(Dis)Passion of Deduction

Roshnee Chaterjee, Veda D, Vibodh Nautiyal

07/03/2021

Introduction and Research Question

We were interested in seeing how detectives in crime fiction in the 19th & 20th century were written and perceived at the time. A commonly observed stereotype portrays detectives as cold, calculating and emotionally distant. These traits are largely associated with individuals of superior intelligence. There is a common belief that these rational characters do not allow emotions to clout their judgement. We would like to examine whether NER, mapping, and sentiment analysis can help us confirm this hypothesis.

Research Question

Do detective protagonists in detective fiction novels in the 19th and 20th century have a more neutral emotional state compared to other characters?

We feel this may be the case as they would be more focused on the logical aspect of the mystery rather than the emotional upheaval that arises from it.

Research Hypothesis

Detective protagonists would have a more neutral emotional disposition as opposed to the other characters in their story.

Corpus Description

Our corpus consists of detection fiction novels written in the 19th and 20th century by authors like Agatha Christie, Arthur Conan Doyle, H. Beam Piper and Ernest Bramah. These novels were selected because they have a central private investigator character. We aim to do a comparative study between the range of sentiments of the central detective character and other characters on a book by book basis. While we would have liked to add more novels to our corpus, there were a limited number of novels with a central private investigator available on Gutenberg. We also tried to limit our corpus selection to single stories as much as possible, as identifying characters in a collection of short stories would be extremely tedious — though we did have to go through this process for one of our books Max Carrados.

Summary

Here are some general statistics about our corpus (See Fig 1.1):

1. Total number of words in the corpus: 348,054 words
2. The 3 longest books in the corpus:

   i. The Secret Adversary: 76,203 words
   ii. Max Carrados: 68,809 words
   iii. Murder in the Gunroom: 68,288 words

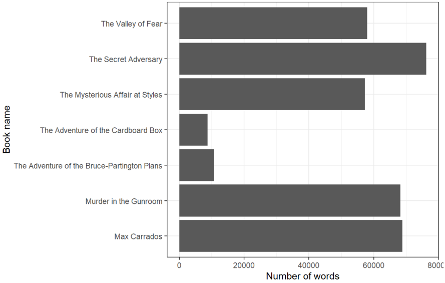
3. The 3 shortest books in the corpus:

   i. The Adventure of the Cardboard Box: 8,695 words
   ii. The Adventure of the Bruce-Partington Plans: 10,775 words
   iii. The Valley of Fear: 57,993 words

All of these statistics were obtained using the count function in R.

Number of words per book

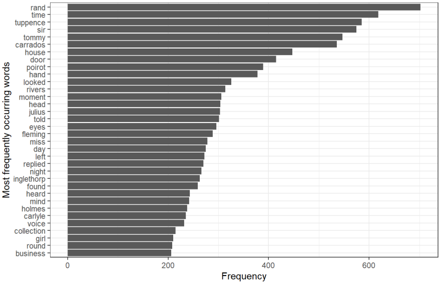
Figure 1.1



Next, we'll look at the frequency of words across the corpus. We began by removing the stop words from our corpus. Currently, we are examining words that have a frequency of greater than 200 (See Fig 1.2). The most frequently occurring word is Rand which appears close to 700 times in the corpus. This is because the protagonist in Murder in the Gunroom is Colonel Jefferson Davis Rand. Holmes, Poirot, Tommy, Tuppence, Carrados and Rand — all the detectives appear in the frequency table as they are the protagonists of their respective books. Other frequently occurring words such as "house" (approximately 440 times) could be a reference to the locations of suspects and the murder scene. The high frequency of the word "doors" (referenced approximately 410 times) could suggest frequent movement, with characters moving through locations to solve mysteries.

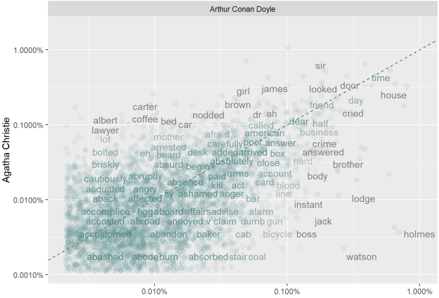
Words occuring >200 times in the corpus

Figure 1.2



We also created a graph which plots the common usage of words between Doyle and Christie's work within the corpus. As seen in Figure 1.3 many words are used across both authors' works. The words that are close to the x = y line have a similar frequency in both sets of text. Seeing as there is such a dense gathering around the line, we could infer that Doyle and Christie use a lot of similar words and might have a similar literary style. Certain outliers such as Holmes and Watson are specific to Doyle's work and are singled out in the bottom right corner (as seen in Fig. 1.2) to reflect as such. It is interesting to note that the focal detective characters in Christie's work such as Poirot, Tuppence and Tommy are not singled out in the same way.

Figure 1.3



Data Visualization #1

We did a comparative analysis of the sentiments of the protagonist of each novel in comparison to other primary characters in the same text. Doing so would allow us to view how each detective is perceived in their novel(s) within the corpus. Using NER and sentiment analysis, we were able to construct the graphs shown below. We calculated the positive and negative sentiment score of each character, and then found the difference between the two. The mean was taken to normalize the scores for comparison.

In order to see how the protagonists of each novel fare in relation to other characters, we have chosen a few characters from each novel for comparison. We chose the side characters on the basis of how central they are to the plot and how frequently they are mentioned in the text, often choosing the culprit, the victim, or the victim's family members/friends. Our rationale behind this was that the most frequently occurring characters would provide the best control test for our hypothesis. The graphs we generated show on the whole, the detectives seem to have a sentiment score of close to 0 as compared to other characters. This does seem in line with our hypothesis that detectives have a lower difference between positive and negative emotions as compared to other characters.

Overall, all detectives throughout the corpus show a positive sentiment value of less than 1, barring Sherlock Holmes in two of the books - The Adventure of the Cardboard Box and The Valley of Fear (with values 2.3 and 2.4 respectively). The low positive score of the detectives suggests that their positive and negative emotions are counterbalanced, resulting in a score that is close to 0. It is important to note that a sentiment score close to 0 only suggests that the detectives have a similar positive and negative score.

On the other hand, antagonists seem to be expressing a net positive sentiment in multiple cases (Oberstein: 2.2 (See Figure 2.1), Jim Browner: 1.8 (See Figure 2.3), Mr Brown: 1.2 (See Figure 2.6)). While one might expect the antagonist to have a net negative sentiment score, authors do not make it obvious who the culprit is in a whodunit. Rather, they try to surprise you by presenting the antagonists in a relatively positive light. Other characters in the corpus who turned out to be murderers had a score relatively close to 0 (Alfred Inglethorp: -0.15 (See Figure 2.5) and Fred Dunmore: -0.9 (See Fig 2.7)), possibly due to the same reason.

The characters that are victims of crimes also seem to have a high negative sentiment value (Lane Fleming: -1.1 (See Fig 2.7) or Emily Inglethorp: -0.4 (See Fig 2.5)). The characters may not be negative. It is possible that the negative sentiments being associated with them is a result of these characters being referenced in sentences with negative connotations. Their negative sentiment score may not be a statement on their personality.

Observing the sentiment analysis scores of our non-protagonist characters, and the possible reasons for the same, helped to solidify our understanding of how the protagonist sentiment scores were playing out. In comparison, the protagonists in these novels have a lower net positive sentiment score compared to other characters. This would imply that they seem to have a more neutral disposition which is in line with our hypothesis.

Drawbacks

Sentiments to a character are assigned by the line that the character is mentioned in. This could lead to inaccuracies when multiple characters are mentioned in the same line, as they would be assigned the same sentiment value. All the sentiments in a sentence may not correspond to all the characters mentioned equally. Further, the sentiment score of a character does not merely refer to their own emotions. It also refers to the manner in which the character is referenced by others.

NER also fails to pick up on pronouns. As a result of this, the sentiment analysis we perform only accounts for the times when the characters were explicitly mentioned by their name. We noticed that Watson (the narrator of the Sherlock Holmes books) was barely picked up by NER, as he is rarely mentioned by name.

Figure 2.1

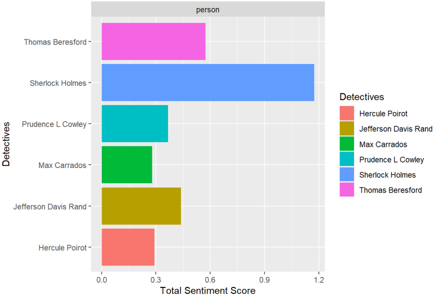


Figure 2.2

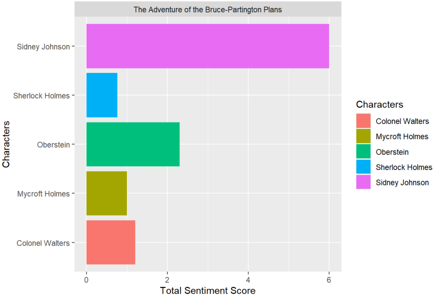


Figure 2.3

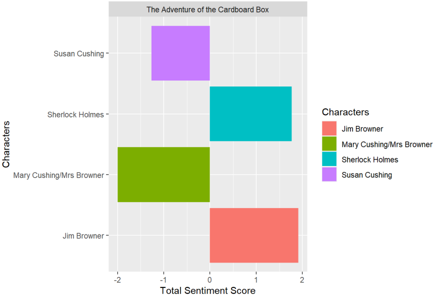
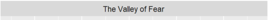


Figure 2.4



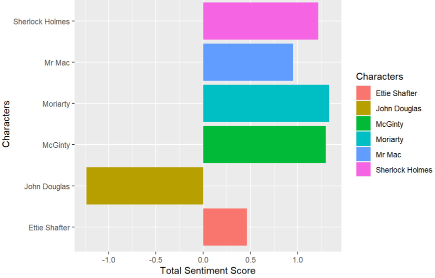


Figure 2.5

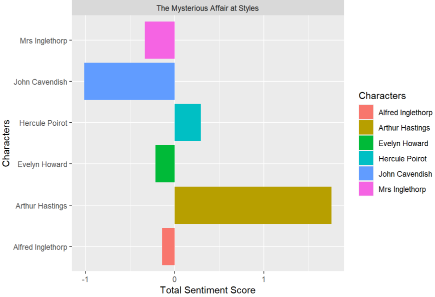


Figure 2.6

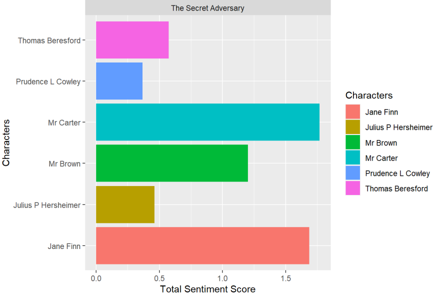
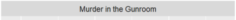


Figure 2.7



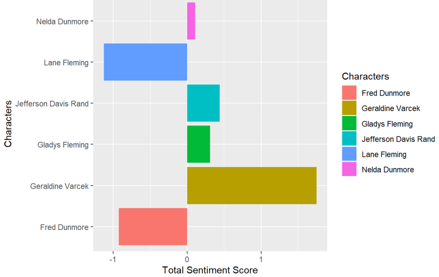
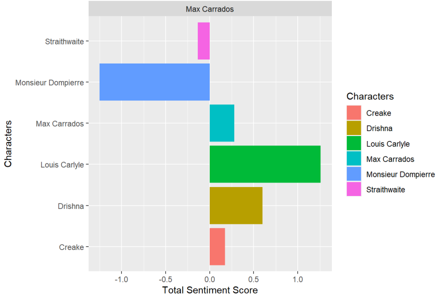


Figure 2.8



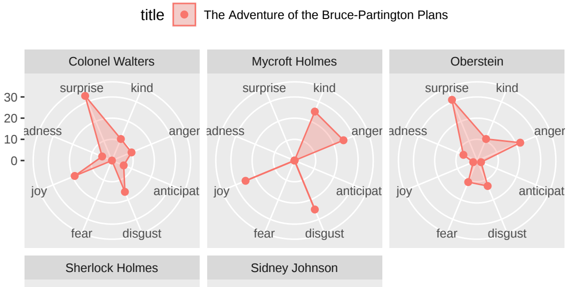
Data Visualization #2

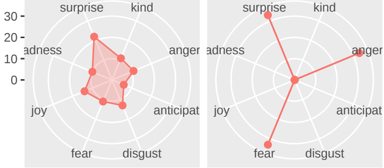
We constructed radar plots for the detectives and main characters in each novel. The radar plot measures the relative frequency of the different sentiments of a character. For example, if a character has a much higher sentiment score for fear than other sentiments, then there would be a spike towards fear in the graph.

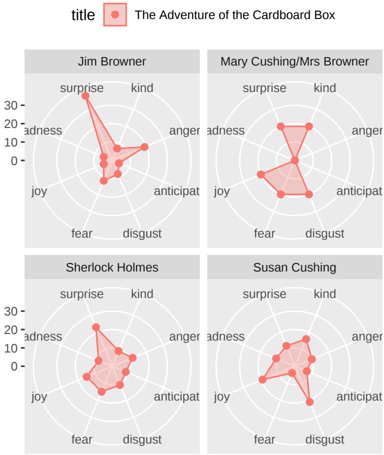
Characters such as Sherlock, Thomas Beresford, Prudence Cowley and Poirot show a relatively constricted emotional radius compared to other characters in their novels (approximately 10). However, detectives such as Jefferson Davis Rand and Max Carrados have a higher radius of approximately 20 for surprise (See Fig 3.1 to 3.7). Overall, the detectives seem to have a more balanced radar plot compared to other characters, though there are some exceptions in Jefferson Davis Rand and Max Carrados. With these exceptions in mind, we can say that the detectives do not seem to veer towards any one emotion.

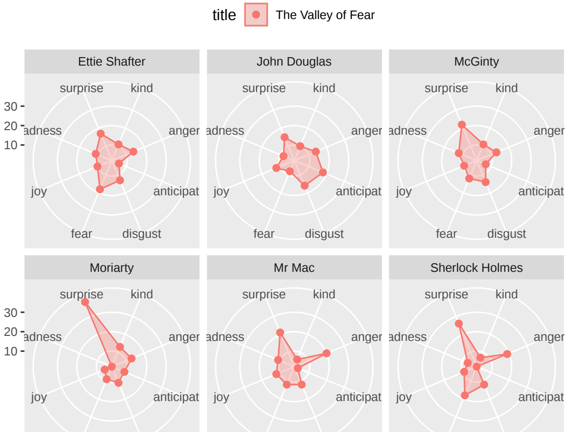
Most detectives have their highest peak in their sentiment radar towards surprise. This may simply be a function of the genre and indicate them discovering new clues in regular intervals. Figure 1.2 suggests that Doyle and Christie have similar styles of writing. This may have resulted in the emergence of similar characters in their books, with similar sentiments associated with them.

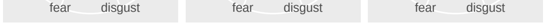
However, the few exceptions that we have prevent us from declaring conclusively that detectives always have a constricted radar plot. We would need a larger corpus to do so.

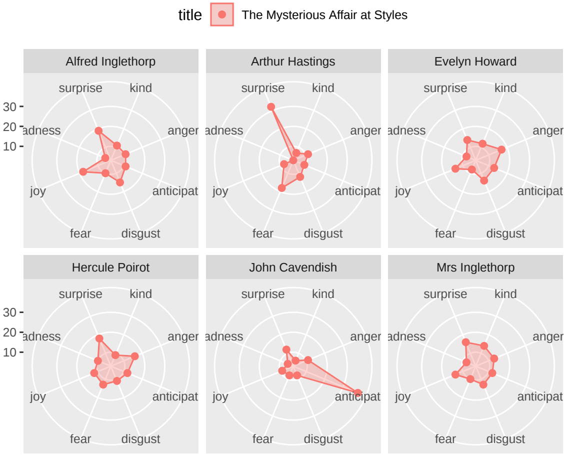


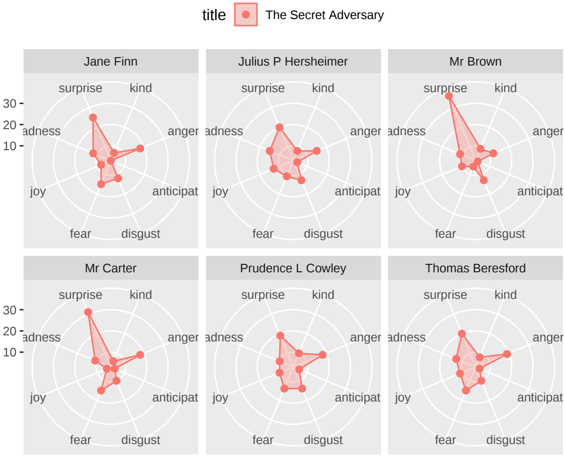


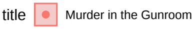


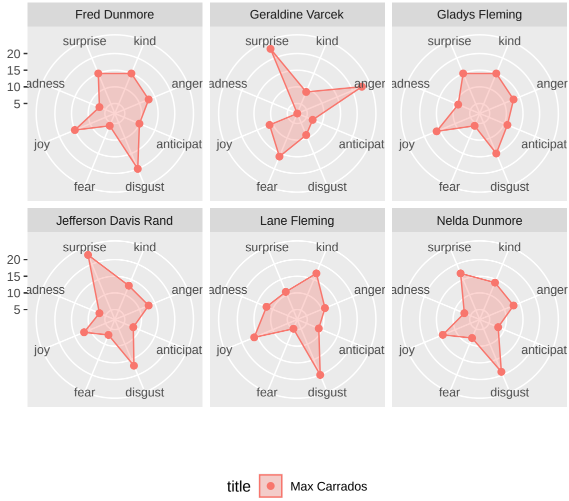


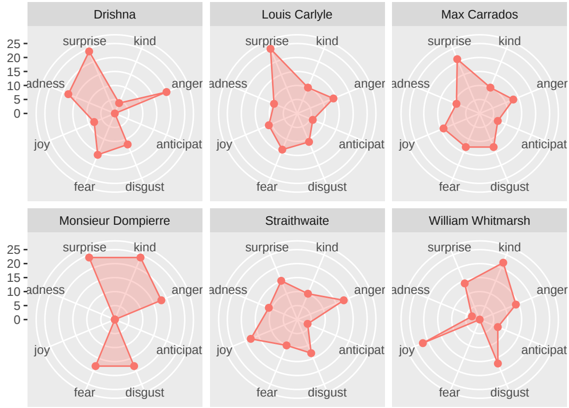












Conclusion

The inspiration for our hypothesis stemmed from the recent portrayal of Sherlock Holmes in television. The cold and reserved nature of a character seems to have a correlation with intelligence and logical deduction in popular culture. Our analysis gave us insights into other detectives in these novels, and Sherlock Holmes in particular, that we were unaware of earlier. For example, Sherlock had the largest net positive sentiment score. This could show that there is a significant difference between the Sherlock Holmes portrayed in 21st century television, and the Sherlock Holmes that Arthur Conan Doyle created in the 19th century.

The detectives in our corpus seem to have a balanced set of sentiments associated with them. They have a similar positive and negative sentiment score, and a relatively constricted radar plot. This suggests that the various sentiments associated with them are equally distributed. We believe that this supports our hypothesis that the detectives in this genre have a neutral disposition. However, we recognize that our corpus does not allow us to prove our hypothesis concretely. With more time and resources, we could examine hundreds of detective fiction books and see whether the results would differ.

Reflection

Keeping our data accurate was very time consuming and tedious. The SQL Join function would associate an entity with a sentence even if the sentence had the letters of the entity as a suffix or prefix in a word. Fixing this required us to fix the entities that the NLP library was identifying upstream. This required judgement calls on which entities were required and which weren't. There were similar subjective calls that we had to make when manually cleaning our data as it was not feasible to check every character in the original plaintext. We also had to manually sort the characters to make sure that they only appeared in their respective books. The NLP library also often tagged a person as a location and vice versa, which was also a pain to clean up.

Using R for our text analysis allowed us greater flexibility to single out characteristics that we were interested in. For example, we were able to isolate the characters that we were interested in, and make charts just on them. We could filter through the data as we wished. In Voyant, it would not have been possible to get so specific in our analysis. To summarize: R is superior for drilling down on specific details. R is more suitable when your hypothesis asks a very specific question about the corpus, whereas Voyant is limited to broader questions about the corpus.

However, the scope of sentiments in the NLP package limits our understanding of complex characters and their emotions. Crimes of passion have more complex motives than just anger and disgust. Understanding such emotions often needs human intervention which was not accounted for by the sentiment analysis. Had the range of emotions been more complex and the NLP technique more sophisticated, we may have had more accurate results.