In Quest of the Main Character - Natural Language Processing and the Film Scripts of *The Lord of the Rings* Trilogy

by Maximilian Michel and Marina Lehmann

1. Introduction - What Makes a Main Character?

> The Ring remains hidden. And that we should seek to destroy it has not yet entered their darkest dreams. And so the weapon of the enemy is moving towards Mordor in the hands of a Hobbit. Each day brings it closer to the fires of Mount Doom. We must trust now in Frodo. Everything depends upon speed and the secrecy of his quest. Do not regret your decision to leave him. Frodo must finish this task alone. *(Gandalf in Jackson, The Two Towers Ext. Version, 01:01:58 - 01:02:30)*

With these words - directed at Aragorn in the second part of the film trilogy *The Lord of the Rings* - Gandalf expresses the importance of Frodo's role in their venture to defeat Sauron and the power of his ring. No doubt, Frodo is central to the plot. But can we therefore automatically assume that the little hobbit is the main character of the film series? Aragorn in his reply to Gandalf already casts doubt on this assumption:

> He's not alone. Sam went with him. *(Aragorn in Jackson, The Two Towers Ext. Version, 01:02:34 - 01:02:36)*

Sam is there as well, protecting Frodo from harm several times. And so are this friends of the fellowship and Gandalf who distract armies and fight bravely so that Frodo and Sam can reach Mordor without being discovered by Sauron. They receive at least as much screen time as Frodo and Sam.

This paper wants to approach the question "Who is the main character in *The Lord of the Rings?*" from a quantitative perspective by testing several means for "measuring" character significance: Statistics of word counts, character mentions and also sentiment analysis will be used to this end. These methods were chosen based on the following three hypotheses about character significance:

1. A character who speaks a lot is certainly important as he or she usually receives a lot of screen time. The significance of a character could therefore be measured by the number of words that the character utters over the course of the action.

2. A character who is mentioned a lot by other characters is also important, because other characters take interest in his or her actions. The significance of a character could therefore be measured by how many times this character's name is mentioned. The number of mentions in the stage directions is considered as well because if the stage directions refer to a certain character many times, it suggests that this character acts much and therefore has a lot of screen time.

3. The emotional development of a character could also be an indicator for how central a character is. Important characters undergo more changes during the action, they might experience more ups and downs and their emotional state at the end of the story might be significantly different from their state at the start. This is of course hard to measure. We try to use sentiment analysis to examine the emotional state of the characters. Based on the results from the statistical analysis (1. and 2.) we select the characters who are most likely among the central characters and analyse their dialogue text with

sentiment analysis techniques. On the one hand we examine their emotional profile over the course of all three films and on the other hand we compare their emotional state at the beginning to their state at the end.

For the analysis we used data from two sources: The stage directions were webscraped from the *Internet Movie Script Database* (IMSDb) (cf. Walsh, Boyens and Jackson, The Lord of the Rings. The Fellowship of the Ring; Jackson, Walsh and Boyens, The Lord of the Rings. The Two Towers, Jackson, Walsh and Boyens, The Lord of the Rings. The Return of the King), while the dialogue data were provided by the scholar and data scientist James Tauber (cf. Tauber and Palladino, Digital Tolkien Project). It should be noted that Tauber's data contain the film lines from the film's extended version, whereas IMSDb only offers the film scripts of the theatrical version.

In terms of data pre-processing we met various challenges of different kinds which will be explained in detail in the following chapter. The dialogue data from James Tauber we could more or less directly use for our analysis. The few adaptions we had to make could be executed directly in Excel. The webscraped film scripts needed more attention since the data had to be cleaned first on the character level (e.g remove erroneous characters) and on the word level (e.g. remove superfluous information like page numbers or camera instructions).

Furthermore, the scripts are not consistently formatted. Major differences exist between the second film script and the other two. Minor differences can also be found between the first and third script. Apart from data cleaning the challenge in pre-processing was to separate dialogue text from stage directions and save the latter to a csv file (see chapters 2.1 and 2.2). In the second film script stage directions are placed in square brackets. In the other two films scripts there is no such distinction. The only typographical characteristic of the stage directions is their lack of indentation: Stage directions are left-aligned whereas dialogue text is center-justified. We therefore used the criterion *presence or absence of whitespace before the first word of a paragraph* for separating stage directions form dialogue text.

For both data pre-processing and analysis we worked with Python 3. More specifically the following Python libraries were used:

- **os** for file operations like opening and saving csv files
- **re** for working with regular expressions in the data cleaning process
- **numpy** and **pandas** for working with dataframes
- **bs4 (BeautifulSoup)** for webscraping the film scripts
- **spacy**, **TextBlob** and **spaCyTextBlob** for NLP analyses
- **wordcloud**, **matplotlib** and **seaborn** for plotting and visualization

2. Data Cleaning and Pre-Processing

2.1 Stage Directions of Script 1 and 3

The process of data cleaning and preparation for script 1 (*The Fellowship of the Ring*) and 3 (*The Return of the King*) overall followed a seven-step process:

1. The data was scraped from the webpages via **BeautifulSoup**.
2. The text data was split into blocks using two blank lines in a row (`\r\n\r\n`) as separator for the blocks.
3. It was established whether each block contained dialogue data (block starts with more than one whitespace character) or stage directions (block starts directly with text, no whitespace at the

beginning). Only the stage directions were then processed in the remaining steps.
4. Specific blocks containing irrelevant data were deleted. These were for example blocks starting with `CONTINUED` which mark page transitions and contain information like page numbers or information about the script revision, but also blocks starting with `EXT.` or `INT.` which contain information about the place where the scene is set.
5. The data was cleaned.
6. The data was written to csv.
7. The stage directions were extracted from the csv file in order to obtain a coherent text.

Going through these steps we met several challenges. We discovered for example that not all stage directions were placed in a separate block. We also found instances in which stage directions directly followed a speaker's text.

> **BILBO** I know. He'd probably come with me if I asked him. I think, in his heart, Frodo's still in love with the Shire, the woods and the fields... little rivers. Bilbo stands gazing out of the kitchen window. *(Script 1, The Fellowship of the Ring)*

These special stage directions could not be extracted automatically from the script since there was no rule-based way to determine where in each case the dialogue text ends and the stage direction starts. The only solution would have been to read through both scripts and identify all such cases. We considered making a list containing those directions which could then be added during the data-preparation process. But considering the great amount of time needed for such an undertaking, we decided to omit these cases altogether. Therefore, in the following analyses not all stage directions occurring in the script were included, but only those ones which could be identified automatically through our python script.

A second problem that could not be solved entirely was the occurrence of random whitespace between words. Since identifying all the cases would have also involved a lot of manual work, we decided to only replace the known cases within character names like in `A RW EN` or `SME AGO L`.

Camera instructions (e.g. `LOW ANGLE`, `PAN OFF` or `SLOW MOTION`) formed another challenge because there are many of them and each can be spelled in different ways, e.g. `ANGLE ON` vs. `ANGLES ON` or `ON THE SOUNDTRACK` vs. `ON SOUNDTRACK`. We removed these instructions because we did not want them to distort our data analysis since words like "camera" or "angle" would probably appear among the most frequent words in an analysis but would not tell us anything about the plot or the characters. At that point in time, we did not have a clear concept of what analyses exactly we would conduct with the stage directions. Retrospectively this cleaning step would not have been necessary for our use case since we only used the stage directions to identify how many times characters were mentioned. Nevertheless, it is an important point to consider for anyone interested in working with this data.

Ellipses (`...`) are frequently used in the scripts though mostly not in their typical use case of indicating that one or more words have been omitted. Ellipses can appear at the beginning, middle or end of a sentence. Sometimes they could be best replaced by a full stop, sometimes by a space character. We decided to delete ellipses if they occurred at the beginning of a sentence, to replace them with a full stop if they occurred at the end of a sentence and to replace them with a space character when they occurred in the middle of a sentence.

2.2 Stage Directions of Script 2

For the data cleaning of the script *The Two Towers* we needed a different approach, since there were some differences in the script's typography compared to the other two films. First of all, the stage directions were

placed between squared brackets, in-speech as well as between speech lines. We assigned stage directions that were completely delimited by two speech lines a “Regie” value in the equivalent “Sprecher/Regie” column cell of the to-be-created CSV file in order to be extracted more easily in a later step. In-speech stage directions were harder to handle since very often multiple stage directions were mixed within the actor's speech text (`[With a twinkle in his eye] I am Gandalf the White. [Aragorn grins] And I come back to you now at the turn of the tide.`). In these cases, we deleted the text surrounding the squared brackets with the help of regular expressions. Both the in-speech and between-speech stage direction were written to CSV and converted to a plain text file that did not contain dialogue data anymore.

A second problem which occurred only in the second film's script was the flawed encoding of special characters, for example in proper names (*Barad-dûr*, *Théoden*, *Lothlórien*), but also lots of apostrophes which are very common in the English language (*Sam's*, *you're*, *can't*). Since all of these occurrences were replaced by the same character (�), they had to be cleaned by grouping them into categories like *genitive with proper names* (*Sam's*, *Frodo's*, *Gandalf's*) or *grammar* (*you've*, *didn't*, *he'll*) and by replacing them with regular expressions. To avoid problems later on, we decided to standardize proper names by removing all special characters (*Smèagol -> Smeagol*). The remaining cases were hard to put into a certain pattern and were therefore replaced as a whole, string by string.

2.3 Dialogue Data by James Tauber

The dialogue data were provided by the scholar and computer linguist James Tauber (cf. Tauber and Palladino, Digital Tolkien Project). These data contain the film lines from the film's extended version and were not based on the print film script, but extracted from the film's subtitles. Therefore, also Elvish speech was part of the original data. For our analysis, we cleaned the data in a way that suited our analysis target best: We removed all content between squared and round brackets, HTML tags that indicated voice over or other emphasized text passages, any additional text next to the speaker names in the “SPEAKER” column, and the timestamp columns. Concerning the Elvish passages, we decided to use the English translation (which had been provided in an extra column) instead of the original Elvish text since some passages in Elvish could contain interesting content (e. g. the dialogues between Legolas and Aragorn). With this approach, of course, some in-line Elvish words with important meaning would be lost (e. g. the use of *mellon* in front of the entry of Moria in the first film). However, the benefit of it predominates by far.

3. Text Analysis

Our analysis is divided into three parts: First of all, we attempted a general characterization of the film's characters by analyzing word frequencies as well as their causes. Second, we used word counts and the number of mentions of the character's names to narrow down the circle of the possible main characters. And third, we performed a sentiment analysis for these characters' dialogue text in order to find out to which extent emotional or personal changes could be observed.

3.1 Characterization with Wordclouds

The first part of the analysis consists of calculating word frequency tables for every important characters and visualizing the outcome with the help of wordclouds. When observing the wordclouds, it should be noted that the visualizations on their own could be slightly misleading since they only visualize the relative frequency of a character's words and not the absolute one. Take the word frequency tables of Gandalf and Legolas: While the former's most used word "ring" has a count of 39, the latter's equivalent "aragorn" appears only eight times in the character's overall speech – and still the sizes resemble each other.

But despite the inherent bias, the procedure shows some interesting results: In many wordclouds some of the biggest (= most used) words are the names of other characters. Frodo's most used word, for example, is Sam and *vice versa*. The same applies to the second duo of hobbit friends, Merry and Pippin. Considering the central role Frodo plays in all three films it is not surprising that his name takes a prominent position in the speech of almost every other main (and important minor) character. In Sam's particular case it is interesting to see that "mr" is another central word since it is often combined with "frodo" – a phrase mirroring Sam's employment as Frodo's gardener which is kept for most of the time, even during their journey as pairs and the development of their friendship. Yet, a closer look at the use of "frodo" and "mr. frodo" over time could bring new findings. While it is somehow logical that characters often say the names of their companions throughout the film (Merry-Pippin, Frodo-Sam, Pippin-Gandalf), another reason to keep in mind for the omnipresence of the word "frodo" is his role as ring-bearer. Leading characters like Gandalf or Aragorn often say his name in his absence because they bear him and his well-being in mind – even though it cannot be said with certainty without additional effort.

Speaking of Gandalf, his vocabulary is what you could call the closest to overall plot line, since he is a major figure in all three films. Therefore, his vocabulary does not only mirror his role as driving force who constantly stresses the urgency of quick action, like "must" (28 times) or "now" (23), but also contain important key words related to the plot, like "ring" (39), "sauron" (20), "king" (16), "mordor" (12), and, of course, "frodo" (32).

Another interesting case is Gollum (or Smeagol), whose vocabulary differs from the other characters' in two dimensions. First, his most used words contain "precious", his famous term for the One Ring, and his own name (both "Smeagol" and "Gollum") respectively the matching personal pronoun ("us"). This shows his obsession with the One Ring and, more important for this analysis, his self-referentiality and split personality. Second, a strong change between the two personalities is observable: While Gollum's most used words are "precious" (23) and "us" (21), Smeagol's favourites are "master" (27) and "smeagol" (23), which corresponds with his (temporary) character change during the second film: On the one hand, he shows an increased servile behaviour towards Frodo by calling him "master", and on the other hand, he tries to support his rediscovered old character by addressing himself with his old name. Keeping the films in the back of the head, the term "gollum" is more a tick than an actual reference to his name anyway.

When creating the wordclouds we were surprised that even Sauron appears as a speaker in the data, although with an uncommunicative attitude, so we decided to include him in our analysis. Not surprisingly, "one" (9) and "ring" (9) are his most used words, but also "baggins" (5) and "aragorn" (3) are mentioned. To understand Sauron's role as speaking character, it has to be considered that his speech is never directly uttered but a voice over to underline the overall presence of the Dark Ruler. It mostly shows up in scenes where the ring is either in action or otherwise the centre of attention. For example, during the debates concerning the ring in Elrond's council a voice recites the famous lines of the ring's creation in the Dark Language ("One Ring to rule them all..."), and when Frodo offers the ring to Aragorn in the end of the first film, Sauron tries to tempt Isildur's heir to take it and whispers his name. Although this might be an interesting discovery (without the data extracted by the subtitles we would not have classified the voice over as the voice of Sauron himself), this case does not contribute too much to our analysis because of its small quantity and therefore lesser significance.

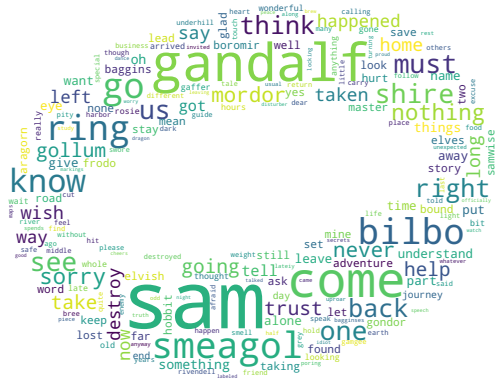


Fig. 1: Wordcloud Frodo

Fig. 2: Wordcloud Sam

Fig. 3: Wordcloud Gandalf

Fig. 4: Wordcloud Pippin

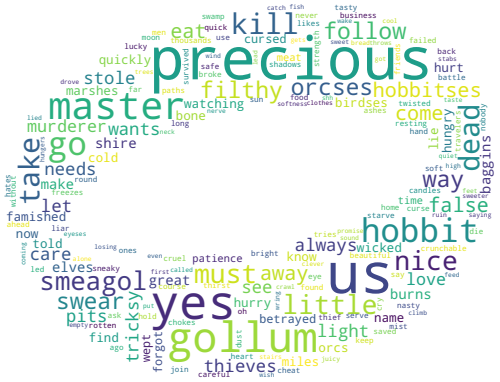


Fig. 5: Wordcloud Gollum

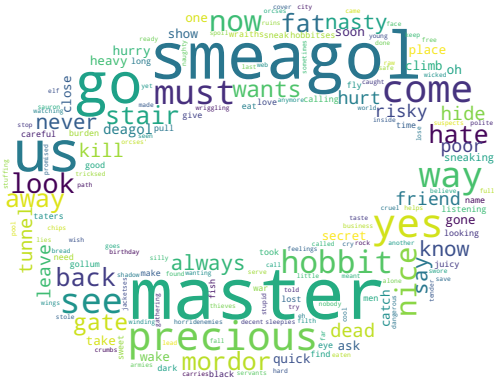


Fig. 6: Wordcloud Smeagol

Fig. 7: Wordcloud Sauron

	Word	Frequency
0	ring	39
1	frodo	32
2	now	28
3	must	23
4	one	23
5	saruman	21
6	sauron	20
7	king	16
8	bilbo	14
9	come	14
10	yes	14
11	lord	13
12	know	13
13	back	13
14	mordor	12

Tab. 1: Word frequency Gandalf

	Word	Frequency
0	aragorn	8
1	come	5
2	gondor	4
3	something	4
4	dead	4
5	must	3
6	one	3
7	may	3
8	need	3
9	men	3
10	heir	2
11	nothing	2
12	gandalf	2
13	shadow	2
14	near	2

Tab. 2: Word frequency Legolas

	Word	Frequency
0	one	9
1	ring	9
2	baggins	5
3	rule	5
4	find	4
5	death	3
6	aragorn	3
7	see	2
8	mordor	2
9	elessar	2
10	hide	1
11	life	1
12	theid	1
13	build	1
14	army	1

Tab. 3: Word frequency Sauron

	Word	Frequency
0	precious	23
1	us	21
2	yes	14
3	gollum	13
4	master	11
5	go	8
6	kill	6
7	dead	6
8	little	5
9	nice	5
10	must	5
11	smeagol	5
12	filthy	4
13	take	4
14	orcses	4

Tab. 4: Word frequency Gollum

	Word	Frequency
0	master	27
1	smeagol	23
2	go	17
3	us	16
4	precious	15
5	yes	15
6	come	14
7	see	11
8	must	11
9	way	10
10	now	8
11	nice	8
12	mordor	6
13	back	6
14	away	6

Tab. 5: Word frequency Smeagol

Since most of the character names are quite dominant among the overall speech, it seems appropriate to not only use word counts by character but also the number of character mentions as benchmark for the question who might be the main character of the films. For the further analysis, we decided not to restrict ourselves to the dialogue data but also include the stage directions. While dialogue data illustrate who is *spoken about* or addressed directly, the stage directions focus on who is *acting* and therefore has a lot of screen time – an important factor for the analysis of a medium that has a high amount of nonverbal communication. These two approaches combined should give us a better basis for answering our research question.

3.2 Who is the Main Character in *The Lord of the Rings?*

3.2.1 Word Counts

When we look at the overall character word count, at first sight Gandalf seems to be the main character by far with 4862 spoken words. The next most talkactive characters are Sam (2557), Aragorn (2393) and Frodo (2320), after whom is another major gap in the ranking of word counts. At the very end and way behind all other characters is Sauron, whose occasional voice overs which we covered in the previous chapters cannot compete with the rest. Since Gollum and Smeagol are de facto one acting character, their spoken words sum up to 1820, so he can be counted to the leading group we could call the "Top Five" of our main character candidates. It is interesting to see that, despite his role as driving force in the second and third film, the lion's share of Gandalf's speech is uttered in the first film, nearly as much as Sam's overall word count. Sam, on the contrary, has his "communicative focus" in the second and third film, where he is not part of the fellowship full

of talkative Gandalfs and Aragorns anymore but the sole companion of Frodo (temporary alongside Gollum/Smeagol). Anyway, the second and third film are a fountain of word counts for formerly "silent" or non-existent characters, take again Gollum/Smeagol, Theoden (1334) and Eowyn (640).

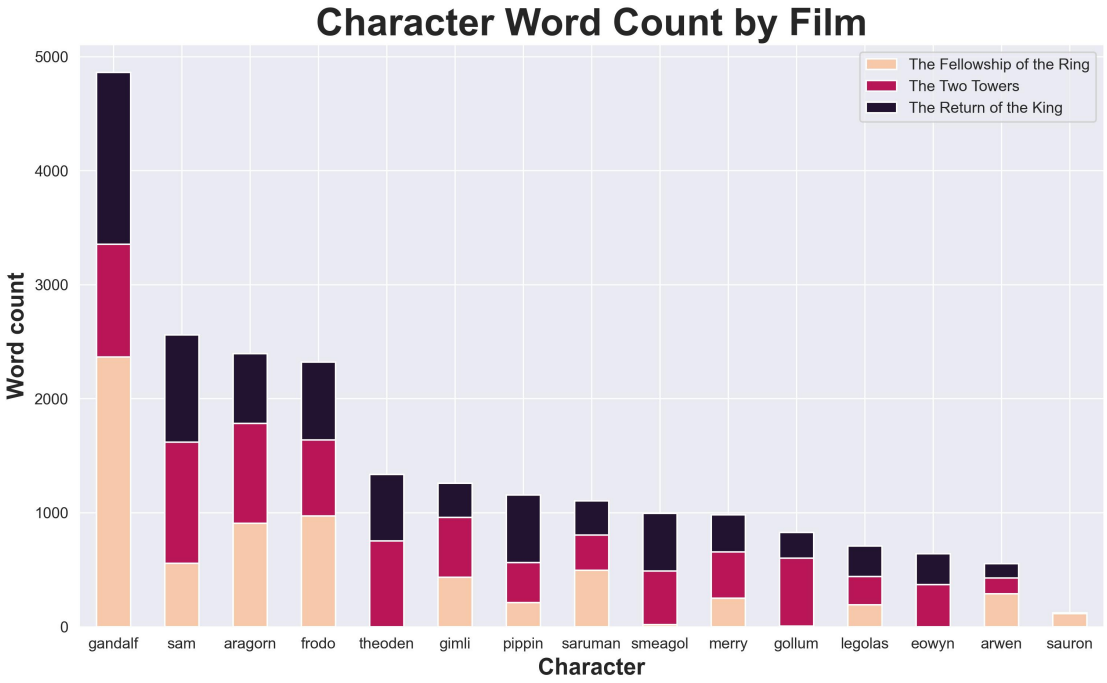


Fig. 8: Character Word Count by Film

	Character	The Fellowship of the Ring	The Two Towers	The Return of the King	Total
0	gandalf	2366	989	1507	4862
1	sam	554	1064	939	2557
2	aragorn	906	876	611	2393
3	frodo	972	664	684	2320
4	theoden	0	752	582	1334
5	gimli	434	524	300	1258
6	pippin	212	351	591	1154
7	saruman	493	310	301	1104
8	smeagol	19	469	505	993
9	merry	251	404	325	980
10	gollum	4	595	228	827
11	legolas	190	249	269	708
12	eowyn	0	370	270	640
13	arwen	289	139	123	551
14	sauron	114	1	5	120

Tab. 6: Character Word Count by Film

The mentions bar plot distribution also seems to have a clear favourite in counts, but this time it is Frodo and not Gandalf (802 mentions overall), followed with some distance by Sam (439). This is not surprising considering Frodo's presence in the overall dialogue, in addition to the previously mentioned role of nonverbal elements in films: The venture of Sam and Frodo in the second and third film takes a lot of screentime which is full of direct addresses as well as stage directions describing the circumstances since there are no other characters who could step into the breach when the conversation flow calms down. This is certainly the reason for the way higher number of mentions in the stage directions than in direct speech, since names and objects are the essential for specifying what is going on.

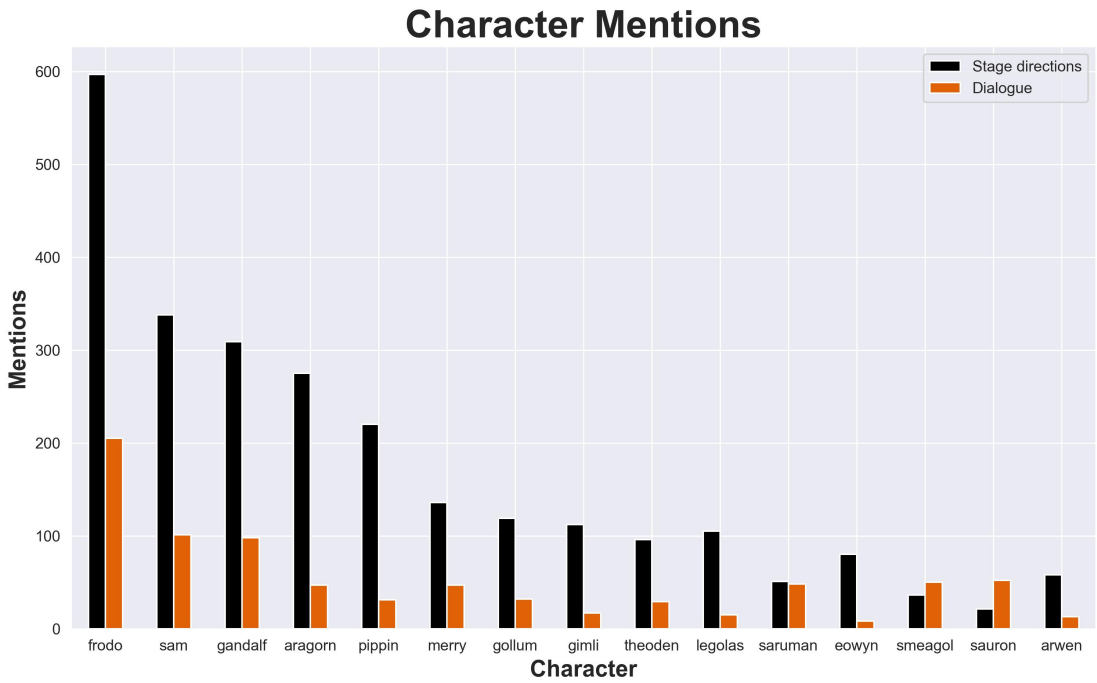


Fig. 9: Character mentions

	Character	Mentions_Total	Mentions_Dialogue	Mentions_Stage_directions	Ratio
0	frodo	802	205	597	2.9
1	sam	439	101	338	3.3
2	gandalf	407	98	309	3.2
3	aragorn	322	47	275	5.9
4	pippin	251	31	220	7.1
5	merry	183	47	136	2.9
6	gollum	151	32	119	3.7
7	gimli	129	17	112	6.6
8	theoden	125	29	96	3.3
9	legolas	120	15	105	7.0
10	saruman	99	48	51	1.1
11	eowyn	88	8	80	10.0
12	smeagol	86	50	36	0.7
13	sauron	73	52	21	0.4
14	arwen	71	13	58	4.5

Tab. 7: Character mentions in dialogue and stage directions

Not necessarily an essential part for our analysis, but yet informative, is the ratio between the mentions in dialogue and stage directions, which is not clearly depicted in the above diagram. Therefore we created a ratio based on the formula `Ratio = mentions stage directions / mentions dialogue`. The higher the ratio, the more a person tends to be mentioned in the stage directions compared to dialogue (Eowyn's name, for example, with a ratio of 10 appears ten times more often in the stage directions than in dialogue). Nearly all of the characters have a ratio higher than 3, while only two characters are below 1 (= mentioned more often in the dialogue): Smeagol and Sauron. Smeagol's ratio might be caused by the facts that the stage directions tend to describe his actions as those of "Gollum", and Frodo almost exclusively calls him "Smeagol" thus supporting him with character change difficulties. Sauron, however, again lives up to his special role in our character squad. His rare mentions in the stage directions cannot compensate that the other characters refer to him as dark menace in the background more often, especially Gandalf in his stories and reports about the ring. Also interesting to see are the high rankings of Gimli and especially Legolas, one of the least talkative characters. This might be the consequence of them taking part in many description-needy action scenes (Helm's Deep, Moria) and being an entertaining fighting duo in them, combined with a less important role in meetings and conversations (beyond just being present).

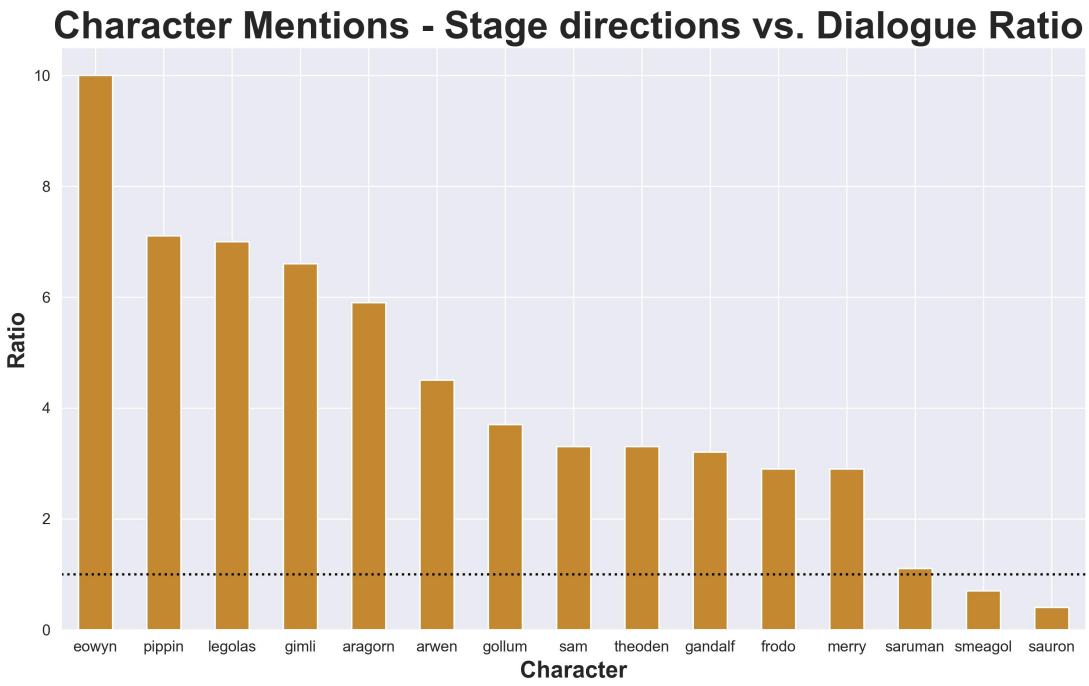


Fig. 10: Character mentions comparison between stage directions and dialogue

	Character	The Fellowship of the Ring	The Two Towers	The Return of the King	Word_Count_Total	Mentions_Total
0	frodo	972	664	684	2320	802
1	sam	554	1064	939	2557	439
2	gandalf	2366	989	1507	4862	407
3	aragorn	906	876	611	2393	322
4	pippin	212	351	591	1154	251
5	merry	251	404	325	980	183
6	gollum	4	595	228	827	151
7	gimli	434	524	300	1258	129
8	theoden	0	752	582	1334	125
9	legolas	190	249	269	708	120
10	saruman	493	310	301	1104	99
11	eowyn	0	370	270	640	88
12	smeagol	19	469	505	993	86
13	sauron	114	1	5	120	73
14	arwen	289	139	123	551	71

Tab. 7: Character mentions count

3.2.2 Sentiment Analysis - Emotional Character Development

The statistical analyses showed that there is a top-five group of characters with high counts in spoken words as well as in mentions: Frodo, Gandalf, Sam, Aragorn and Gollum/Smeagol. The latter was included even though Gollum and Smeagol do not appear among the top five in the bar charts presented in the previous chapters. In the charts Gollum and Smeagol were counted seperately. However, taking both together their word count and number of mentions is higher, reaching fifth place in word counts and a very close sixth place after Pippin in mentions. Since Pippin only reaches seventh place in word counts - after Theoden and Gimli - he is not included in the top-five group selected for sentiment analysis.

We used `spaCyTextBlob` for the sentiment analysis. `SpaCyTextBlob` is a component based on the library `TextBlob` which can be added to the spacy nlp pipeline. It allows to determine polarity (whether a given text is positive or negative) and subjectivity (whether a given text is written subjectively or objectively). For our analysis only polarity is relevant. The polarity values can easily be accesed via the extension attributes `._.polarity` which returns a polarity score between -1.0 and 1.0 or `._.assessments` which returns a summary containing the words on which the sentiment analysis was based, the polarity score as well as the subjectivity score. (cf. Open Source Libs)

Overall, there are three approaches to sentiment analysis: 1) sentiment analysis based on lexicons i.e. dictionaries which contain fixed polarity values for a large number of words, 2) sentiment analysis based on machine learning which uses text classification algorithms trained on large datasets to determine polarity and 3) hybrid approaches which combine both techniques. (cf. D'Andrea et al., pp. 26-27)

By default `spaCyTextBlob` uses the `PatternAnalyzer` from the `pattern` library to classify words as positive or negative. `PatternAnanlyzer`uses a lexicon based approach. It is important to keep in mind that this method starts with the polarity of individual words and then sets out to calculate sentiments for sentences and whole texts based on this initial input. (cf. Kuzminykh)

Our first goal in sentiment analyis was to establish emotional profiles for each of the top-five characters. Our initial idea was to use line plots to capture each character's individual development. However, those line plots were hard to interpret since no clear patterns or phases could be identified. It resembled a seismographic record with alternating ups and downs in short intervals. For some characters - like Frodo - phases of mostly positive sentiments (e.g. from 25 to 50) and mostly negative sentiments (e.g. from 50 to 110) could be identified. But since this information was detatched from the words which had caused these sentiments as well as from the moment in the plot when the words were uttered, this kind of plot did not prove very useful. Neither did, of course, the line plot combining information from all characters. Each character had a different number of words to which a polarity had been assigned, therefore each plot has a different length which made them incomparable.

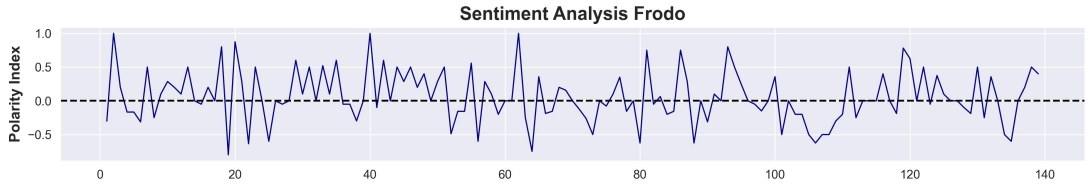


Fig. 11: Line Plot Frodo

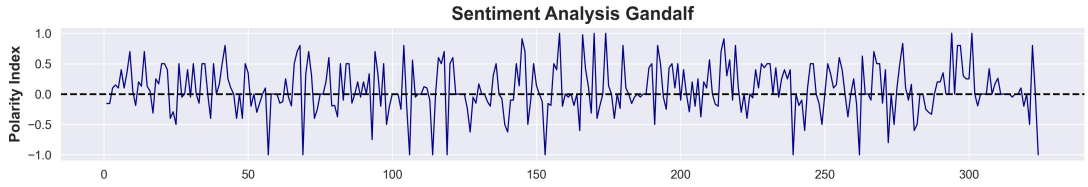


Fig. 12: Line Plot Gandalf

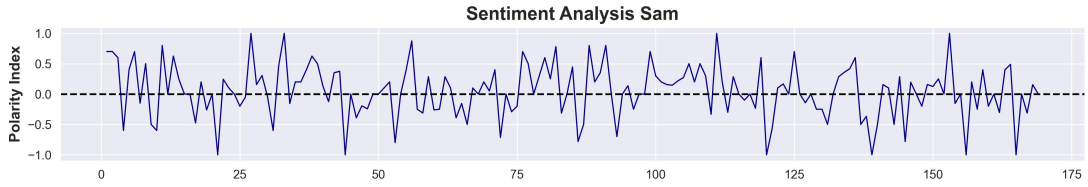


Fig. 13: Line Plot Sam

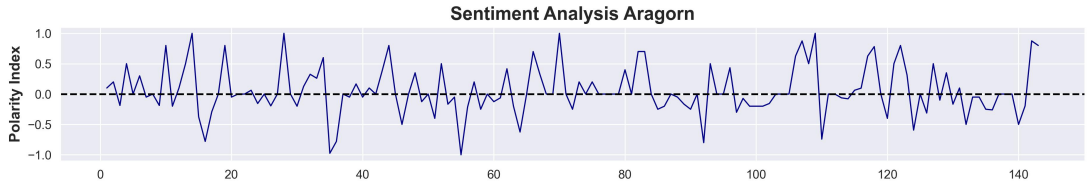


Fig. 14: Line Plot Aragorn

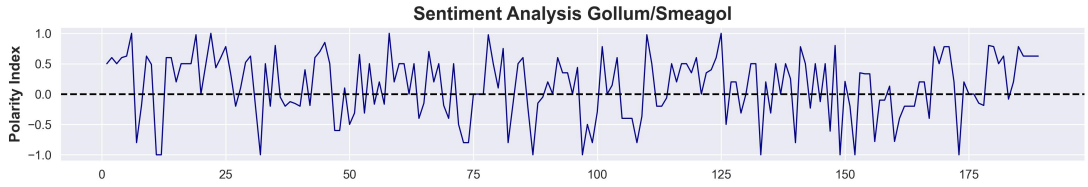


Fig. 15: Line Plot Gollum/Smeagol

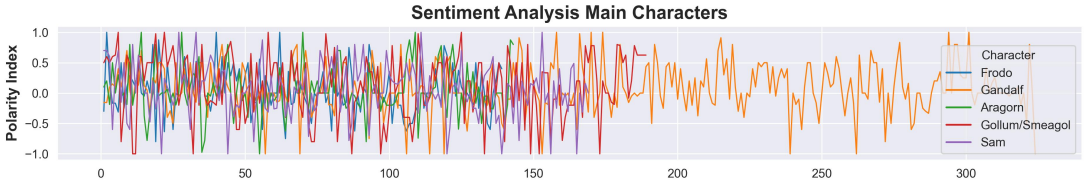
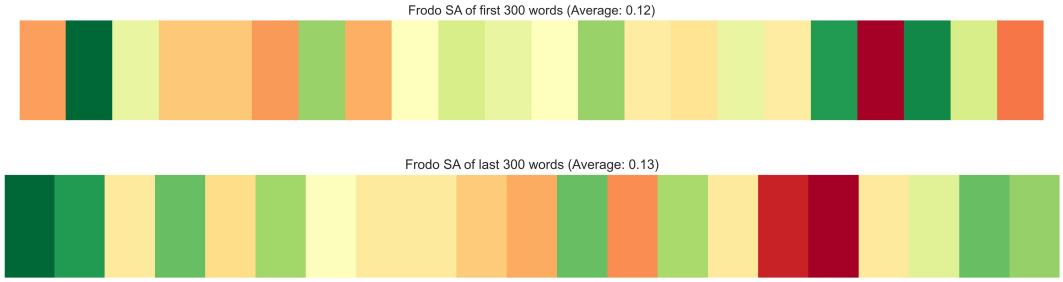
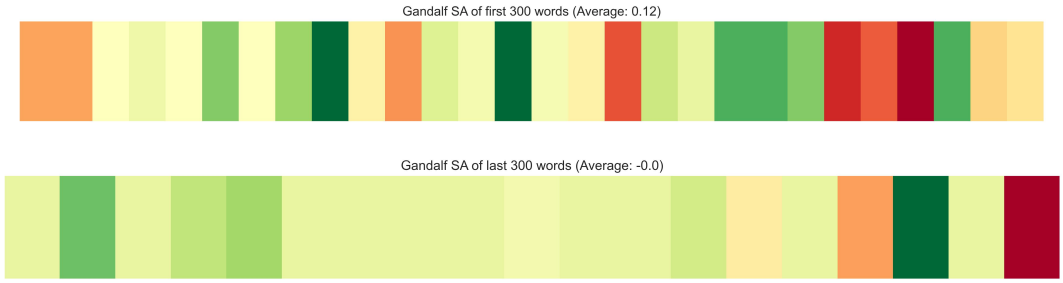


Fig. 16: Line Plot All Characters

Our second goal was to use sentiment analysis on the first and last 300 words of each character and then compare the results. Maybe a significant emotional change could be observed? We calculated the average polarity for the start and end word group and subtracted the difference. Additionally, we used barcode plots to encode polarities as either red (negative), yellow (neutral) or green (positive). The barcode plots were inspired by a student project that used NLP to compare the books and films of *The Lord of the Rings* (cf. Cannistrà).

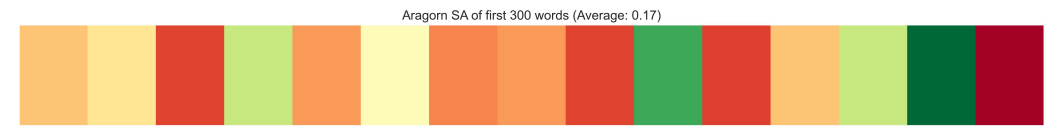


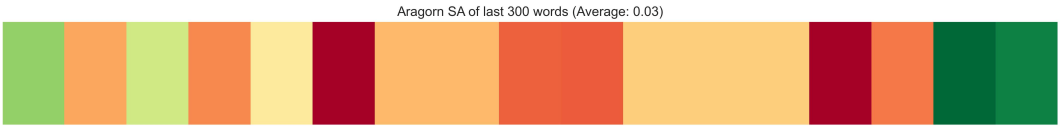
| | **Fig. 17**: Barcode Plots Frodo |



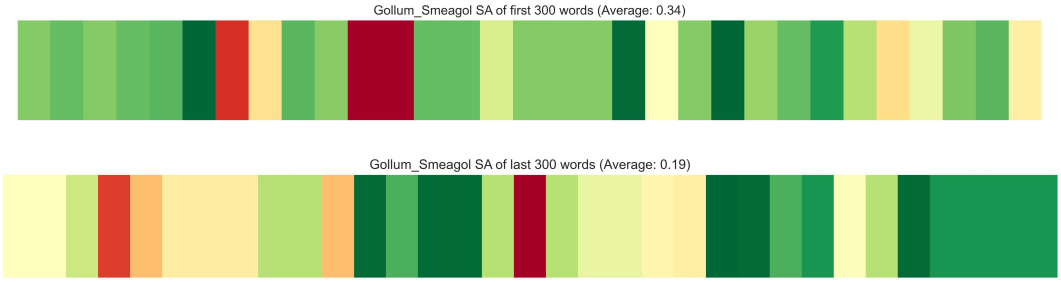
| | **Fig. 18**: Barcode Plots Gandalf |

| | **Fig. 19**: Barcode Plots Sam |





| | **Fig. 20**: Barcode Plots Aragorn |



| | **Fig. 21**: Barcode Plots Gollum/Smeagol |

However, no significant emotional change could be observed for any of the characters. Even though the barcode plots provided clearly distinguishable visual profiles for each text group, the polarities for almost all text groups (start and end) averaged out around zero. Therefore, also the differences between start and end average of each characters were very low. The results produced the following (not very informative) polarity ranking:

Rank	Character	Difference (average start - average end)
1	Sam	-0.2 (0.1 → -0.1)
2	Gollum/Smeagol	-0.15 (0.34 → 0.19)
3	Aragorn	-0.14 (0.17 → 0.03)
4	Gandalf	-0.12 (0.12 → 0.0)
5	Frodo	0.01 (0.12 → 0.13)

Tab. 8: Polarity Differences

The comparatively positive start average of Gollum/Smeagol of 0.34 seems interesting as well as the fact that Frodo appears last in this ranking with amost no difference between his start and end average. Apart from Frodo everyone apparently ended with slightly more negative feelings than at the start. But can those results actually tell us something meaningful? Why is there so little development since all values are centred around zero?

To find out why the results of the sentiment analysis are not informative, we investigated the numbers behind our visualizations and averages and also had a look at which words were actually incorporated in the sentiment analysis.

Frodo's first 300 words of text, for example, consist of 63 sentences. However, only 17 received a polarity other than 0.0. Looking at the word level, the reduction is even stronger: From 300 words only 20 have a polarity different from zero. Those words are:

> wonderful, whole, half, odd, mean, more, usual, old, right, very, unexpected, glad, long, really, proud, idiot, good, right and hidden

What was Frodo talking about? - Hard to tell. To compare, these are the first 300 words of text, the words with polarity are highlighted:

> you're late. it's **wonderful** to see you, gandalf! you know bilbo. he's got the **whole** piece in an uproar. **half** the shire's been invited. and the rest of them are turning up anyway. to tell you the truth, bilbo's been a bit **odd** lateiy. i mean, more than usual. he's taken to locking himself in his study. he spends hours and hours poring over old maps when he thinks i'm not looking. he's up to something. all **right**, then. keep your secrets. but i know you have something to do with it. before you came along, we bagginses were **very** well thought of. never had any adventures or did anything **unexpected**. whatever you did, you've been officially labeled a disturber of the peace. gandalf? i'm **glad** you're back. go on, sam. ask rosie for a dance. oh, no, you don't. go on! bilbo? bilbo, have you been at the gaffer's home brew? bilbo. bilbo, watch out for the dragon! speech! bilbo! bilbo! he's gone, hasn't he? he talked for so **long** about leaving i didn't think he'd **really** do it. gandalf? where are you going? what things? you've only just arrived. i don't understand. and **proud** of it. cheers, gaffer. don't worry, sam. rosie knows an **idiot** when she sees one. **good** night, sam. what are you doing? nothing. there's nothing. wait. there are markings. it's some form of elvish. i can't read it. mordor! bilbo found it. in gollum's cave. but he was destroyed. sauron was destroyed. all **right**. we put it away. we keep it **hidden**. we never speak of it again. no one knows it's here, do they? do they, gandalf? shire. baggins. but that would lead them here! take it, gandalf! take it! you must take it! i'm giving it to you!

Three problems can be identified here: 1) 20 words with polarity out of 300 is a very low score. 2) 20 words are far too few to calculate a reliable average. 3) Furthermore, the selected words do not characterize very well what Frodo says. In fact, from those 20 words the content of Frodo's speech could not be guessed.

These problems do not only concern Frodo, but all characters. The mean polarity for (the complete text of) each character is close to zero: Frodo: 0.06, Gandalf: 0.08, Sam: 0.05, Aragorn: 0.06, Gollum/Smeagol: 0.13. Looking at the polarity distributions of each charachter makes clear why this is the case: For almost all characters, most words have a polarity around zero, whereas few words with stronger polarities (higher/lower than 0.5/-0.5) can be identified. Only with Gollum/Smeagol the polarities are more evenly distributed. Most words have a polarity of 0.5. -1 and 1 also appear more frequently than with the other characters.

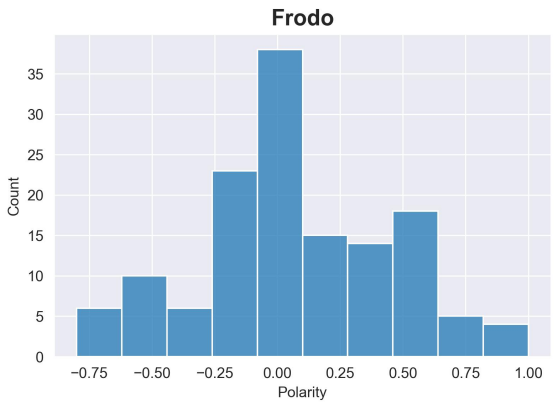


Fig. 22: Polarity Distribution Frodo

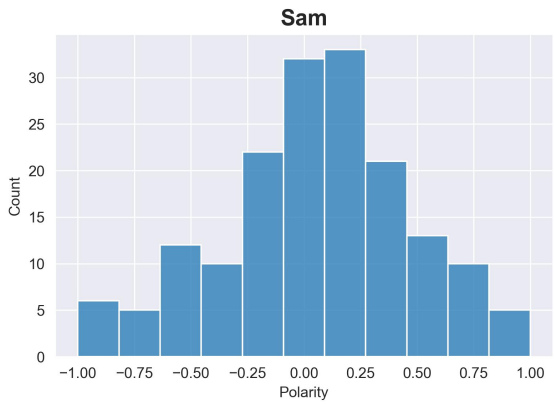


Fig. 23: Polarity Distribution Sam

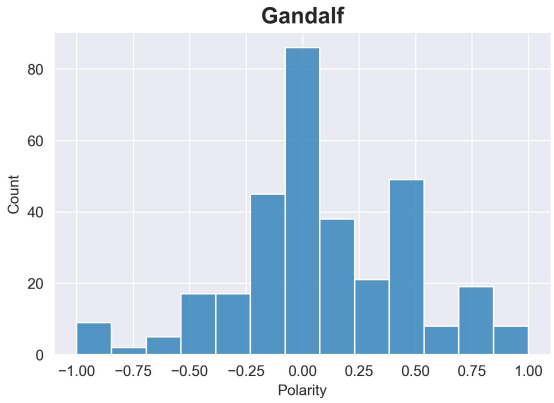


Fig. 24: Polarity Distribution Gandalf

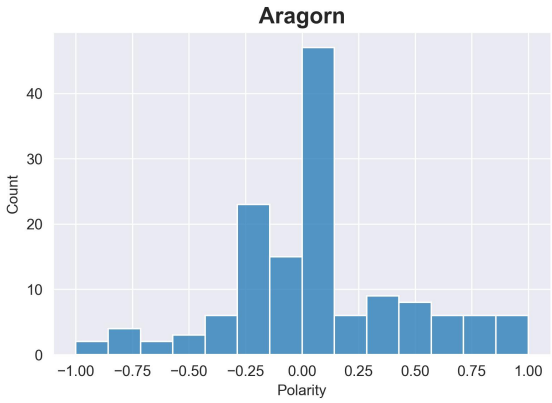


Fig. 25: Polarity Distribution Aragorn

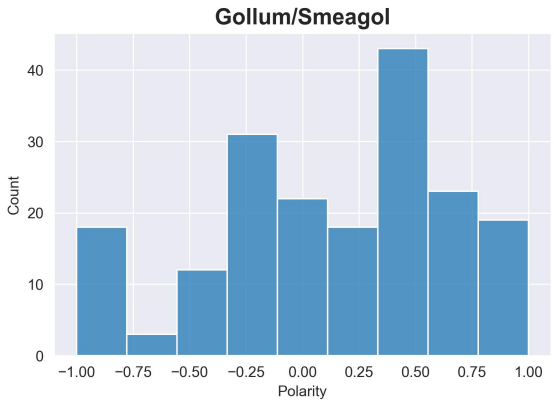


Fig. 26: Polarity Distributino Gollum/Smeagol

A wordcloud of Gollum's words which have been considered for sentiment analysis sheds light on this phenomenon:

Fig. 27: Sentiment Analysis Wordcloud Gollum/Smeagol

Gollum very often says "my precious" and "my love" (adressing either himself or the ring) which both have a positive polarity. Apart from that he uses "very nice" as filler words and also often talks about who "wants" the ring which both also have a positive polarity. This explains why Gollum's results are more positive than those of the other characters, but we cannot conclude that Gollum is a more positive creature since the words are considered out of context and can therefore mislead.

Fig. 28: Sentiment Analysis Wordcloud Frodo

Fig. 29: Sentiment Analysis Wordcloud Sam

Fig. 30: Sentiment Analysis Wordcloud Gandalf

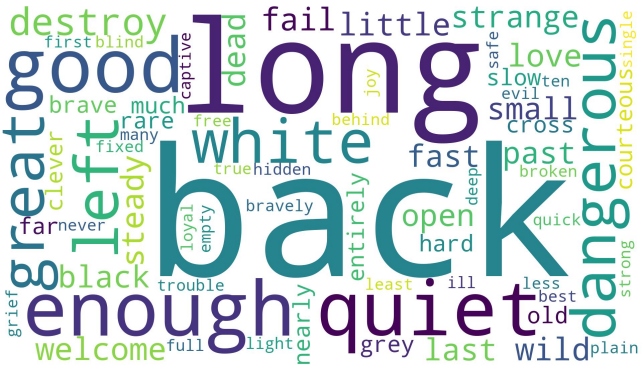


Fig. 31: Sentiment Analysis Wordcloud Aragorn

The equivalent wordclouds of the other characters strengthen the hypothesis that the words found in the sentiment dictionary are not representative of what the characters actually say. Among the most common words are always rather generic adjectives:

Character	Most frequent sentiment analysis words
Aragorn	back, long, enough, good, quiet, left
Frodo	right, back, sorry, left, long, destroy
Gandalf	old, back, dark, long, many, deep, good, first, safe
Sam	back, mean, right, good, far, dead, quiet, much, really

Tab. 9: Most frequent sentiment analysis words

We can therefore identify several problems with this sentiment analysis:

1. Many words from the dialog texts are not in the dictionary. Especially verbs are missing. Based on the words used for sentiment analyis it is hard to determine what the text is about.
2. There is a high number of zero-rated words which are not very helpful for the analysis.
3. The words that are in the dictionary and can therefore be considered for sentiment analysis are mostly adjectives.
4. However, adjectives do not appear very often in the dialogue text and if they appear, they do not convey the meaning of the text.

Since sentiment analysis with the `PatternAnalyzer` from `TextBlob` could not provide satisfactory results, we tried two more approaches based on machine learning.

`TextBlob` incorporates a second Analyzer, the `NaiveBayesAnalyzer` which is based on "an NLTK model trained on a film reviews corpus" (cf. Kuzminykh). Its output is different from the output of the `PatternAnalyzer`: It returns for each input (document, sentence or word) a pos-value, a neg-value and an overall classification which is either positive or negative. The advatage is that each word can receive a classification as either positive, there are no polarities of zero/neutral. However, there is a tendency to favour a positive classification because we noticed that if the pos-value and the neg-value are equal, the classification results in "positive". Moreover, the process is very slow, it takes around 2 seconds for each computation. It was consequently impossible to test all data with this Analyzer. We thus tested the `NaiveBayesAnalyzer` only for Frodo's first 300 words. Looking at the output of the analysis per sentcence, it seems that this model got many of the unamibigous sentences right.

Correctly classified as positive:

- it's wonderful to see you, gandalf
- half the shire's been invited
- i'm glad you're back
- before you came along, we bagginses were very well thought of
- never had any adventures or did anything unexpected
- ask rosie for a dance
- and proud of it
- do not worry, sam

However there are actually many sentences, for which even humans would be in doubt whether to classify them as positive or negative. For example:

- he spends hours and hours poring over old maps when he thinks i'm not looking
- he's up to something
- but i know you have something to do with it
- he talked for so long about leaving i did not think he'd really do it
- where are you going
- what are you doing
- it's some form of elvish

This indicates that having a classification "neutral" would make sense after all. Two sentences were especially interesting: "but he was destroyed. sauron was destroyed". They were classified as positive even though destruction is usually a bad thing. Since the model was based on a film review corpus, there were probably

also reviews about *The Lord of the Rings*. Did the model learn that Sauron was evil and his destruction therefor a positive event? We cannot know for sure, but it is an interesting hypothesis.

The second alternative was Stanford CoreNLP. This model was also trained on a film review corpus and is based on a recursive neural network (cf. NLP Stanford, Information). It works with a sentiment treebank (cf. NLP Stanford, Sentiment Treebank) as a corpus, i.e. it can take the context within a sentence into account. The classification is more detailed than the NaiveBayes classification: The categories "very negative", "negative", "neutral", "positive" and "very positive" are distinguished.

We used the live demo (cf. NLP Stanford, Live Demo) to get a first impression of the results, again using the first 300 words of Frodo's dialogue. 66 sentences were identified, 34 with neutral polarity, 32 with polarties other than neutral. However, only two sentences received strong polarities: "it's wonderful to see you, gandalf" and "and proud of it".

For each sentence a sentiment tree is displayed which can be expanded to get more details about which words are associated to which sentiment. The visual output is clearly an advantage, it is easy to get a first impression of the analyisis: How many sentences are neutral? Do strong polarities appear often? Unfortunately, the sentiment calculations are not displayed. Polarity is only expressed with colors not with the calculated sentiment values. For a more detailed analysis with this tool, it would be necessary to download the NLPCore and use it for example within a Python environment (cf. Frei). This should definetly be considered for future work on this *Lord of the Rings* corpus.

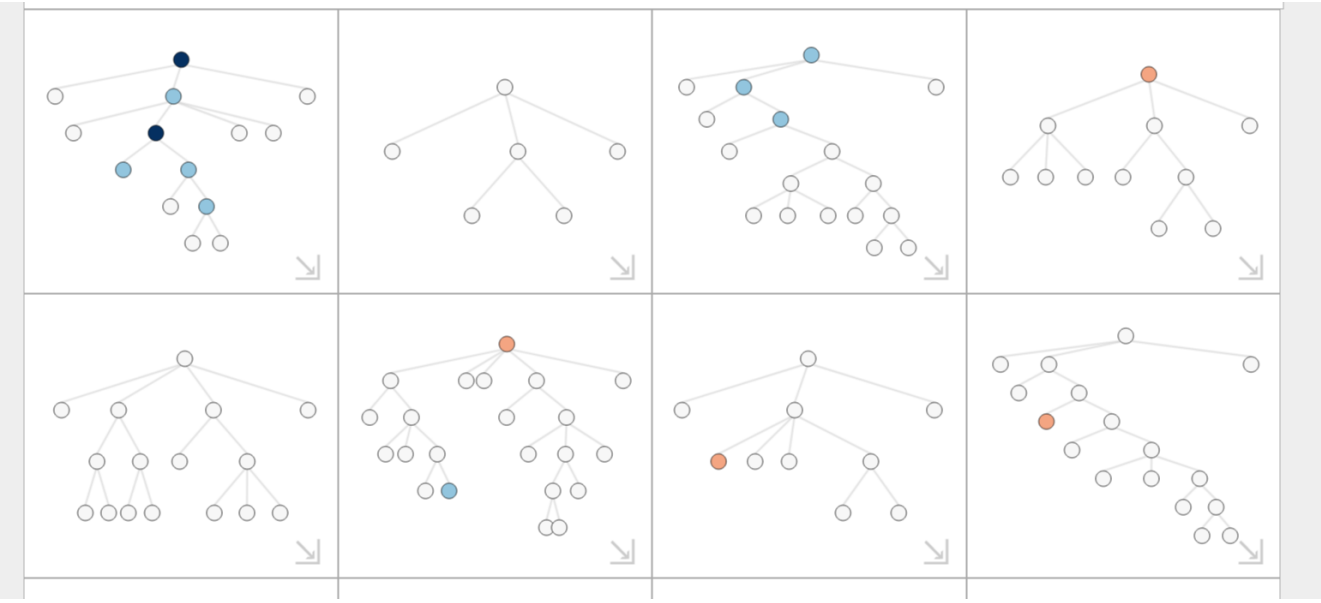


Fig. 32: Stanford CoreNLP - live demo output for sentences from Frodo's dialogue

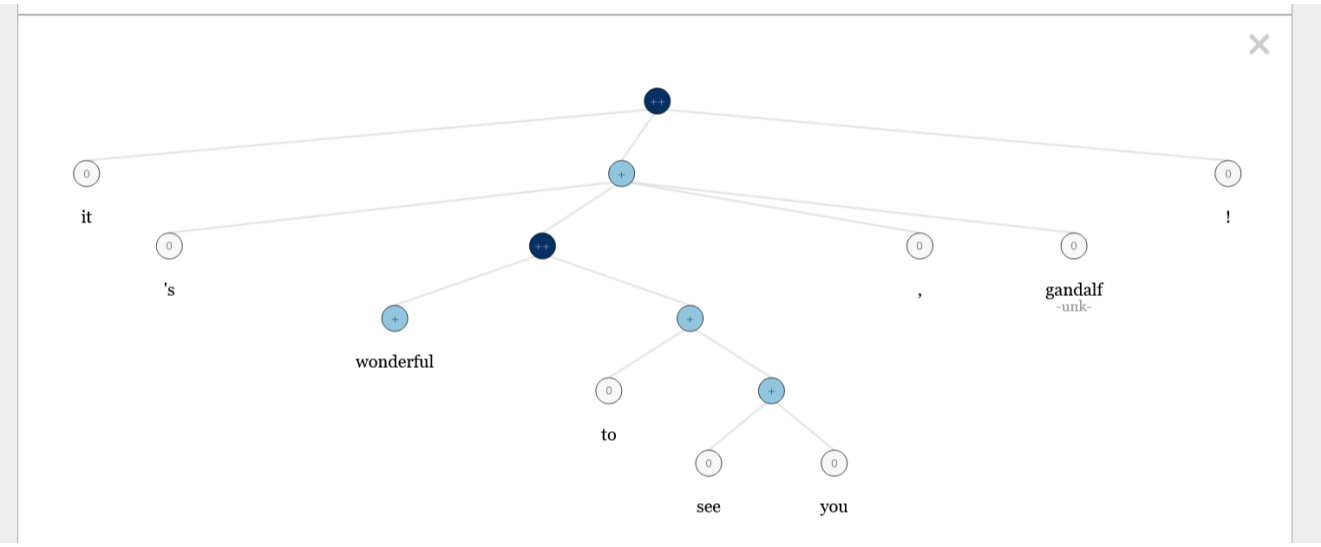


Fig. 33: Stanford CoreNLP - detailed view of one sentence

4. Conclusion

Through our analyses we definitely came closer to answering our research question *Who is the main character in Lord of the Rings?* The analyses based on word counts and mentions proved more useful to this end than sentiment analysis. Either Frodo - the most talked of character - or Gandalf - the most talkaktive character - could be considered as the main character. The results of our sentiment analysis experiment could not be used for answering the question, because it did not yield reliable results about the emotional development of the characters. However, it clearly showed the problems and limits of lexicon-based approaches to sentiment analysis. The quality of the results depends on the quality and scope of the dictionary. If the majority of words occurring in the text are not covered by the sentiment dictionary, the results of the analysis cannot be reliable.

Machine Learning based approaches could be more promising and should definetly be considered for further analyses since we could only test some of them on a very superficial level. However, their downside might be that more computing power is needed for those approaches and therefore more time for calculations.

5. References

Cannistrà, Riccardo (2016). The Lord of the Rings. In-depth Analysis. https://rickystream94.github.io/social_graphs/.

D'Andrea, Alessia et al. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. International Journal of Computer Application, 125(3), 26-32.

Frei, Lukas (2019). Natural Language Processing Using Stanford's CoreNLP: Analyzing Text Data in Just Two Lines of Code. https://towardsdatascience.com/natural-language-processing-using-stanfords-corenlp-d9e64c1e1024.

Kuzminykh, Natalia (2020). Sentiment Analysis in Python With TextBlob. https://stackabuse.com/sentiment-analysis-in-python-with-textblob/.

Jackson, Peter (2001). The Lord of the Rings: The Fellowship of the Ring. Extended Version. Jackson, Peter (2002). The Lord of the Rings: The Two Towers. Extended Version. Jackson, Peter (2003). The Lord of the Rings: The Return of the King. Extended Version.

Tauber, James and Palladino, Chiara: Digital Tolkien Project. https://digitaltolkien.com/.

Jackson, Peter; Walsh, Fran and Boyens, Philippa (2001). The Lord of the Rings. The Fellowship of the Ring. Film Script Draft. https://imsdb.com/scripts/Lord-of-the-Rings-Fellowship-of-the-Ring,-The.html

Jackson, Peter; Walsh, Fran and Boyens, Philippa (2002). The Lord of the Rings. The Two Towers. Film Script. Draft. https://imsdb.com/scripts/Lord-of-the-Rings-The-Two-Towers.html.

Jackson, Peter; Walsh, Fran and Boyens, Philippa (2003). The Lord of the Rings. The Return of the King. Film Script Draft. https://imsdb.com/scripts/Lord-of-the-Rings-Return-of-the-King.html

NLP Stanford (2013). Information. https://nlp.stanford.edu/sentiment/index.html. NLP Stanford (2013). Sentiment Treebank. https://nlp.stanford.edu/sentiment/treebank.html. NLP Stanford (2012). Live Demo. http://nlp.stanford.edu:8080/sentiment/rntnDemo.html.

Open Source Libs. spaCyTextBlob. https://opensourcelibs.com/lib/spacytextblob.

6. Our GitHub Repository

The code for data cleaning, statistic analysis and sentiment analysis as well as the visualizations can be found in our GitHub repository: nlp-lotr.