Jane AI-sten: What is Sentiment Analysis's Connection to Best-Selling Literature?

Casey Leach with Professor Chun and Professor Elkins
IPHS 200 Programming Humanity
Kenyon College

Abstract

Through sentiment analysis, we can explore the intersection of technology and literature, specifically the ways in which literary insight can be gained through data analysis. Focusing on Jane Austen's six major novels, I used sentiment analysis on the raw text of each to examine the novels' plot structures. Comparing the sentiment analysis graphs with the sales rank of each novel, as well as the typical plot structures that accompany best-sellers, I was able to further understand how literature is more uniform and predictable than one might think. Certain plots and classic literary tropes are more popular than others, and Austen is no different from other authors in her use of them. Throughout the project, it becomes clear that literature can be understood through a technological lens even if those in the humanities resist the idea that literature is too subjective and variant to be confined to one specific plot that sells well. The human urge for comfort and predictability is fought against, yet literature like Jane Austen's works proves that we love drama but also happy endings.

Introduction

There is a variety of opinions on the involvement of technology in literary studies. Some are skeptical of whether or not AI and programming have a place in analyzing literature, while others are adamant that we can gain valuable insight. Researchers have used AI to locate themes, perform character analysis, and even mimic certain authors' writing styles. Jodie Archer and Matthew L. Jocker have done extensive research to figure out which plots make best-selling novels. For this project, Archer and Jocker's analysis is key to finding out if there is an underlying reason for Jane Austen's novels' success: can we track why some of her books are more popular than others?

Austen published six major novels, four in her lifetime and two published posthumously. *Sense and Sensibility, Pride and Prejudice, Mansfield Park, Emma, Northanger Abbey,* and *Persuasion* are all successful novels, but to varying degrees. In the mainstream media and popular culture, *Pride and Prejudice* and *Emma* are the most talked-about works of Austen's. Both novels have movie adaptations and movies that take inspiration from Austen's words and apply them to a slightly deviated plot, like *Bridget Jones's Diary* and *Clueless.* What made these novels more prevalent than the rest of Austen's novels? Can we find evidence in the sentiment analysis of every text? Though a total number of copies sold is not available for these texts, Barnes and Noble keeps a sales rank of their texts. The bookstore ranks books daily based on rolling six months of sales data. If there are no sales during a particular amount of time, or if other titles have more sales, a title can lose its ranking. Using the Penguin Classic versions of Austen's texts, here are their most recent sales ranks in 2021:

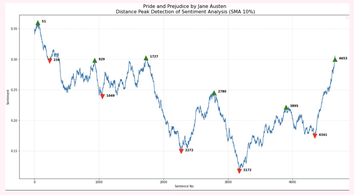
Pride and prejudice - 3,170
Emma - 55,595
Northanger abbey - 91,778
Sense and sensibility - 110,310
Mansfield park - 185,618
Persuasion - 873,484

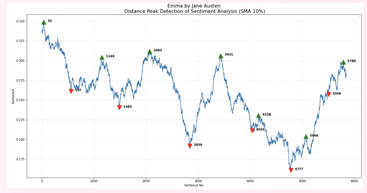
Each novel has a rank, indicating successful sales even today despite being published in the 1800s. However, there are obviously large differences in the current reception of the texts. Looking into the sentiment analysis of these novels can shed light on what makes novels popular, and also makes a broader point regarding how literature and technology can coexist and create a richer analysis than one would have otherwise.

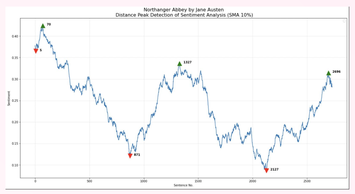
Methodology

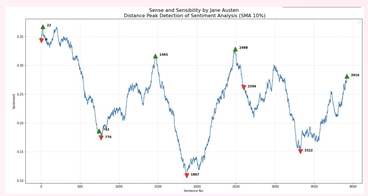
With the help of Professor Chun, I uploaded the raw texts of Austen's novels from projectgutenberg.com into a Google Colab notebook. The notebook cleaned the texts of any headers and footers, as well as sentences shorter than two words. Next, I used VADER and TextBlob to run the sentiment analysis of each text. The graphs of the sentiment analysis plotted sentiment on the y-axis and sentence number on the x-axis. Though both VADER and TextBlob were used for the original sentiment analysis graphs, I then moved into only using VADER within the Scipy program to look closer at the sentiment analysis along with peak and valley detections throughout the novels. Scipy takes the original sentiment analysis from VADER and allowed me to look into more detail at the specific moments in the text where a peak or valley was detected. Finally, I compared the sentiment analysis of each book to better understand what made one story more popular than another.

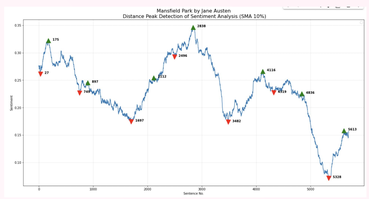
Results

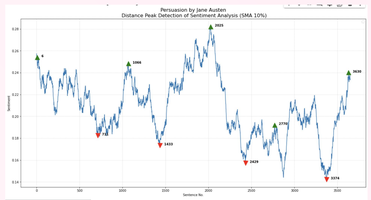












Looking at the sentiment analysis of each novel gives insight into why *Pride and Prejudice* and *Emma* are so much more popular than Austen's other novels. Each novel follows the trajectory of two examples from Archer and Jocker's *The Bestseller Code*: the plots of voyage and return and coming-of-age/rags-to-riches. *Pride and Prejudice*, *Emma*, *Sense and Sensibility*, *Northanger Abbey* and *Persuasion* all follow the voyage and return model: an intellectual or emotional journey for one or more of the characters that end up with a happy ending. This is very common for romance novels. *Mansfield Park* follows the coming-of-age/rags-to-riches model: a storyline where there is a central movement between crisis and success. While both of these plots are typical of bestsellers, it is interesting for evaluating the variance of each novel's individual popularity.
The structure of *Pride and Prejudice* and *Emma*'s plots are extremely similar: they start on a high point, experience some drops and negative moments, but altogether end up on a high at the end of the novel. What sets these two texts apart is the fact that their middle of the story low point is not extremely low, just enough to keep readers on their toes. Both novels also end on high points, i.e., a happy ending, specifically ones where couples end up in love. Though *Sense and Sensibility* follows the same plot model, it has a much lower low point in the middle of the story: all throughout its' low points are much lower than *Pride and Prejudice* and *Emma. Northanger Abbey and Persuasion* also differ from the two most popular Austen novels: they do not have a middle low point but an extremely high one instead, ruining the popular model of having the ending being the happiest point of the story. *Mansfield Park* has a high point in the middle of the story, but ends on a much lower point than all of Austen's other novels.

Conclusion

Examining the sentiment analysis and peaks and valleys of each Austen novel makes it clear that valuable insight into what makes literature popular can be gained from technology. Each novel has an archetypal plot structure that can be seen in many other bestselling novels ranging from the 1800s to today. The sentiment analysis graphs reveal why *Pride and Prejudice* and *Emma* are much more popular than the other novels: they contain just the right amount of drama to keep readers interested but ensure that there will be a conflict resolution at the end of the novel that leaves the characters happy, and in these specific instances, in love. *Sense and Sensibility* takes the successful plot structure of these two novels to an extreme that affects its reception: its low points are too low. Even though *Persuasion and Northanger Abbey* follow a similar model to the most popular of Austen's novels, they lack drama and tension. *Mansfield Park* has high points, but at the wrong moments: it fails to provide that middle of the story conflict that readers would want to see resolved in a happy ending, and instead puts too high of a point at the middle and ends on a low. *Northanger Abbey* is the sole outlier. It is similar to *Persuasion* in terms of structure but ranks as Austen's third best-selling novel. This is exemplary of why literary analysis still needs its own field of study separate from the tech world to understand why this may be.

This research project raises questions about the role of technology in literary analysis. Literary scholars are resistant to create a space for coding, AI, and technology in the literary field because they fear it may take away from the merit of being a reader or a writer. However, literary studies and technological advancements can coexist without one overtaking the other. Technology like sentiment analysis looks at literature in a new way, supplementing the work that literary scholars have already done. It should be celebrated that literature is gaining a new way to be understood instead of fearing it.

The role of AI and technology in the humanities is a much disputed one mostly because of the human urge to fear change. All we have known is the separation of these fields, and it seems overwhelming that now they can combine. This project serves to illuminate the intersection between the two disciplines and how we can come to look at novels and other texts in a new way that might make us question why authors write the way they do. Could a book become a bestseller without following a classic plot structure? Are authors aware of the models they follow as they write? Will new plot structures emerge? Is it a bad thing that most best-sellers follow a similar pattern or does it speak to the innovation of classic ideas? All of these questions lie at the crossing point of literature and technology, and can only be answered by the two working together in the future.

Acknowledgements

Archer, Jodie, and Matthew L. Jockers. *The Bestseller Code: Anatomy of the Blockbuster Novel*. New York, St. Martin's Griffin.
"Barnes and Noble." https://www.barnesandnoble.com/. Accessed 14 December 2021.
Chun, Jon. "SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs." *arXiv. Cornell University*. Accessed 14 December 21.
Mani, Inderjeet. "How AI is revolutionising the role of the literary critic." *Aeon*, 6 December 2016, https://aeon.co/essays/how-ai-is-revolutionising-the-role-of-the-literary-critic. Accessed 14 December 2021.
Mckenzie, Paige. "Exploring Jane Austen novels with text analysis." *UT Austin*. *Github*, https://p-mckenzie.github.io/2018/01/11/Jane-Austen/. Accessed 14 December 21.
Phillips, Stephen. "Can Big Data Find the Next 'Harry Potter'?" *The Atlantic*, 12 September 2016, https://www.theatlantic.com/technology/archive/2016/09/bestseller-ometer/499256/. Accessed 14 December 2021.
Reagan, Andrew J., et al. "The emotional arcs of stories are dominated by six basic shapes." *SpringerOpen Journal*, 2016. *EPJ Data Science*. Accessed 14 December 21.
Silge, Julia. "If I loved Natural Language Processing less, I might be able to talk about it more." 2017. *Open Data Science*, https://opendatascience.com/if-i-loved-natural-language-processing-less-i-might-be-able-to-talk-about-it-more/. Accessed 14 December 2021.