Exploring the Harry Potter universe through Network Analysis and Natural Language Processing

Georgia Tsoukala[a, 1], **Ioannis Tselios**[a, 2], **and Konstantina Freri**[a, 3]

In modern science, integration and interdisciplinarity are the key to discovering solutions for several challenging tasks. The current paper focuses on two active research areas, Network Science and Natural Language Processing. The integration of Network Analysis and Natural Language Processing (NLP) stands at the forefront of contemporary machine learning research. We take an interest in a specific subfield of NLP, Sentiment Analysis, which has also long been a topic of interest in recent years. Based on worldwide popular novels about Harry Potter written by J. K. Rowling, this study provides an extensive investigation by building networks and using sentiment analysis. Our main question is: Is it possible to identify the nature of relationships between the characters, in other words, to define whether two characters are friends or foes, allies or enemies, by constructing a network of character interactions? We build the characters network assigning weights on the edges based on the sentiment value of the text connecting the characters. The results do not show any direct correlation between the nature of the characters relationships and the generated graph. However, we express our thoughts on further analysis, discussing about the limitations of this study and we anticipate our study to be the starting point of different approaches.

network analysis | natural language processing | sentiment analysis | word clouds

Significance Statement

We investigate the integration of Network Analysis, Natural Language Processing (NLP), and Sentiment Analysis, focusing on character interactions in the Harry Potter books. The significance lies in exploring the potential of these interdisciplinary approaches to unveil the nature of relationships between fictional characters. Our research question of whether or not we can identify the nature of the relationships between characters, allows for an in-depth study of the Harry Potter series and requires the combination of network and sentiment analysis. Our findings, although inconclusive, pave the way for future inquiries into the connection between narrative sentiment and character relationships in literature, contributing to the evolving landscape of computational literary studies.

Network Science has become a thriving field of study due to its remarkable ability to represent very intricate phenomena using very simple models (1). The popularity of Network Science lies in its effectiveness in capturing and representing the complexity of various systems through the lens of interconnected nodes and edges, providing valuable insights into the structure and dynamics of networks across diverse domains. At the same time, Natural Language Processing is experiencing rapid growth as its theories and methods are deployed in a variety of new language technologies (2). As machines strive to understand not only the syntactic structures of language but also the nuanced emotions embedded in textual content, Sentiment Analysis serves as a crucial component of NLP applications. By incorporating sentiment analysis into NLP models, researchers aim to uncover the emotional tone, attitudes, and subjective information conveyed in text. This has been especially used in computational literary studies where it can be used to infer the relationships between fictional characters. (3)

This study explores the Harry Potter universe through text analysis of the books. There are two main reasons why we chose to extract networks from fictional novels. One is that defining the relationships in the real world is difficult because relations in the real world are fuzzy and human's emotions are variable and complex. The second is that characters in movies and novels experience their emotions openly, without being reserved. Hence, it should be easier to analyze networks of fictional material. (4) It is also true that the relationships found in books are not too distinct from the ones in real life, allowing the use of clever text mining and natural language algorithms to catch the main features of the social network depicted in the novel. (5)

In this paper we combine the fields of Network Science and NLP by implementing sentiment analysis on the Harry Potter books and assigning the values as weights in the edges connecting the characters. We start by constructing our dataset of characters and then we research our main question. We propose a method of generating networks of characters from the books, that includes connecting characters with an edge every time they appear in the same sentence. In our network, we incorporate the sentiment values of the sentences that connect the characters and explore whether or not this is in line with the nature of the characters relationship. We perform further analysis on the Harry Potter universe by detecting

Author affiliations: [a]Technical University of Denmark

Author contributions:
Preprocessing the books: Author 3
Creating the dataset of characters: Author 2, Author 3
Creating the networks: Author 1, Author 2, Author 3
Analyzing the networks: Author 1
Sentiment Analysis: Author 3
Communities: Author 2
Word Clouds: Author 1
Analyzing Spells: Author 1

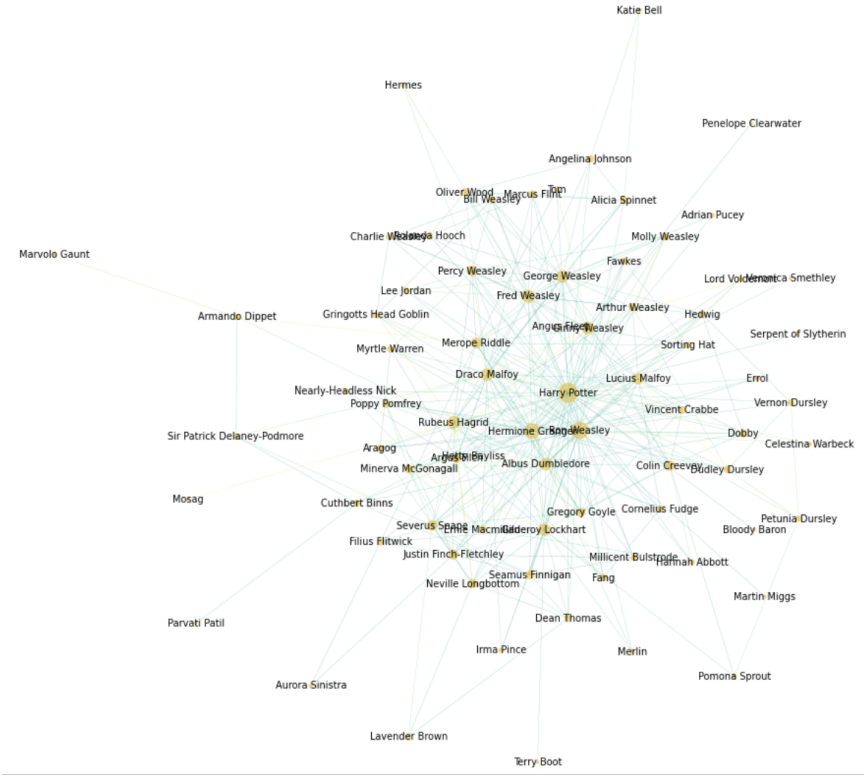

Fig. 1. The network of Harry Potter and the Chamber of Secrets. This network consists of 77 nodes and 342 edges. The size of the nodes denotes their degree and the color of the edges is associated with the sentiment value of the sentences that connect the characters.

communities, creating wordclouds and exploring the use of various spells.

In *Results* we describe all our process and findings in detail. We conclude this paper with a discussion of results and some possible directions for future work.

Results

Datasets. Our analysis was based on textual data extracted from the text versions of the Harry Potter book series. The primary dataset consists of the complete text of each book of the Harry Potter series, providing the foundation for character interaction and sentiment exploration. The total size of the books' dataset in ".txt" format amounted to 6.67 MB.

In order to be able to create our network there was the need to compile a list of characters from the Harry Potter universe. To achieve this, we employed web scraping techniques to extract information from the Harry Potter Fandom page. For each book, in Fandom, there is a page listing all characters in the order of their appearance. Through this process, we successfully compiled a dataset comprising a total of 717 characters from across the books. We expanded our dataset by scraping additional information for each character. This included details such as species, blood status, house affiliation, and loyalty, indicating whether a character belonged to Dumbledore's Army, the Death Eaters, or the Order of the Phoenix. Based on the name of the character, we also kept separate information for first name and last name, so that we will be able to find these characters in the book no matter how they are mentioned. After processing and cleaning the dataset our dataframe has 557 entries with 9 attributes.

Additionally, to dive deeper into the intricacies of the wizarding world, we aimed to analyze the progression of spell usage throughout the series. We sourced a supplementary dataset of spells through web scraping. This dataset has 178 entries and each entry includes essential information about the spells, such as the Incantation, Type, and Resulting Effect.

Network Analysis. By combining both our network knowledge and sentiment analysis we wanted to construct a network of the Harry Potter characters based on the interactions with each other and see if we can examine the nature of their relationships - if they are friends or enemies.

To do that we had to create a network for each of the books. Firstly, we had to accurately identify the characters within the text to create the nodes. To achieve this, we implemented a mapping function that associated first names or last names with the full names of characters, utilizing the NLTK library for named entity recognition and cross-referencing with our characters dataset. Then, we had to find out who interacted with who in order to create the edges. To do this, we consider every instance of two characters appearing together in the same sentence, as an interaction, resulting in the creation of edges connecting the characters in the network. Finally, we assigned sentiment values as weights to the edges based on the sentiment for the aggregated sentences of the characters' pairs.

This way, and with the help of NetworkX, we created 7 networks, one for every book. For the first three books, the networks exhibit relatively smaller sizes (nodes and edges wise), reflecting the fact that the story mainly takes place at Hogwarts, and there aren't many new characters introduced. But in the fourth book, "Goblet of Fire," things change. The Triwizard Tournament brings in a lot of new characters, making the network bigger. This shows how these characters are now connected and interacting more. As the series goes

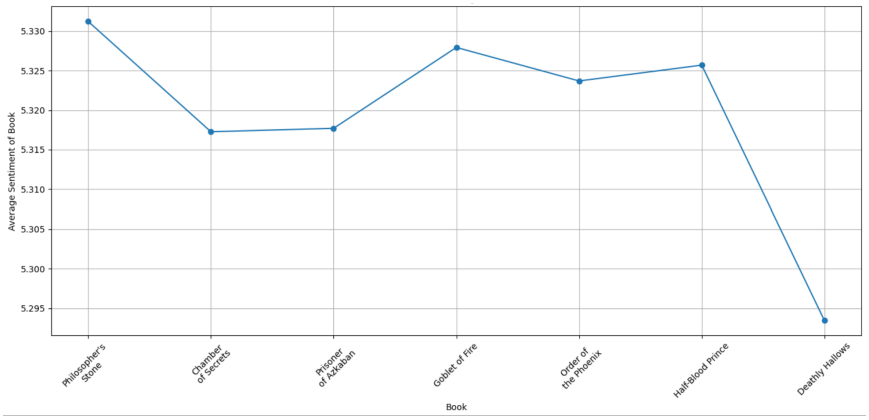

Fig. 2. Sentiment by book. This line plot illustrates the sentiment value of each book. The first book is the happiest one, while the last book has a really low sentiment value, associated with its darker theme.

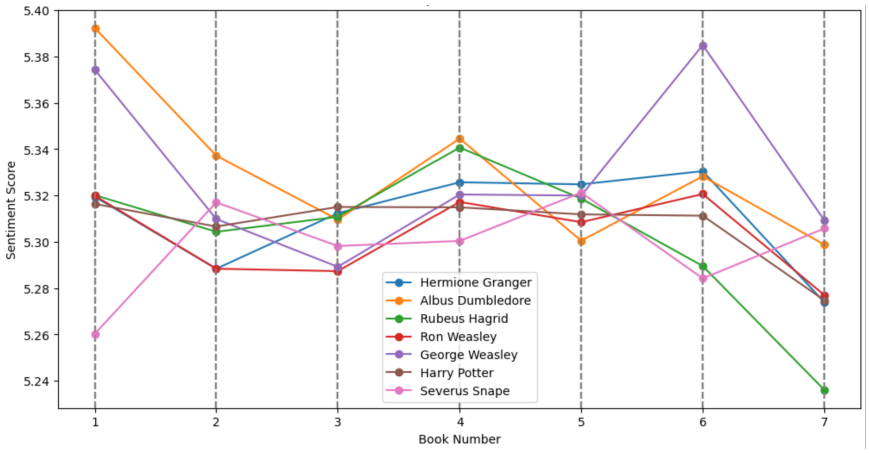

Fig. 3. Sentiment for each character across the books. The graph reveals a remarkably narrow range of values, ranging from 5.24 to 5.40. There is minimal fluctuation across all scores, while the sentiment of the main character, Harry Potter, exhibits the least variation.

on, the networks keep getting larger. The story goes beyond Hogwarts, bringing in lots of new characters like the members of the Order of the Phoenix and Death Eaters. This makes the networks more complex as the story expands outside the familiar school setting.

We also wanted to examine, through node degrees and centralities, if the protagonists of the books are indeed the most connected nodes or if our analysis could reveal something new/different. By calculating the degrees of the nodes and then looking at the characters that appear in the top 20 characters lists of all the books, we can spot 6 that appear in all of them. These are the 3 main friends and protagonists Harry Potter, Ron Weasley and Hermione Granger, Professor Albus Dumbledore as well as Professor Severus Snape who play a critical role in the series. Another one is Rubeus Hagrid, that helped the trio in their adventures. The presence of George Weasley can be considered a small surprise since he doesn't play a central role in a lot of the happenings throughout the books, but it seems that he interacted with a lot of the characters in all of the books.

Sentiment Analysis. In our quest to understand the relationships between the characters, after identifying the most central ones within the constructed networks, we proceeded to create some pairs consisting of some of these characters. This allowed for an in-depth examination of sentiment values derived from the weights of the edges connecting these pairs/nodes, aiming to discern patterns that could offer insights into the nature of their relationships (whether characterized by friendship or enmity).

By printing the sentiment values of the pairs in a table, as shown in Fig. 4, we observe that there is a lack of variation in sentiment scores as the values generally revolve around 5.3 and that main character pairs maintain consistent sentiment values across books, such as Harry Potter with his friends, Hermione and Ron. Also, the sentiment between Harry Potter and Voldemort is not negative besides the fact that they are the main adversaries of the series and the sentiment between Ron and Malfoy is pretty similar to that of Ron and Hermione which does not showcase that the former are rivals and the latter are friends. This means that, unfortunately, we are not able to draw any conclusions on the nature of the relationships between the characters.

We also tried to explore the evolution of sentiment of some main characters throughout the series by aggregating all sentences mentioning each character's name and subsequently calculating the sentiment scores for the compiled text. This analysis was visually represented through a line graph, as shown in Fig. 3, illustrating how sentiment scores fluctuated for each character across the series. Surprisingly, the findings indicate a remarkable consistency in sentiment scores, with values clustered within a tight range between 5.24 and 5.40 throughout the entire series.

To further explore the books, we decided to conduct sentiment analysis to find out about the progress of the sentiment throughout the Harry Potter series. The sentiment analysis graph of the books, as shown in Fig. 2, suggests that the Harry Potter series gets progressively darker, with the last book, "Deathly Hallows," having the lowest sentiment,

	Couple	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6	Book 7
0	Harry Potter - Hermione Granger	5.304617	5.290217	5.296287	5.309499	5.317433	5.321088	5.252946
1	Harry Potter - Ron Weasley	5.324008	5.293698	5.282727	5.313431	5.295629	5.310020	5.256729
2	Hermione Granger - Ron Weasley	5.286474	5.262610	5.272791	5.321687	5.295491	5.309458	5.243725
3	Draco Malfoy - Ron Weasley	5.318511	5.307553	5.328837	5.357805	5.384626	5.341586	5.127974
4	Harry Potter - Lord Voldemort	5.112972	5.259908	5.269448	5.218844	5.281680	5.260552	5.235836
5	Harry Potter - Rubeus Hagrid	5.300481	5.262121	5.267441	5.326789	5.300931	5.280681	5.217647
6	Albus Dumbledore - Minerva McGonagall	5.343368	5.550345	5.184651	5.383855	5.323475	5.496000	NaN
7	Albus Dumbledore - Severus Snape	5.365714	5.449130	5.300051	5.369363	5.304359	5.266156	5.292765

Fig. 4. Table of sentiment values for characters pairs. There is a lack of wide variation and the scores revolve around 5.3. Main character pairs maintain consistent sentiment values across books. The sentiment of rivals such as Harry and Voldemort or Ron and Malfoy does not differ from the sentiment of friends such as Harry and Hermione.

reflecting its heavy themes. The last book is the peak of the fight against Voldemort that leads to significant character deaths and a final battle. The unexpected lower sentiment in "Chamber of Secrets" and "Prisoner of Azkaban" compared to "Goblet of Fire" and "Order of the Phoenix" might be due to the complex emotions and conflicts in these books, which affect sentiment but not necessarily the overall darkness.

In order to understand why this is happening we decided to explore the sentiment per chapter so that we could better identify the events that influence the overall sentiment of each book. We created a graph for each book and noticed that the average sentiment levels are relatively stable. Some notable dips in sentiment correspond to important and grim events. For instance, the sharp drop in 'Goblet of Fire' aligns with Voldemort's return and Cedric Diggory's death. In 'Order of the Phoenix,' the sentiment drops when Sirius Black is killed by Bellatrix Lestrange. 'Half-Blood Prince' shows a significant decline close to the end, which coincides with the tragic death of Dumbledore at the hands of Snape. The last book, 'Deathly Hallows,' shows the lowest sentiment compared to its predecessors, indicating the series' intense and somber ending with its increased conflict and numerous losses.

Communities. To further explore the networks of the Harry Potter books and the relationships between characters, we used the Louvain community detection algorithm. Through our findings, we observed the consistent presence of a community in each book consisting of characters who are athletes or have some affiliation with Quidditch, which is a significant element in the series. Additionally, we note that in the first three books, most communities revolve around Hogwarts and we don't detect any communities with Voldemort's followers. This aligns with our expectations, as Voldemort's existence is not widely known during this period, although his name appears in some communities in references to the past. In the fourth and fifth books, the presence of Voldemort becomes more pronounced. Specifically, in the fourth book, there is a community associated with the inner circle of Voldemort. In the fifth book, a substantial community emerges, encompassing many significant figures and characters from the Ministry of Magic. It is logical for those books to highlight Voldemort's presence, as it marks his return.

Word Clouds. In order to gain more insights into the Harry Potter universe, we sought to create word clouds highlighting the most frequent and distinctive words for each book compared to the entire series. After preprocessing the books, we applied a TF-IDF vectorizer to capture the significance of words within the context of each book and then, individual word clouds were generated for each book, offering a visual representation of these terms.

Upon examination of the resulting word clouds, key characters, including professors and creatures that played pivotal roles in specific books as well as some objects, emerged as prominent features in the visualizations. The distinct presence of these characters or objects within the word clouds served as a visual testament to their significance within each book's narrative landscape.

Fig. 5. Word Cloud for Harry Potter and the Prisoner of Azkaban. The main words observed are: "lupin", referring to Remus Lupin, an old friend of James Potter and new Defence Against the Dark Arts teacher, "pettigrew" and "scabbers", referring to Peter Pettigrew, an old friend of James Potter, who is also Ron's rat Scabbers and "sirius", referring to Sirius Black, an old friend of James Potter, godfather of Harry and wrongly convicted for the murder of the Potters.

Spells. In our endeavor to analyze the evolving use of spells throughout the Harry Potter series and identify the most frequently employed ones, we used our dataset of all known Charms and Curses from the wizarding world. We counted the occurrences of each spell in the text of each book in order to find the 10 most used ones. Our investigation into the ten most frequently used spells revealed the existence of spells useful in everyday life but also curses and defensive spells. However, a compelling aspect of our analysis focused on the occurrence of curses. In contrast to the expectation that the number of curses cast would consistently rise as the plot darkened, our findings revealed a nuanced pattern. In the initial three books, there is a relatively low frequency of curses (3 in the first book, 1 in the second and 0 in the

third), however, a notable surge in curse usage is observed in the fourth book (26 curses). In the sixth and seventh book a high level of curse utilization is maintained (23 curses in both).

Discussion

In this paper, we observed minimal divergence in sentiment scores due to a multitude reasons. One reason for that could be the way we perform sentiment analysis as using a dataset of words associated to a sentiment score, misses out a lot of forms of speech or expressions found in books that could reveal different sentiments. Another reason might be that the books are mainly targeted towards teenagers and even thought they might describe some dark scenes, they do not use a lot negatively charged or harsh words. Finally, due to the way books are structured a lot of the sentences where the characters appear might not contain all the information about their relationship and the feelings to each other. This information might lay in dialogs that usually do not contain both names in one sentence, so we miss it. Also, it might be found in descriptive parts of the text or they are conveyed through third-party characters, resulting in loss of crucial information.

In order to solve the last problem that is mentioned, we considered focusing on the dialogs within the books, but the construction of an algorithm to extract all dialogues proved unfeasible. This is because, dialogs differ from, for example, movie scripts, as they do not always explicitly mention who said what. In order to solve the sentiment analysis issue, we could enhance the sentiment analysis methodology within the texts. Rather than drawing conclusions solely based on the words used, exploring more sophisticated sentiment analysis techniques could provide more nuanced insights.

Concerning the communities, we gained an overall sense of communities comprising both positive and negative characters, but the analysis has its limitations. In such an extensive text with numerous descriptions and dialogues, individuals may be linked merely because their names appear together in a passage without signifying a substantial connection. Also, numerous characters may connect coincidentally or be part of a side story and that influences the communities of other characters.

Future plans may include comparing and integrating results obtained from the books with those derived from the scripts of the movies, providing a comprehensive understanding of the Harry Potter universe across different mediums. Another idea could be to narrow down the number of characters analyzing their interactions first, before expanding to include a broader range of characters.

1. M Newman, *Networks*. (Oxford University Press), (2018) Google-Books-ID: YdZjDwAAQBAJ.
2. S Bird, Natural Language Processing with Python. (2018).
3. A Zehe, J Arns, L Hettinger, A Hotho, HarryMotions – Classifying Relationships in Harry Potter based on Emotion Analysis. (year?).
4. IEEE Xplore Full-Text PDF: (year?).

Materials and Methods

Dataset. We used BeautifulSoup for HTML parsing in order or extract the characters, their attributes and the spells from the web. All the characters and attributes are extracted from the Harry Potter Fandom page (6) and the spells from pojo (7).

Detecting the characters. We used the Natural Language Toolkit (NLTK), a library in Python, to perform Named Entity Recognition (NER) and part-of-speech tagging on a given text. We filtered the named entities using our list of characters and our name-mapping function, to identify mentioned characters in the book.

More specifically, we start by tokenizing the input text (in this case the book) into a list of words. We continue by performing part-of-speech tagging on the tokenized words, which assigns a grammatical category (like noun, verb, adjective, etc.) to each word in the text. Then we use the ne_chunk function for named entity recognition, to identify chunks that correspond to named entities, such as people, organizations, locations, etc.

At the same time, we create a function called name_mapping that maps the first and last name of characters to their full name. This way we are able to filter our named entities and identify the characters names. The name_mapping function, of course, results in mapping one last name to many full names, since, for example, Weasley family has a lot of members. The way we resolve this issue is to keep a variable with the full name of the most recently mentioned character with the same last name. So, if a last name is mentioned in the text, it will be mapped to the full name of the last mentioned character from the list produced by the name_mapping function.

Community detection. We used the Louvain community detection algorithm to compute the best partition for our networks. This algorithm aims to find a partition of the nodes into communities that maximizes modularity, a measure of the quality of the partition. The algorithm operates by iteratively optimizing the modularity through a two-phase process: a local optimization step where nodes are moved to improve modularity within their local neighborhood, and a global optimization step where small communities are aggregated into a single node, and the process is repeated.

Word clouds. To implement our word clouds we started by preprocessing the text: setting the text to lowercase, performing tokenization, removing stopwords and verbs. We then used the TfidfVectorizer from scikit-learn to convert a collection of processed books (text data) into a TF-IDF (Term Frequency-Inverse Document Frequency) matrix. We set the max_df parameter to 0.9 to ignore the terms that appear in more than 90% of the text. This helps in removing terms that are too frequent and may not be informative. We finally use the WordCloud library to visually represent the words based on their frequencies.

Sentiment analysis dataset. We used a dataset that contains happiness evaluations of over 10,000 individual words and based on this we calculated the average sentiment value of a text passage. The dataset, referred to as Data Set S1, was sourced from (8).

5. MC Waumans, T Nicodème, H Bersini, Topology Analysis of Social Networks Extracted from Literature. *PLOS ONE* **10**, e0126470 (2015) Publisher: Public Library of Science.
6. Harry Potter Wiki (year?).
7. Harry Potter Spell List - All Spells On One Page! - Pojo.com (year?).
8. PS Dodds, KD Harris, IM Kloumann, CA Bliss, CM Danforth, Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLOS ONE* **6**, e26752 (2011) Publisher: Public Library of Science.