine tuned transformer model. Therefore, we just use the family sanchez as classes because the other characters talk way less in the episodes as already mentioned in (**Fig. 2.1**). In order to get rid of this imbalaced form, we decided to duplicate the amount of spoken lines of jerry, beth and summer for the modeling part.
For this classification problem, we used the DisitllBERT transformer (Huggingface n.d., n.p) as a base model consisting of the DistillBERT Tokenizer and a Sequence classifier. As the test error stops to decrease at epoch 4 as visible in (**Fig. A.7**), we fine tuned the model with 4 epochs. Doing this, we created a model that lead to 61% accuracy on the test set.
To demonstrate the model and differentiate the characters' speaking habits, the model identifies frequently spoken words for a character in a simple way. For example, the prompt "That's weird" results in Jerry, while "That's weird burp" results in Rick.
In addition, more complicated sentences such as "I love Jessica" are also output as the result of Morty, which is most likely also a quote from him. Also, the phrase "Pluto is a planet" is a phrase that most likely came from Jerry and was correctly predicted by the model, for example, as he and Rick had some discussions about Pluto.

The overall results are visible in following confusion matrix (**Fig. 5.5**) and interpreted with the metrics in table (**Tab. A.2**).

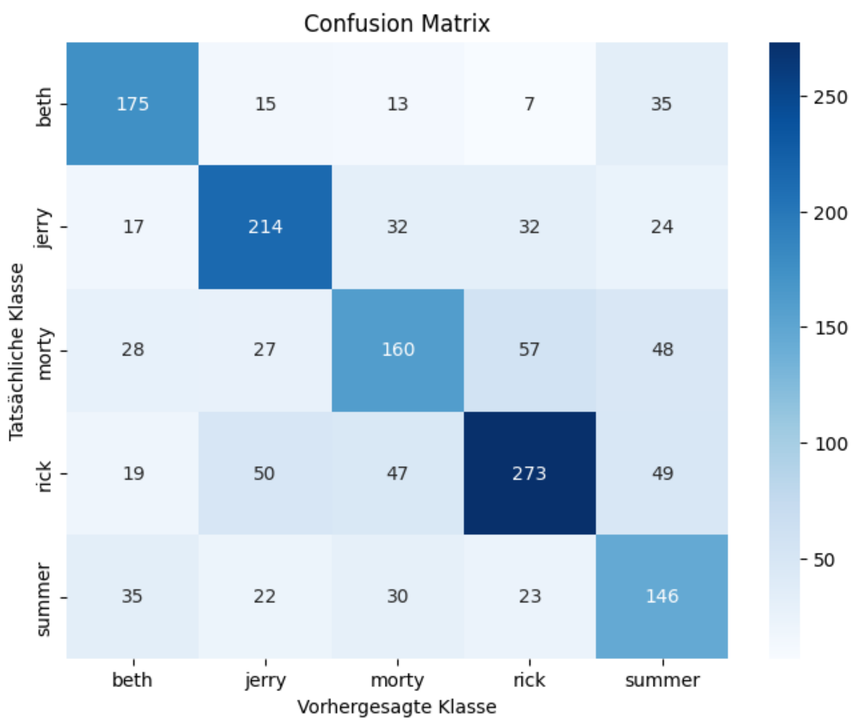


Figure 5.5: Confusion Matrix speaker detection

6 Discussion

By summarizing the topic modeling, we were able to extract relevant content that plays an important role in the Rick and Morty series. Topics such as family or topics describing outer space were taken up throughout the series. A closer look at the topics reveals that most of the found topics only refer to one episode, which allows the conclusion that each episode is hardly connected to other episodes.
This conclusion is strengthend by our sentiment analysis results since the sentiment graphs exhibit significant volatility, characterized by sharp spikes, indicating that the sentiment varies greatly from episode to episode and even within a single one. Aditionally we were able to connect certain events in the story to the graph progrenssion providing support for our results.
As for IMDB prediction, our hypothesis proved true that it is very difficult to predict IMDB ratings based on description texts. Each classification approach had some difficulty in predicting whether the episode will have a high or low IMDB prediction.
The regression model is a bit more accurate and with that we were able to extract words like mortytown that possibly are more popular and result into higher IMDB ratings.
Based on some experiments, the model has learned which character combinations have a potentially higher IMDB rating than others. This could lead to the hypothesis that frequent combinations such as Rick and Morty or Rick and Jerry generate a higher IMDB rating than those with lower term frequency such as Rick and Birdperson. The model also learned that more characters can also lead to a higher IMDB rating, as the same set without more characters results in a lower IMDB rating.
With the finetuning of the DistillBERT transformer, we were able to classify the speakers with a test accuracy of 61%, which could lead to the conclusion that the speaking behavior and topics of the individual characters differ from each other.

A Abbildungen

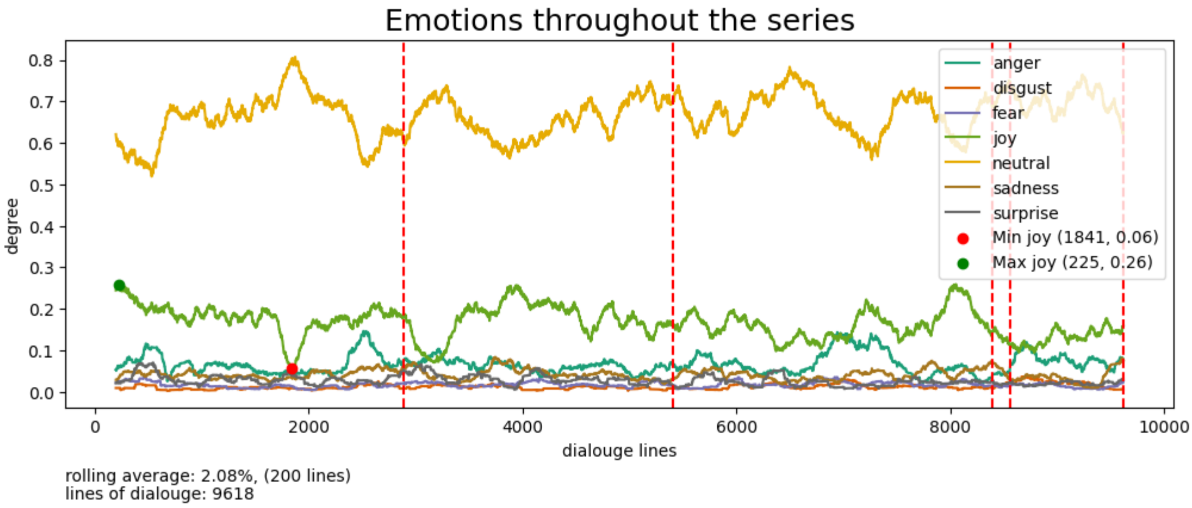


Figure A.1: Full sentiment analysis

mean average error
0.4996431767940521,
0.4830387234687805,
0.5242437124252319,
0.4312141239643097,
0.49884870648384094

Table A.1: 5 Fold Cross Validation

	precision	recall	f1-score	support
beth	0.64	0.71	0.67	245
jerry	0.65	0.67	0.66	319
morty	0.57	0.50	0.53	320
rick	0.70	0.62	0.66	438
summer	0.48	0.57	0.52	256
accuracy			0.61	1578
macro avg	0.61	0.62	0.61	1578
weighted avg	0.62	0.61	0.61	1578

Table A.2: Metrics for speaker detection

epoch	Mean sqaured Error
15	0.45
25	0.479
35	0.51
45	0.43
55	0.50

Table A.3: Test scores accuracy LSTM regression

feature	**coeff**
president	8.821506
saber	5.209245
one	4.810728
churry	4.542387
time	4.510295
suit	4.186404
things	4.167846
curve	4.155876
rift	4.114602
lord	3.994916
fluid	3.846431
mortytown	3.577070
point	3.569252
universe	3.547126
hands	3.388094
family	3.386688
jesus	3.336255
rhett	3.309392
mr	3.251881
chaos	3.172340

Table A.4: Top Coefficients Linear Regresssion (highes score)

feature	**coeff**
destruction	-2.168899
head	-2.116017
success	-1.995059
cop	-1.455023
version	-1.347783
him	-1.310323
fact	-1.238303
locos	-1.191900
jellybean	-1.096124
scene	-1.040677
ricks	-0.976909
robot	-0.916737
court	-0.887538
rock	-0.692056
hand	-0.635960
home	-0.604017
garage	-0.596317
diane	-0.525239
people	-0.487156
gun	-0.471000

Table A.5: Top Coefficients Linear Regresssion (lowest score)

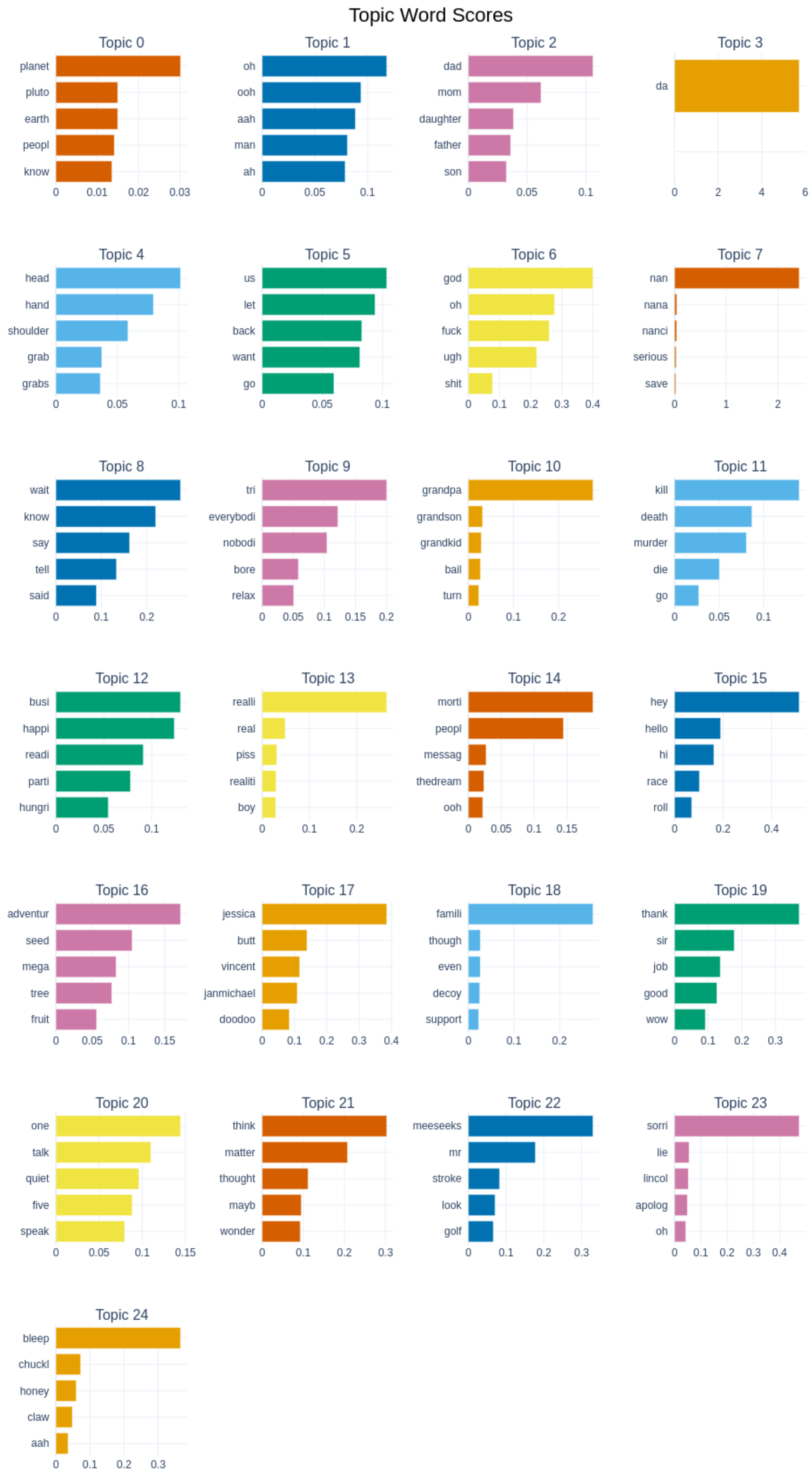


Figure A.2: Top 25 BERT topics for transcript (all seasons)

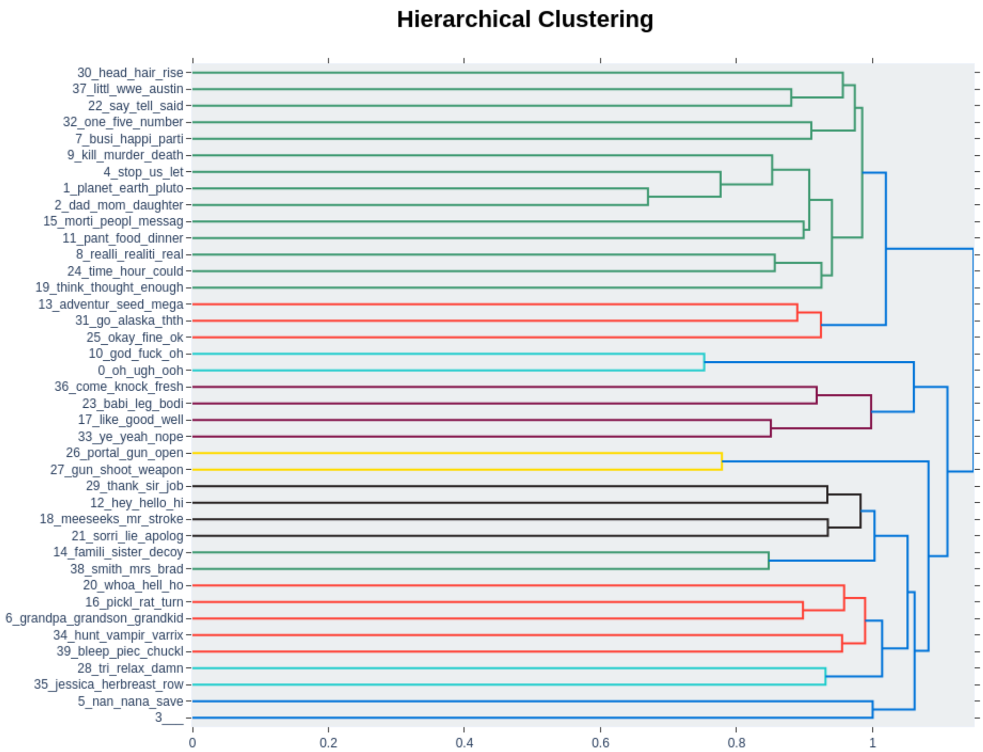


Figure A.3: Hierarchial Clustering topics on transcript

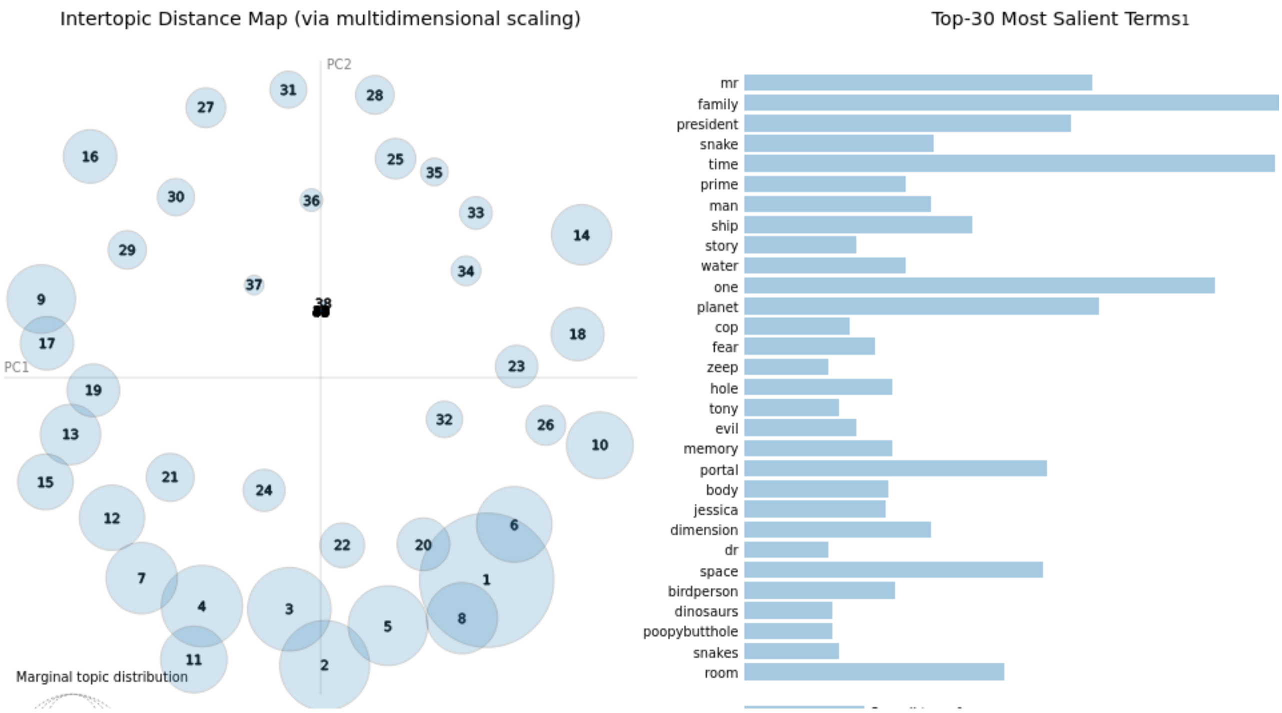


Figure A.4: All Topics LDA for description (all seasons)

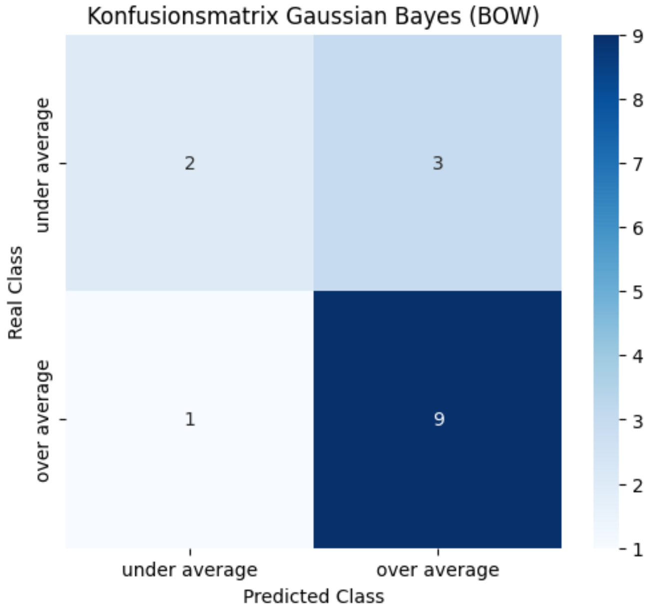


Figure A.5: Confusion Matrix Bayes Classification with BOW

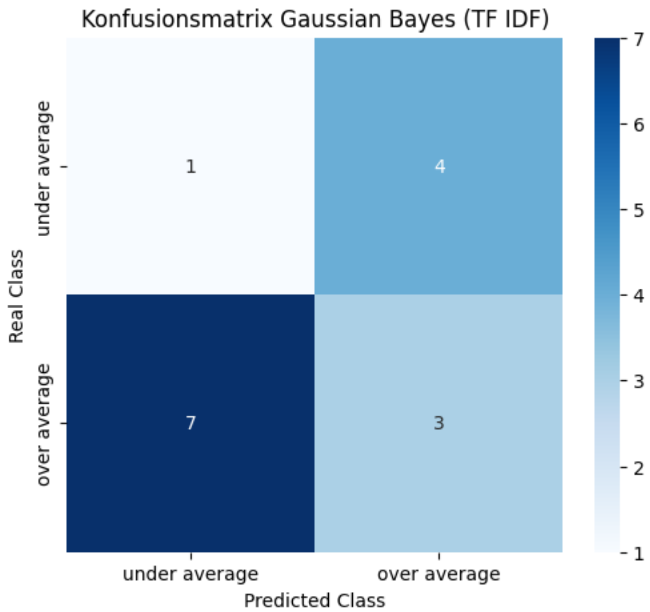


Figure A.6: Confusion Matrix Bayes Classification with TF IDF

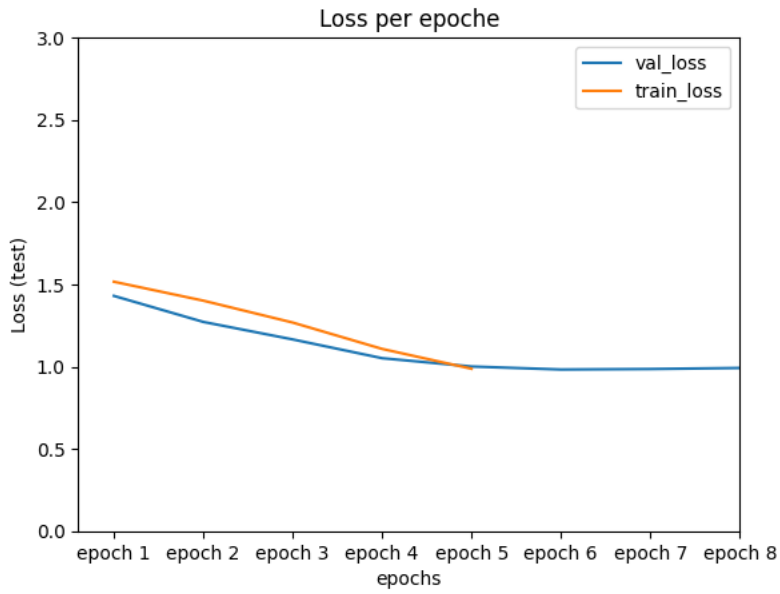


Figure A.7: Cross Validation Speaker Detection DistillBERT

References

Albanese, Nicolo (Sept. 2022). "Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison". In: *Towards Datascience*. URL: https://medium.com/towards-data-science/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5.

Grootendorst, Maarten (Mar. 2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794 [cs]. DOI: 10.48550/arXiv.2203.05794. URL: http://arxiv.org/abs/2203.05794 (visited on 03/06/2025).

Huggingface (n.d.). *DistillBERT Documentation on Huggingface*. englisch. URL: https://huggingface.co/docs/transformers/model_doc/distilbert.

IMDB, company (Jan. 2025). *Rick and Morty Episodenguide IMDB*. URL: https://www.imdb.com/de/title/tt2861424/?ref_=fn_all_ttl_1.

Mikolov, Tomas et al. (Sept. 2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs]. DOI: 10.48550/arXiv.1301.3781. URL: http://arxiv.org/abs/1301.3781 (visited on 03/06/2025).

Mueller, Andreas (n.d.). *Word Cloud Documentation Python*. URL: http://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html.

Prarabdha, Srivastava (2025). *Rick&Morty Transcripts*. URL: https://www.kaggle.com/datasets/prarabdhasrivastava/rickmorty-transcripts (visited on 01/31/2025).

Rick and Morty wiki, fandom (Jan. 2025). *Rick and Morty episode list*. URL: https://rickandmorty.fandom.com/wiki/List_of_episodes.

Sushmita, Katam (Dec. 2014). *The Porter Stemmer*. englisch. URL: http://cs.indstate.edu/~skatam/paper.pdf (visited on 03/03/2025).

Selbständigkeitserklärung
Anton Geiger

Ich versichere hiermit, dass ich die vorliegende Seminararbeit mit dem Thema

Burp... NLP! A multidimensional analysis through the characters of Rick and Morty

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum

Unterschrift

Selbständigkeitserklärung
Cedric Fechner

Ich versichere hiermit, dass ich die vorliegende Seminararbeit mit dem Thema

Burp... NLP! A multidimensional analysis through the characters of Rick and Morty

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum	Unterschrift

AI Usage

KI basiertes Hilfsmittel	Einsatzform	Betroffene Teile der Arbeit
ChatGPT 4.o	We used Chat GPT for code generation	All Chapters
DeepL	Translating phrases from german to english	All chapters
ChatGPT, Deepseek, Deepl	style and tonality of the work	All chapters