The World of Harry Potter: How well do the Films Represent the Books?

Helga Þórey Björnsdóttir (s213615)[a], Rebekka Jóhannsdóttir (s212963)[a,b,1], and Sara Húnfjörð Jósepsdóttir (s212952)[a]

[a]Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark; [b]Department of Technology, Management and Economics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

This manuscript was compiled on December 14, 2022

The ability of network analysis to reveal intricate and frequently emergent patterns and dynamics has led to its widespread adoption in the scientific community (1). Text analysis uses natural language processing (NLP) to read and understand human-written text using computer programs to gain insights (2). Research on fictional novels has combined these two fields to show how successful novelists have been able to replicate social networks in their works, which possess complex network properties that social networks in the real world adhere to (3). One such network is the Harry Potter character network. Here we show that there are differences between the social networks introduced in the books compared to the one from the films, owing to the fact that the films represent only a subset of the characters in the books. This was to be expected, as the screenwriters had to eliminate characters and change the storyline of the books to avoid confusion and adhere to time constraints. Both networks, on the other hand, follow a power-law distribution with scale-free and small-world properties. Furthermore, text analysis showed variations in the most common words appearing in the books and films as well as the character's sentiment or happiness scores. However, the films seem to have captured the overall feel of the books as the sentiment analysis of the character's happiness showed similar patterns.

Social Network | Harry Potter | Sentiment Analysis | Text Mining

Complex networks, including social networks, have become an increasingly important part of science (1). It is fascinating to explore how successful novelists, such as J.K. Rowling, are capable of recreating social networks that exist in reality (3). The Harry Potter series, which spans seven books and eight films, comprises multiple narratives that can be challenging to follow at first. Although the films follow the book's storyline, we argue that there are distinct differences between the social networks and the writing.

The purpose of the research presented in this paper is to better understand the difference between the films and the books by utilizing network science tools and natural language processing (NLP) for text analysis. We intend to conduct a network analysis to see all of the existing relationships and determine whether the films manage to maintain the structure of the character's social network. Text analysis was then performed to determine if the films accurately conveyed the tone of the books. This allows us to begin understanding the complicated nature of the stories in the books.

It is obvious that the Harry Potter films cannot include every sub-storyline because they would end up being far too long. Therefore, it will be interesting to examine whether the networks of the films and books both replicate real-world networks because the film's social network is a subset of the book network. Numerous characters appear in the series, but it mostly centers on Harry Potter and his two closest friends, Hermione Granger and Ronald Weasley. It will be interesting to see whether certain network properties such as degree centrality, which indicates node importance, will change between the books or films and how they develop throughout the series. It is also interesting to investigate the difference in character dialogues between films and books by doing sentiment analysis as well as comparing the overall happiness scores. The difference will be further explored by looking into the most significant words appearing in the books versus the movie scripts. We will also look into gender bias, whether females or males are portrayed as happy or sad.

The results of our analysis will be presented in the following section. Then, in the *methods* section, we will explain the definitions used in our analysis and calculations in greater detail. In the *discussion* section, we will attempt to answer our hypothesis regarding whether the films were able to replicate the world of Harry Potter as it is portrayed in the books.

Results

The Harry Potter Character Network. In general, the strength and robustness of a network can be measured using metrics based on simple concepts such as a node's number of links, links between neighbors of a node, or the number of paths that could be established (4). The character network in Harry Potter was constructed using articles from the wiki hosting cite *Fandom* (5). A visualization of the network can be found in (SI Appendix: Fig. S1). If the name of another character was mentioned on the character's page, the character was linked to it. The resulting network was then divided into two smaller networks, one containing characters appearing in

Significance Statement

We investigate the Harry Potter series using network science and text analysis to determine whether the films manage to capture the complex nature of the books. Many articles have touched upon the subject of comparing films based on books. Nevertheless, our goal was to conduct more thorough research that demonstrated, among other things, that removing characters from films alters the connections between characters, causing the social network to change. Comparative analysis for each book, film and character in the network was performed, which revealed that the films did a good job portraying the feel of the books. However, we saw that fairly happy but minor characters did not make it to the films.

Author Contributions: H.Þ.B and R.J. performed text analysis and S.H.J. created and analyzed the network. H.Þ.B, R.J. and S.H.J wrote the paper.

The authors declare no conflict of interests.

[1]To whom correspondence should be addressed. E-mail: s212963dtu.dk

the books and another for characters appearing in the films. Table 1 provides a summary analysis of the networks. As was expected, the film network contains the fewest nodes as not all characters from the books are included in the films. On the other hand, it has the highest average degree which indicates that characters in this network have more connections than in the book network.

Table 1. Comparison of the Harry Potter character network as a whole and divided into two separate networks for characters appearing in the books and in the films.

Network	No. nodes n	No. links L	Avg. degree $\langle k \rangle$
Complete	462	7108	30.77
Books	436	6597	30.26
Films	310	6263	40.41

The Small-World Effect. Previous research on fictional character social networks, such as Harry Potter and the Lord of the Rings trilogy, has revealed complex network properties such as small-world and scale-free properties (3). This prompted us to investigate whether the film network possessed these properties.

The *small-world effect* is the observation that the network has a small average path length and large clustering coefficient (6). The average shortest-path length for the film and book networks was $\langle d \rangle_{films} = 2.11$ and $\langle d \rangle_{books} = 2.29$, respectively. The average clustering coefficient was $C_{films} = 0.62$ and $C_{books} = 0.59$. This is consistent with a study on a Harry Potter network generated from the books, which yielded similar results, with $\langle d \rangle = 2.78$ and $C = 0.42$ (7). The networks thus have a considerably small shortest path length and a large clustering coefficient, indicating that the characters have a large tendency to cluster together. This confirms that both the book and film networks are small-world networks. This was expected, as the story expands around just a few main characters and it is evident that most of the characters can be found through Harry Potter and Lord Voldemort - resulting in a small average path length.

The Scale-Free Feature. A network is scale-free if the node degrees obey the power-law degree distribution. According to reports, many scale-free networks have a degree exponent, γ, in the range of 2 to 3 (8). This is especially true for social and computer networks, e.g., the internet. These networks are resilient against random decay; removing a node causes them to shrink but not fall apart. Furthermore, a degree exponent under 3 means that the scale-free network can continue to remain connected indefinitely (9).

Figure 1 shows that the degree exponents for the power-law fit were $\gamma_{books} = 2.79$ and $\gamma_{films} = 2.76$, respectively. This is within the desired range, and it is clear from this and the degree distribution plots that the degree distributions of the two networks satisfy the power-law and are thus scale-free. On the other hand, the book network has more power-law relationship than the film network. It is evident this is due to the character network being altered when the books were converted to movie scripts, as all characters couldn't be included. Even major characters such as Charley Weasley, Ron Weasley's brother and the second oldest son of Arthur and Molly Weasley, and the mischief-making ghost Peeves (10).

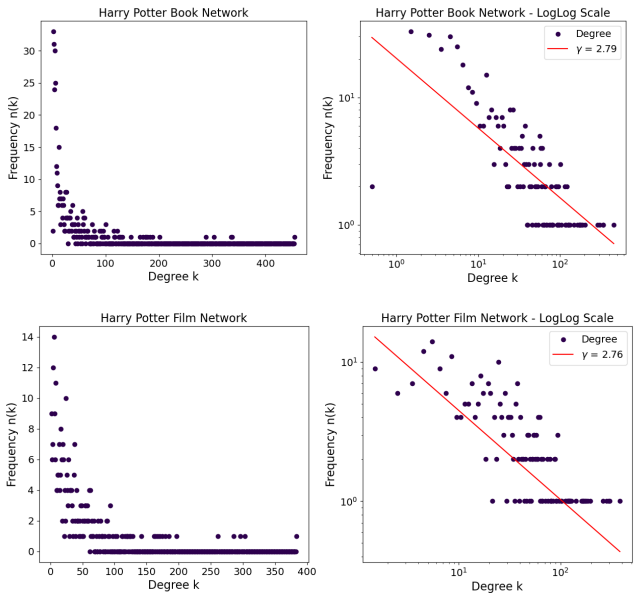

Fig. 1. The degree distribution of the book and film networks in normal and log-log scale. Both networks have a power-law distribution with degree exponent, γ around 3 which conforms with the property of scale-free networks. The film network exhibits less power-law behavior because it does not accurately represent Harry Potter's original social network as the books do. The analysis confirms that both networks are scale-free.

Character Importance. Character importance can be analyzed by looking at a property called degree centrality C_d, which can be computed for each node, d, as the fraction of nodes it is connected to. Another property, the betweenness centrality, C_B, tells us the most influential node in the network. That is, nodes that are *bridges* between other nodes. We computed the degree and betweenness centrality for the characters. Not surprisingly, Albus Dumbledore was among the top three characters, with $C_B = 0.12$ and $C_d = 0.77$ in the books and $C_B = 0.10$ and $C_d = 0.96$ in the films. He is the headmaster and the heart of the books, so he is the main *bridge* between the students and other important characters.

Figures 2a and 2b depict the evolution of character significance for the three main characters, Harry, Ron, and Hermione, as well as Rubeus Hagrid and Sirius Black. We wanted to look into the latter two characters in particular based on our reading and film experience and analysis of overall character centralities. This is because Hagrid appears to play a larger role in the books than in the films, and Sirius Black does the opposite. As can be seen, the degree centrality in the films is higher than in the books. Ron and Hermione exhibit similar fluctuations, appearing to have the greatest importance in the first few books before plummeting dramatically in the middle of the film franchise. Interestingly, Sirius Black appears to fluctuate more throughout the films, peaking at film 5 for betweenness centrality when he dies at the hands of Bellatrix Lestrange. According to the analysis, everyone appears to be more important in the last film than in the book. Furthermore, Hagrid's higher overall degree centrality in the books indicates

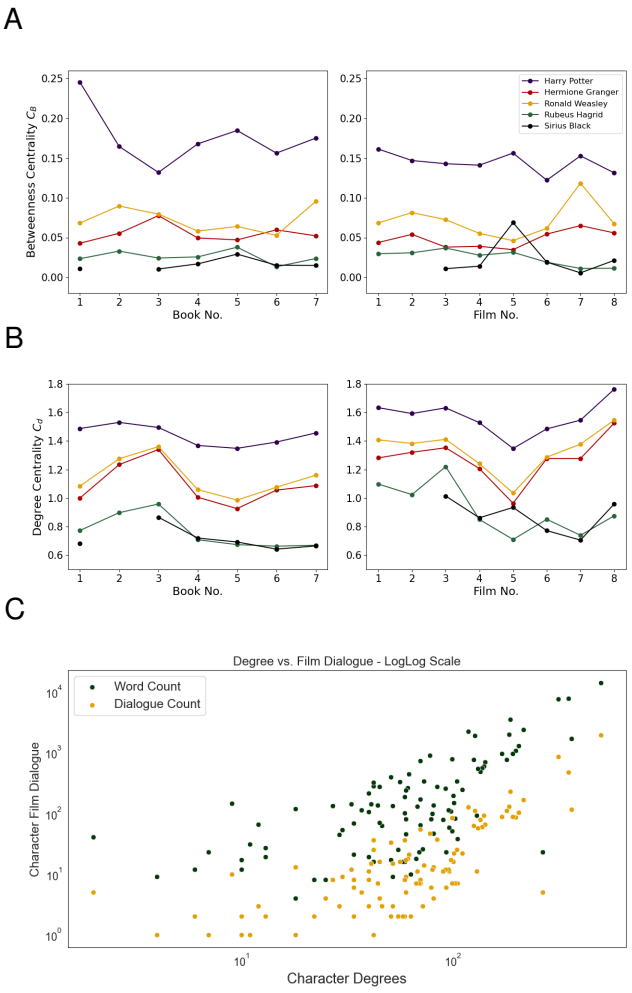

Fig. 2. Harry Potter character network analysis. (A) Analysis of most influential characters throughout the series by means of betweenness centrality. Sirius Black peaks in importance during the 5th film, *The order of the Phoenix*, when he dies but this peak is not present in the books. (B) Most important characters throughout the series, by means of degree centrality. Higher values of centrality are observed in the film than in the books. (C) Correlation of degree and film dialogue. There is a slight correlation between the degree and number of dialogues and words of characters in the films.

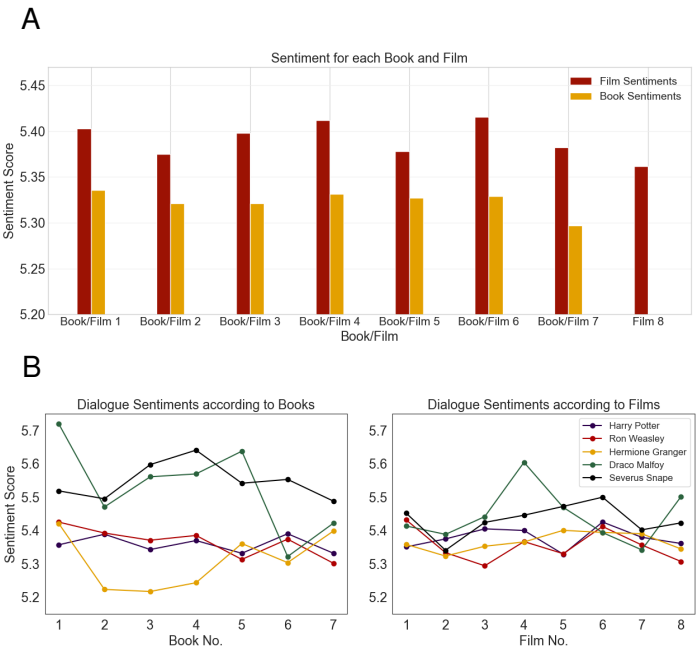

Fig. 3. Sentiment analysis on each book and film. (A) Overall sentiment scores show that the films and books share similar patterns across different film/book numbers.(B) The dialogue sentiment scores for each character show that Draco and Snape have higher sentiment scores than the other characters. Hermione has the lowest score of the bunch according to the books but Ron according to the films.

that the characters left out of the films appear to have had a stronger connection to Hagrid than Sirius.

The characters with the highest degree would be expected to have more dialogue in the films. Figure 2c is a log-log plot and shows that there is a slight relationship between the two. Relationships of the form $y = ax^k$ appear as straight lines in a log-log graph. From that, we can interpret character word and dialogue count from the films to increase exponentially as the degree increases. Basically, the more connected a character is, the more often and more words it will speak in the films.

Text Analysis. Text frequency analysis was done by calculating TF-IDF on the books and the movie script dataset to investigate whether the films capture the main feel and purpose of the books. Word clouds are shown in the appendix (SI Appendix, Fig. S4, Fig. S5), where larger words have more significance. It is interesting to see which words have more importance, for example in *Prisoner of Azkaban* we see that Buckbeak, the hippogriff, which is essentially a mixture of a horse and a bird, is one of the most important words in the film but not as much in the book. Perhaps the film producers wanted to keep Buckbeak more relevant in the film since making the creature took more than a year for the animators as they wanted to make Buckbeak look as real as possible (11). The second film script appears to place a greater emphasis on Dobby, although the second book does not. We also see that happening in *Deathly Hallows Part 1* film where we don't see his name in the book word cloud. This is probably because Dobby appears in every book except two and only appears in two films, making his TF-IDF score larger in those films. This could also be because the film script only contains character dialogues and Dobby always refers to himself in the third person, so it makes sense that his name would appear more often than others.

Sentiment Analysis. Sentiment analysis for each book, film and character in the network was performed using the LabMT wordlist (12). Having the sentiment values provides the opportunity to analyze various aspects of the Harry Potter universe, such as; if the happiness score of the books and films differs significantly, or if the difference is due to whether the movie's writers omit certain characters to make the film more exciting, and if the sentiment score for the majority of the characters varies from less important characters.

In figure 3a the sentiment scores for each book and film are shown. We observe that the book sentiments are lower overall, but it is hard to compare the two because the books include more text and a storyline while the movie scripts only include character dialogues. However, it is interesting to see that we observe a similar pattern between the two. The first book and film have pretty high sentiment scores, then it decreases, then increases again for *Goblet of Fire*. It is hardly surprising

that *Deathly Hallows (Part 1 and 2)* have the lowest sentiment scores. The final book and films have a more sinister tone as they tell the story of how the war against evil ends.

The character dialogue was extracted from the books by finding who said what with regular expressions. We compared the sentiment scores of the character dialogues and how they evolve for both the films and books. This is shown in figure 3b. We observe that Draco Malfoy and Severus Snape have higher sentiment scores than the other characters. This is interesting as they are one of the darker characters of the series. Perhaps their dialogues are being misinterpreted as happy. For instance, Malfoy's remark, "The Dementors Send Their Love, Potter!" may be taken positively, but Harry Potter fans realize it's actually a dig. It is interesting to see the sentiment score development of Hermione. In books 2 to 4, they are the lowest among the others but then increase in book 5. We also see a kind of a similar increase in her film sentiment values but there she is perceived as happier than Ron. The low sentiment values in the books could be explained by the fact that Hermione is initially portrayed as an annoying perfectionist who hates breaking the rules. Then, when she continues to stand up for her friends and even begins breaking the rules to help them, she evolves into a more likable character (13).

Figure 4b demonstrates that characters with large amounts of text on their *Fandom* page, i.e., major characters, tend to be less happy. In light of this, it was decided to examine the characters with more than 20,000 words on their *Fandom* page to determine the happiest and unhappiest characters; the results indicated that Ginny Weasley is the happiest character and Lord Voldemort is the unhappiest character. Since Ginny Weasley is portrayed as relatively innocent in the story and has romantic feelings for Harry, it makes sense that she would appear in happy scenes. Lord Voldemort is the dark lord and the story's primary antagonist, and thus he was expected to be at the bottom.

Figures 4a and 4c illustrate that the characters in the books are a bit happier on average than the characters in the films. It also demonstrates that the characters in the films have the full sentiment range of the characters in the books. From figure 4b where the sentiment score is plotted against the number of words, it can be seen that characters with sentiment scores of approximately 5.37 or higher have few words on their *Fandom* page. That should imply that they are not the main characters. Looking at the grey box in 4a, the number of characters in the range from 5.37 to 5.42 is much lower in the films than in the books, whereas other characters in different ranges appear in both. This means that the screenwriters decided to eliminate some unimportant characters from the books in order to reduce the size of the cast and prevent confusion.

An examination of the female characters' happiness revealed that there is a bias (SI Appendix, Fig. S6, Fig. S7). Despite the fact that there are more supporting male characters, which should lead to increased average happiness for males, the female characters are generally happier in both films and books. This was to be expected because male characters are typically interpreted as more evil, while female characters are typically interpreted as cheerful (14).

Discussion

Results showed that more connected characters also had more dialogue in the films, as was expected. What's more, the TF-IDF word clouds revealed interesting distinctions between the books and films (SI Appendix, Fig. S4, Fig. S5). In several instances, we saw that words were given a lot more weight in the films than they were in the books and the other way around. However, the word clouds of the book and corresponding film usually share similar vocabulary, indicating that the films were successful in conveying the book's main message.

The network analysis revealed that both networks are small-world and scale-free, as expected from literature (7). The degree analysis indicated that both networks followed power-law distributions. However, the film network distributions are not as well-defined. A plausible explanation is that characters were omitted from the films, generating only a subset of the social network of the books. This confirms that the films capture a subset of the book's network structure, with the exception of the distributions diverging. In addition, the importance of characters showed similar patterns, although being higher in the films than in the books. The significance of a few major characters was diminished as a result of characters being omitted from the films. Sirius Black seemed to have more influence in the films than in the books, especially in the fifth film, implying that the screenwriters decided to make him a more prominent character.

Figure 3a showed that the films and books both have similar patterns in their sentiment scores which confirms that the films managed to capture the feel of the books. A surprising finding was that Severus Snape and Draco Malfoy had higher sentiment scores than Harry, Ron and Hermione throughout the series. Nonetheless, it is interesting to see that this holds for both films and books. Implying that the films comply with the interpretation of these characters in the books, although they failed to mimic the behavior of Hermione, which has lower sentiment scores in the books. The average happiness score of characters in books was higher than that of characters in films. This was expected given it was shown that major characters, who are not omitted from the films, tend to have lower happiness scores than minor ones. It should be noted that the algorithm that was implemented to extract character dialogue from the books could not find which character spoke if the quotations were followed with a "he said" or "she said" but only if a character name was specified. This was not implemented due to a limitation of study.

To summarize, the findings reveal that, while the film's social network is not an exact representation of the book's, the film network does capture the majority of the structure of the character network as well as the overall tone of the books. When it comes to the happiness of the characters the films manage to get the whole spectrum of happy and less happy characters. In addition, the gender bias was not surprising given that the villains in most movies are typically men (14). Future studies could be done to better investigate the behavior of certain groups in the network such as the Hogwarts Houses. To guarantee that every character dialogue is preserved, the algorithm used to extract book dialogues could be improved.

SI Datasets. HPCharactersData.csv, dataset_S01.txt, Books 1-7, Film scripts 1-8, HarryPotterWiki, For Network.

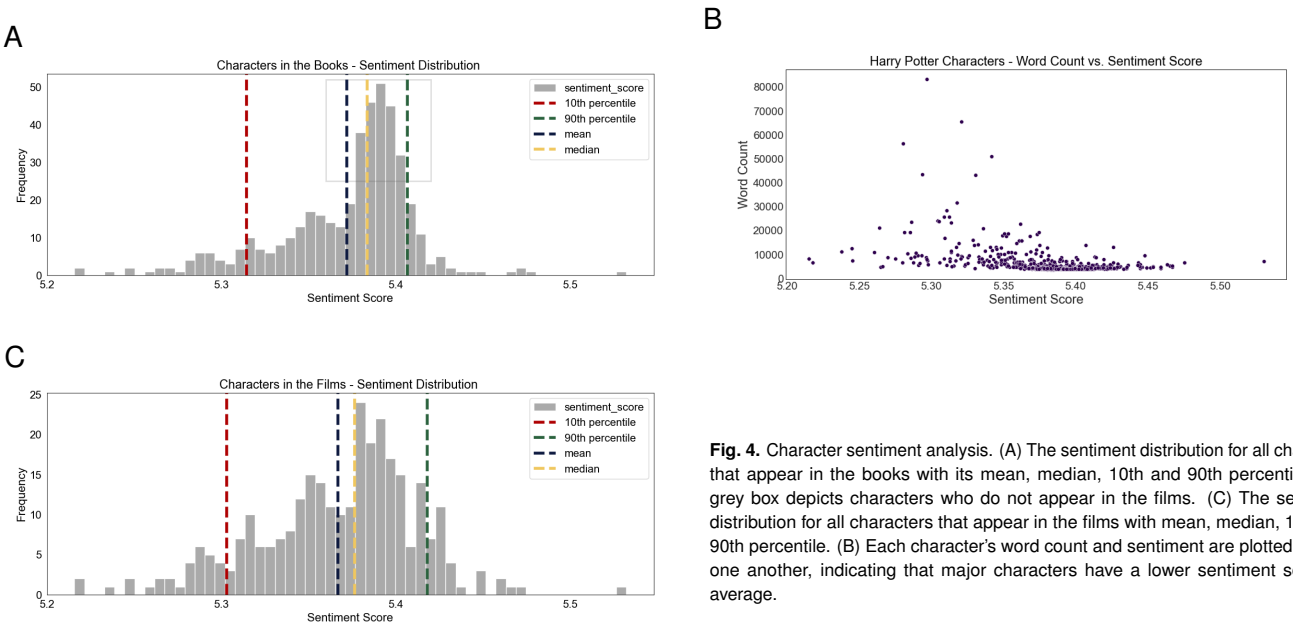

Fig. 4. Character sentiment analysis. (A) The sentiment distribution for all characters that appear in the books with its mean, median, 10th and 90th percentile. The grey box depicts characters who do not appear in the films. (C) The sentiment distribution for all characters that appear in the films with mean, median, 10th and 90th percentile. (B) Each character's word count and sentiment are plotted against one another, indicating that major characters have a lower sentiment score on average.

Materials and Methods

Explainer Notebook. The notebook containing the code along with further explanations of the data process can be found here, also in the GitHub repository.

The Datasets. The datasets used for the analysis were six. A list of Harry Potter characters and a text file for each book were extracted from *Kaggle* (15). Data for the sentiment analysis was obtained from *LabMT* and consisted of 10222 words with computed happiness average (12). The eight movie scripts were extracted from *GitHub* (16). Files containing the text about each character were generated using articles from the wiki-hosting site *Fandom* (5).

Definition of a Power Law Distribution. Given a degree distribution of a network, $n(k)$, i.e., the number of connections each node has, a power-law can be defined as (17):

$$n(k) = k^{-\gamma}, \qquad [1]$$

with exponential cut-off, where γ represents the degree exponent.

Definition of Sentiment Analysis. Sentiment analysis is the most widely used method for determining whether an incoming message is positive, negative, or neutral. The authors of "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter" calculated the sentiment score of a number of individual words where higher values indicate a happier text (12). By calculating the frequency of individual words in a given text, the texts weighted average level of happiness can then be found as:

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i) p_i, \qquad [2]$$

where f_i is the frequency of the ith word w_i for which there is an estimate of average happiness, $h_{avg} w_i$, and the corresponding normalized frequency is defined as $p_i = \frac{\sum_{i=1}^{N} f_i}{\sum_{j=1}^{N} f_i}$.

Definition of TF-IDF. TF-IDF, which is short for term frequency-inverse document frequency, is a metric that calculates the importance of a word to a given document in a corpus. This formula comes in great use when analyzing a book series because it shows which words appear frequently in one book, but not the others. The TF-IDF is a product of two statistics, the term frequency (TF) and the inverse document frequency (IDF). The TF is the relative frequency of a term, t, within a document, d:

$$TF(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}. \qquad [3]$$

The IDF is a measure of how much information a particular word provides, i.e. whether it is common across all documents. Defined as:

$$IDF(t, D) = \log \frac{N}{|d \in D : t \in d|}, \qquad [4]$$

where N denotes the total number of documents in the paper and D is the corpus (18).

1. RJ Fletcher, MA Acevedo, BE Reichert, KE Pias, WM Kitchens, Social network models predict movement and connectivity in ecological landscapes. *Proc. Natl. Acad. Sci.* **108**, 19282–19287 (2011).
2. A Kao, SR Poteet, *Natural Language Processing and Text Mining.* (Springer Science & Business Media), (2007) Google-Books-ID: CVtxFWbKT7wC.
3. J Li, C Zhang, H Tan, C Li, Complex networks of characters in fictional novels in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS).* pp. 417–420 (2019).
4. E Xamena, N Brignole, A Maguitman, Structural analysis of relevance propagation models. *Knowledge-Based Syst.* **234**, 107563 (2021).
5. Harry potter wiki. (2022).
6. DJ Watts, SH Strogatz, Collective dynamics of 'small-world' networks. **393**, 440–442 (1998).
7. J Zhang, H Zhao, Jq Xu, Jf Wang, Small-world and scale-free features in harry potter. **12**, 6411 – 6416 (2014).
8. AL Barabási, M Pósfai, *Network science.* (Cambridge University Press, Cambridge), (2016).
9. AL Barabási, Network science. *Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci.* **371**, 20120375 (2013).
10. Harry potter characters who didn't appear in movies. *The Times India* (2021).
11. A Roker, Behind the magic of 'harry potter'. *NBC News* (2004).
12. 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.
13. Harry Potter and the Sorcerer's Stone: Hermione Granger. *SparkNotes* (2005).
14. A Synnott, *: Heroes, Villains and Victims.* (Routledge, London), (2016).
15. Harry potter books corpora (part 1 - 7). *Kaggle* (2022).
16. L Chauvet, Harry Potter movies datasets (2021).
17. H Ebel, LI Mielsch, S Bornholdt, Scale-free topology of e-mail networks. *Phys. Rev. E* **66** (2002).
18. A Simha, Understanding tf-idf for machine learning (2021).