Text Mining Report

20011262

INST0073 Natural Language Processing and Text Analysis

Dr Andreas Vlachidis

Introduction

This report accompanies a Jupyter Notebook which computationally identifies the main characters of Pride and Prejudice and then visualises the polarity of their associated sentiments within the book. The first part is an explanation of the NLP pipeline's design and the second is an analysis of the results.

NLP Pipeline

1. Processing the text

After retrieving the text from Project Gutenberg (Weeks, Hurst, & Meehan, 2013) using the urllib library's request function (Python Software Foundation, 2025), unnecessary formatting characters were removed in addition to pre and post text extraneous to the book's narrative. It was then processed using Spacy's en_core_web_sm pipeline (Explosion, 2013). Benefits of this pipeline include: it is designed for English language; small enough to run on most machines; and is capable of Named Entity Recognition (NER) which is vital to our task.

2. Detecting the Main Characters

While iterating through the named entities labelled "Person", automatically carried out by the Spacy pipeline, each entity was added to a dictionary and their frequency in the book was recorded. The four characters with the highest frequency were then identified as main characters.

3. Bennet

A view of all the characters and their frequencies revealed that "Bennet", identified as the fourth main character, was actually the surname of multiple characters, including two other main characters. To investigate further, a random sample of 4-token spans including "Bennet" was displayed revealing that most of its usage referred to either Mrs or Mr. Bennet. It is likely that the full stop after their title caused Spacy to interpret them as separate tokens. If this was the case for every occurrence of "Bennet" an edge case for detecting tokens preceded by a title could have been added. However, there were instances of "Miss Bennet" which could have referred to either Elizabeth or Jane (two other main characters) and as marital status can change, a coreference resolution

would be necessary to handle all cases. This was deemed outside the scope of the project, and 3 characters proved to be sufficient data to draw meaningful insights.

4. Sentiment Analysis

The sentiments associated with each character were computed as follows: Iterating through the book's sentences, whenever a main character was referenced the sentiment of the sentence was calculated using NLTK SentimentIntensityAnalyzer (NLTK Project, 2023) and Textblob sentiment polarity (Loria, 2025). Two frameworks were used to make comparison possible. Analyses were conducted on the whole sentence, expecting this to provide sufficient context for accurate scores. For each such sentence, a triple containing the sentence number and sentiment scores (number, NLTK sentiment, Textblob sentiment) was stored alongside the character.

5. Visualisation

First, summary statistics were presented in a table including mean, maximum, minimum and standard deviation for both sentiment analysers as well as frequency for each character.

Second, graphs were produced using each sentence as a data point, its number as its x-coordinate and its sentiment as its y-coordinate. This allowed lines to be plotted comparing the sentiment associated with characters across the book and comparing their sentiment according to different analysers.

Findings

Summary statistics

Main Characters	Elizabeth	Darcy	Jane	Bennet
no. of mentions	628	414	293	325
NLTK mean sentiment score	0.197	0.187	0.214	0.179
NLTK max sentiment score	0.976	0.979	0.977	0.944
NLTK min sentiment score	-0.968	-0.948	-0.952	-0.911
NLTK st dev of sentiment score	0.498	0.48	0.52	0.49
textblob mean sentiment score	0.115	0.103	0.092	0.084
textblob max sentiment score	1.0	1.0	1.0	0.875
textblob min sentiment score	-1.0	-0.65	-0.609	-0.85
textblob st dev of sentiment score	0.253	0.249	0.246	0.254

Figure 1

Comparison of Frameworks

NLTK (NLTK Project, 2023) and Textblob (Loria, 2025) take a similar approach to calculating sentiment, both leveraging a database of sentiment scores for words and applying simple rules for modifying words to recognise negation and intensity. However, the numerical values they arrive at differ significantly. From figure 1 we can see that on average NLTK assessed sentences more positively and, despite Textblob having a greater range, NLTK had a consistently higher standard deviation. This is evident from a comparison of their assessments of sentences containing "Elizabeth" throughout the book, in figure 2.

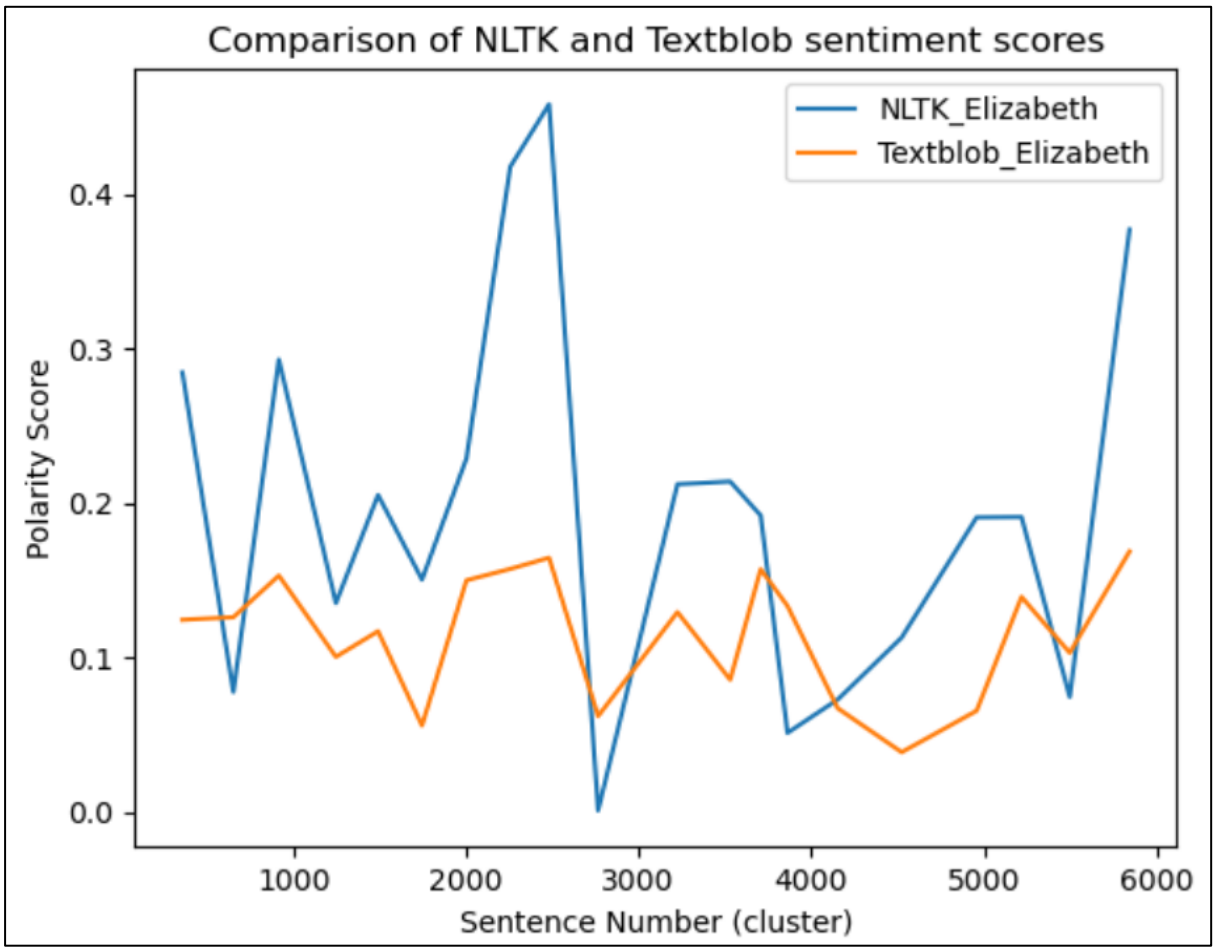

Figure 2

The lines tend to move in the same direction between points implying they interpret sentences similarly however the blue line moves a greater distance meaning NLTK scores have greater magnitude. There is nothing to indicate one framework is superior to another so NLTK will be used hence to present the differences between points more clearly.

Number of clusters

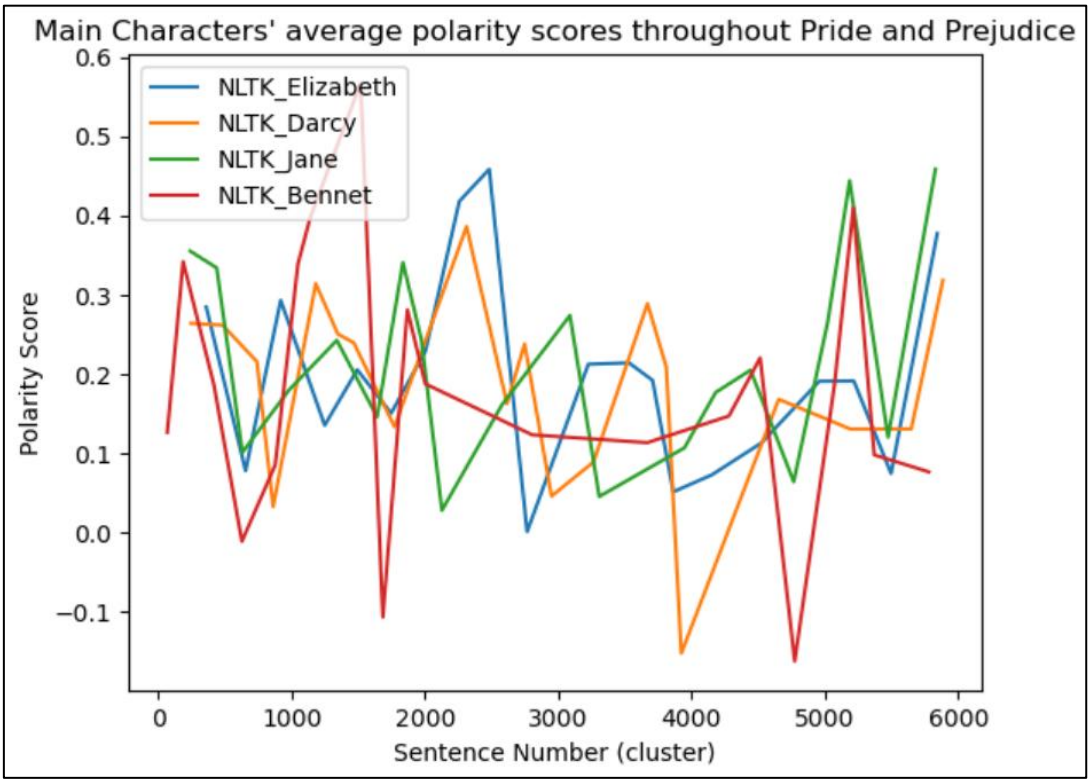

Figure 3

Figure 3 displays the scores of the 4 main characters from sentences in which they are mentioned, divided between 20 clusters. Deciding on the number of clusters to use balances a trade-off between expressiveness and interpretability. A greater number of clusters provides a more granular view of the change of sentiment across the narrative but also produces a more erratic trend, making the graph harder to read. When comparing multiple lines, fewer clusters are favourable. This graph was chosen because it makes the erroneous nature of "Bennet" more visible. The standard deviation for both scores is high but not significantly higher than that of the other characters. The changes between neighbouring datapoints are higher, visible in the sharp and steep jumps. An appropriate statistic was calculated to corroborate this. Averages of the gaps between datapoints are presented in figure 4.

```
Elizabeth = 0.02911130725016441
Darcy = 0.03183033891571489
Jane = 0.03009642816596252
Bennet = 0.06607726196657031
```

Figure 4

This is likely occurring because it is changing which character it references, and so this line will be removed from subsequent analysis.

Narrative Analysis

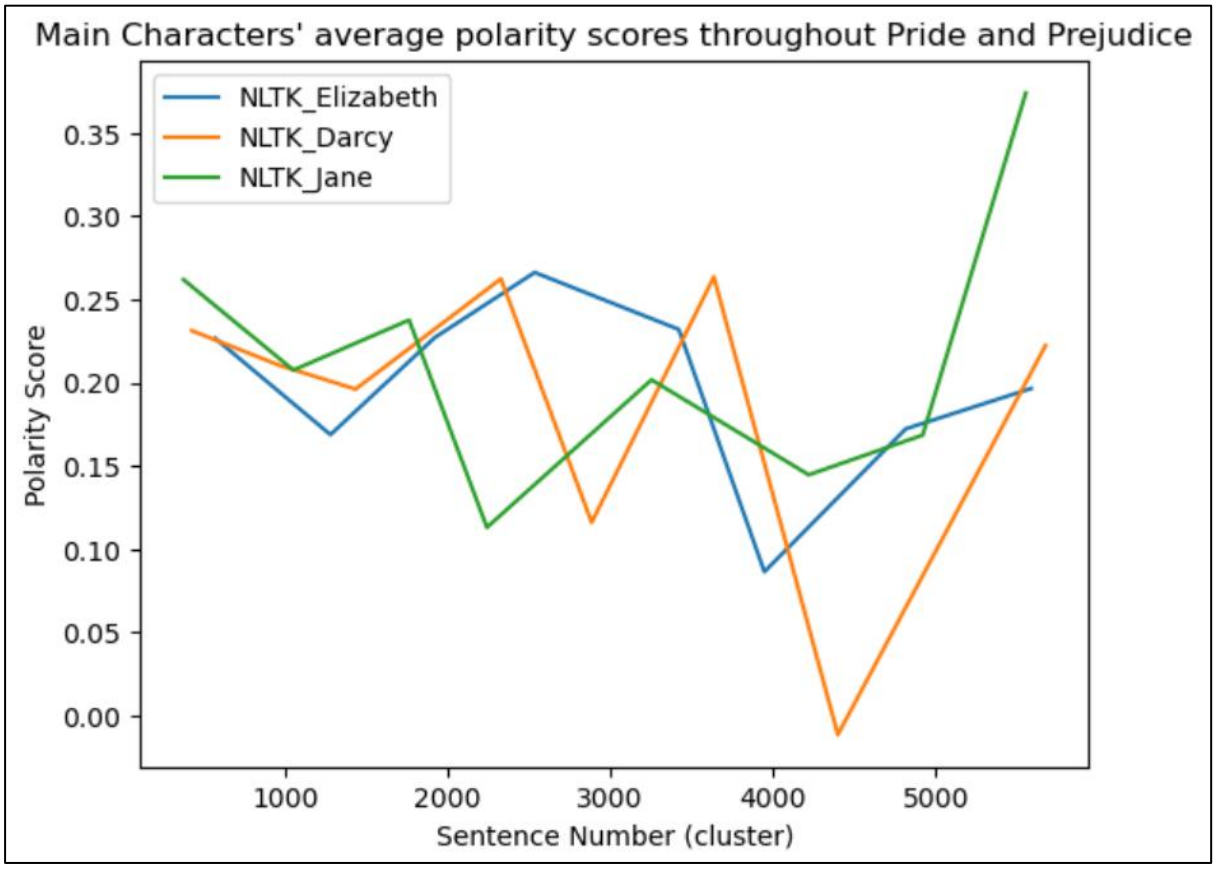

Figure 5

Figure 5 displays the scores divided into 8 clusters. In this graph the trend of the narrative is more evident. The characters begin congregated at a positive score. There is a sharp decline approximately ¾ of the way through, particularly for Jane and Darcy, indicating they face adversity at this point. Finally, all scores rise, indicating a generally happy ending. Since these are the three main characters, it is fair to assume where they behave in a similar fashion reflects the book as a whole, which would mean the climax is roughly ¾ of the way through, preceding a happy ending.

The graphs and summary statistics both show relatively little variation between characters' scores which is not surprising for main characters as they probably interact with each other frequently and appear in similar scenarios. The scores tend towards mild positivity at 0.2 and 0.1 for NLTK and Textblob respectively. This aligns with what is expected of this kind of book and setting, and indicates that words likely to elicit extreme negativity (e.g., violence) are rare or not present.

Limitations and Improvements

Published in 1813, the English used in Pride and Prejudice is contemporary to the time. This probably limited the accuracy of Spacy's pipeline which is designed for modern

English. However, the focus of this kind of research is to develop text agnostic approaches, so general improvements will be discussed.

More comprehensive cleaning and parsing of the text would improve the accuracy of tools like NER, as well as regular expressions to pick up names obscured by grammatical features, like the apostrophes in possessives. More advanced techniques like coreference resolution to identify pronouns relating to a character would also increase the sample size, improving the reliability of results.

The aim of the graphs displayed was to show how sentiment changed over the course of the book, however sentence number is not a great metric to use, because it is difficult to contextualise. A more interpretable approach would be to use (clusters of) chapters after developing a method to detect their starts and ends.

One may argue that one sentence is insufficient context to accurately assess the sentiment at that stage in the book, or that including the surrounding sentences provides valuable information. For each datapoint, including x sentences before and after in the sentiment analysis might provide more insightful results but will become more computationally intensive as x increases.

Scalar sentiment polarity scores are used because they are most widely researched and have proved effective in downstream tasks (Tian, 2018). However, for the purpose of analysing a story or its characters, valuable information is lost by reducing sentiment to one dimension, for example a sad but safe character may be scored similarly to a happy character in peril. Sentiment vectors that capture the spectrum of emotions conveyed in literature could lead to more insightful analysis. Mohammad & Turney (2022) created a lexicon of emotion scores for English words that could be leveraged to implement a similar approach to the one discussed in this report, but with multiple axes (e.g., anger and fear) instead of sentiment polarity. A 3-dimensional plot of this kind could provide much deeper insight into a story's nuance and complexity. Currently these techniques are typically less accurate than sentiment polarity but could become more reliable in the future.

References

Explosion. (2013, October 16). *spaCy Models*. Retrieved from spaCy : https://spacy.io/models/en#en_core_web_sm

Loria, S. (2025, January 13). *Blob Classes*. Retrieved from Textblob : https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment

Mohammad, S., & Turney, P. (2022, August). *NRC Word-Emotion Association Lexicon*. Retrieved from saifmohammad.com: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

NLTK Project. (2023, January 2). *nltk.sentiment.SentimentIntensityAnalyzer*. Retrieved from NLTK Documentation: https://www.nltk.org/api/nltk.sentiment.SentimentIntensityAnalyzer.html?highlight=sentimentintensity

Python Software Foundation. (2025, April 30). *urllib.request — Extensible library for opening URLs*. Retrieved from The Python Standard Library: https://docs.python.org/3/library/urllib.request.html

Tian, L. L. (2018). Polarity and intensity: the two aspects of sentiment analysis. *Grand Challenge and Workshop on Human Multimodal Language* (pp. 40-47). Melbourne, Australia: Association for Computational Linguistics.

Weeks, G., Hurst, J., & Meehan, M. (2013, May 9). *Pride and Prejudice by Jane Austen*. Retrieved from Project Gutenberg: https://www.gutenberg.org/ebooks/42671