A Comprehensive Text Mining Analysis of War and Peace by Lev Tolstoj

Federico Grazi

2023-02-19

1 Introduction

War and Peace by Lev Tolstoj has been recognized as one of the greatest masterpieces of modern literature ever written: with its five volumes and over 300 chapters, Tolstoj takes us on a journey in both time and spirit by describing the Napoleonic Wars and the Russian society that was being attacked. While describing in such unique way the experience of the war, Tolstoj manages to take a deep dive into his characters' consciousness and emerges with some of the most fascinating depictions of human sentiments and feelings.

The aim of this project is neither to sum up all the 2000 and plus pages of the novel (which would be rather difficult, even if one were to use statistical techniques), nor to capture the profound historical and personal meaning of the work, which can only be truly understood through a long and immersive reading of the full book. The aim of this project, instead, is merely to try and capture as much of the intricate plot and to uncover the underlying structure of plot and subplots that define War and Peace.

2 Methodology

As for the methodology used in this project, the main tools used were text mining tools offered by the `tidyverse` and the `tidytext` libraries, as well others statistical techniques - described below - for more complexes analyses. Both sentiment and topic analysis are discussed and confronted with a final cluster analysis: the aim was to confront the Latent Dirichlet Allocation (LDA) with various techniques of clustering to see which would be able to estimate and better uncover the various plots. LDA is a method for unsupervised classification of documents (in the sense the we have no *a priori* knowledge concerning the groups), which is based on two concepts: each document is a mixture of topic, and each topic is a mixture of words; as such, LDA will return a β per-topic-per-word probability (how likely is that a term is part of a certain topic) and a γ per-document-per-topic probability (how much of this document is made of a certain topic in terms of percentages). This allows documents to

"overlap" each other in each document, rather than being separated into discrete groups, as cluster analysis would do. Cluster analysis is, also, an unsupervised multivariate technique that partition our dataset into groups, called "clusters", that are made of homogeneous element within them, and are heterogeneous between them, such that each cluster is distinct and exhaustive. We'll take into consideration both hierarchical clustering and partitive clustering, to make up for the weaknesses of each method. We'll use sentiment analysis and the `NRC` sentiment lexicon, given its wide range of sentiments, to determine the principal characteristics of our result and for an initial descriptive analysis of the book.

3 Dataset Description

The dataset was made available by *The Gutenberg Project* and was imported in `R` using the `gutenbergr` library, which conveniently allowed me to retrieve the dataset using only the `gutenberg_metadata()` function. The novel is fully translated in English, and omits the entirety of the French-language part; names are also fully translated. Once downloaded, the resulting output was the a dataset with one line of the novel per-row. Firstly, I created an index for the books and chapters of the novel: I did so by using the `stringr` package, and by recognizing in the text whether in the row there was either the string marking the start of a new book, or of new chapter. Once I had an variable for volume, book and chapter, the dataset was grouped by chapter and pasted together, in such a way that the format was one-row-per-chapter, making the tokenization process much simpler for keeping track of which word appeared in which chapter.

4 Explanatory Data Analysis

In this section I will outline at some interesting statistics, concerning both the whole novel and its three main characters in the novel, Pierre, Natasha and Andrej. In doing so, we will be able to reach an approximate overall understanding of the novel's arguments, as well as itsm sentiment trends.

4.1 Descriptive Novel Analysis

The novel is composed of 4 story-related volumes and 2 epilogues, with a total of 17 books and 367 chapters. The novel is fairly unevenly distributed by token count as can be observed in Table 1.

The volumes with the least chapters and words are the epilogues, which are rather short compared to others volumes. The main corpus of the text can be found in the middle two volumes, which have higher number of chapters and words. By looking at the overall by volume, displayed in the next box-plot, we can have a clearer view of how the book is structured. In the first volume we have progressively longer chapter and books (word-wise), but they decrease as we read the novel, with the last three books forming the shortest books in the

	Chapters	Total Word Count
volume 1	68	48316
volume 2	99	50911
volume 3	98	55697
volume 4	74	31618
epilogues	28	14033

Table 1: Count by volume

novel. This trend can be compared with the dashed line which represent the mean.

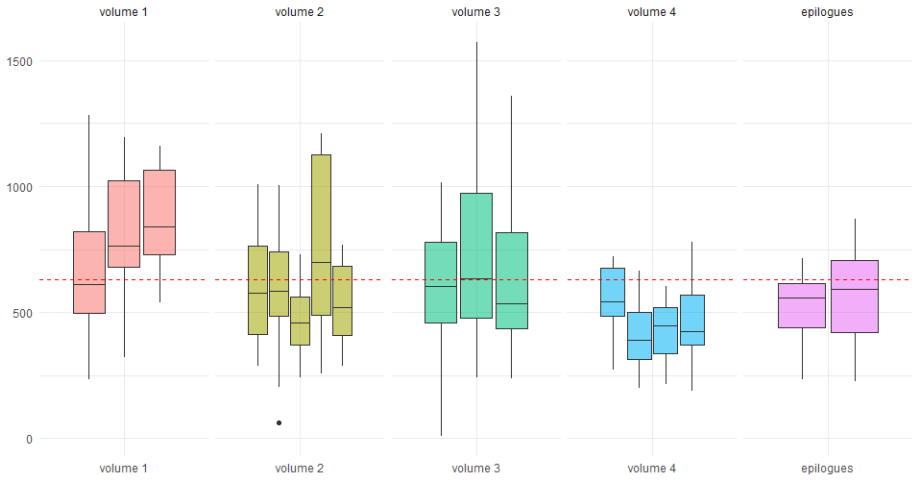


Figure 1: Boxplot of the token count distribution

If we look at the token frequency, we can see the same trend observed in the boxplot, and at the same time take a look at the *term frequency*, introducing the most common words. As Figure 2 shows, many words occur quite rarely during the novel, with only a few appearing more frequently: a common pattern in any text corpus. In the first three volume we have at least 6000 less frequent words, while in the others we have much fewer than 5000. So we would expect the most common words to be related to the first three volumes.

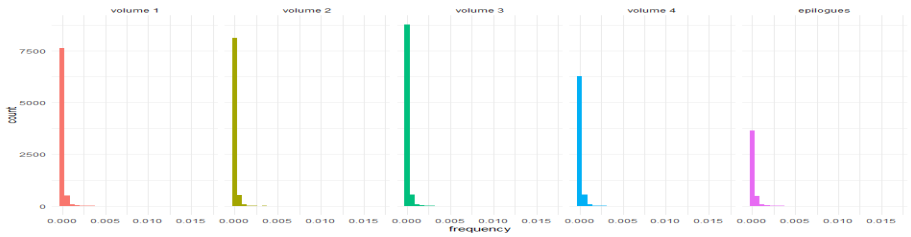


Figure 2: Term frequency by volume

In Figure 3, immediately we are able to spot the main characters: Pierre, Natasha and Andrej (in English, Andrew). In a very coherent way also the words *'prince'*, *'princess'* and *'count'* also appear in the top 10. Coherent, because most of this characters are referred to with these titles, as it was typical in Russia at the time. Rather unexpected is the presence of the word *eyes*: even though Tolstoj is well-known for depicting some of the most beautiful description of all literature, I wasn't expecting such a specific word to be found so frequently. This gives us an interesting insight as to how much Tolstoj values eyes in its writing. Now, we will further analyse the contexts in which the word is used, before returning to our analysis of the whole novel.

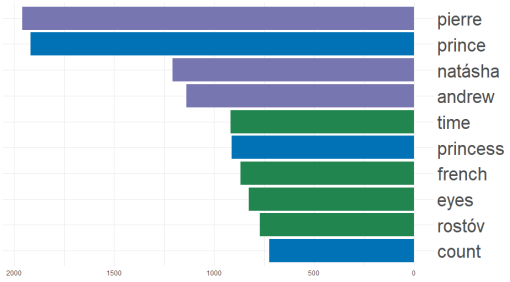


Figure 3: Most common words by count

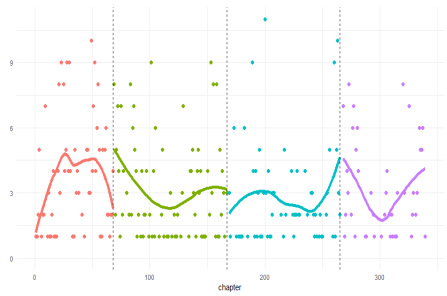


Figure 4: Count of word 'eyes' by chapter

Bigram that have 'eyes'	n
eyes fixed	20
black eyes	16
open eyes	14
blue eyes	12
downcast eyes	12
frightened eyes	12
glittering eyes	11
shining eyes	11

Table 2: Top bigram with eyes

First of all, the epilogues have a significantly lower number of the word 'eyes' (only 31 of the 826 total appearances): in Figure 2 I chose to exclude the last

chapters in order to have a clearer understanding of the trend. As expected the words most often associated with eyes are descriptive words, and by looking at the trend we can see that at the beginning and at the end of the first four volumes (which respectively represent the part of the novel in which the characters and the battles are described) there is a quite high count of; so it seems safe to assume that Tolstoj uses eyes as an important mirror of both the experienced sentiments (e.g. *frightened eyes*) and as an distinctive feature of an individual.

Turning back now to the main analysis, we'll observe the most common and most distinctive words based on volume, and introduce the `bind_td_idf` function, which computes the `term frequency * inverse document frequency` metric, as it multiplies the already seen term frequency with the *idf* value, which is the logarithm of the inverse ratio of the total number of document over the total number of document in which the term appears.

Figure 5: Most common and distinctive words by volume

The three main characters appear at least once in each volume. It is noteworthy, however, that in volume 3 and 4 the terms *French, Napoleon* and *army* are more frequently used, since the 1812 French Campaign is at its climax and the story largely dwells on it. In the second figure there are the most "discriminant" word for each volume, and it's not surprising that they are mostly proper nouns of characters who have a significantly part in the novel, even if only in a short period of time. We can see the General Kutuzov, who leads the Russian Army in the 1805 offensive in Austerlitz and is later called into battle again to combat Napoleon after the Fire of Moscow; likewise, we have Karataev in Volume 4: a

Jesus-like figure whom Pierre meets while imprisoned. The only volume which has different words is the epilogues; in fact the last chapters are a long essay-like philosophical reflection about Time and History, without any more story development.

We will now take a brief look at the overall sentiment count in the novel using, as mentioned above, the `NRC` lexicon, a more structured analysis will be presented later.

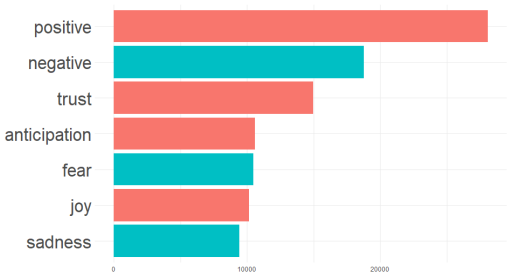


Figure 6: Top sentiment by count

Considering the novel narrates two distinct wars in the 1800s one might expect perhaps more "negative" sentiments to rise to the top. Nonetheless, it's clear form Figure 6 that positives words are much more present than negative ones. Unexpected as it might be, there are two alternatives: either the "peace" aspect of War and Peace outweighs the "war" aspect, or the novel has an overall positive tone, despite the tragic events it depicts. We will discuss this trend shortly.

In order to better understand the distribution of sentiments throughout the novel, we can look at Figure 7, which gives us an overview of the predominant sentiments in both chapters and books. As Figures 7a and 7b suggests, most chapters are positive, with only a small portion being negative: this could be influenced by the fact that in the lexicon is made up by 40% of positive and negative words, as Table 3 shows.

In our up-coming analysis, we will be using weights that value more rare sentiments over more common ones. As of now, I've provided Figures 7c and 7d to make room for other sentiments to prevail and better read the novel sentimets.

Table 3: Sentiment in `NRC` lexicon

sentiment	n	perc	cumulative
negative	3324	24 %	23.9 %
positive	2312	17 %	40.5 %
fear	1476	11 %	51.2 %
anger	1247	9 %	60.1 %
trust	1231	9 %	69 %
sadness	1191	9 %	77.6 %
disgust	1058	8 %	85.2 %
anticipation	839	6 %	91.2 %
joy	689	5 %	96.2 %
surprise	534	4 %	100 %

With the help of the second two graph we can better capture the increase in the sentiment 'trust' that becomes more importante in the last chapter and the last book; in fact, this reflects how the main theme of Tolstoj's philosophical soliloquy is the hope for a better future for humankind and a very powerful

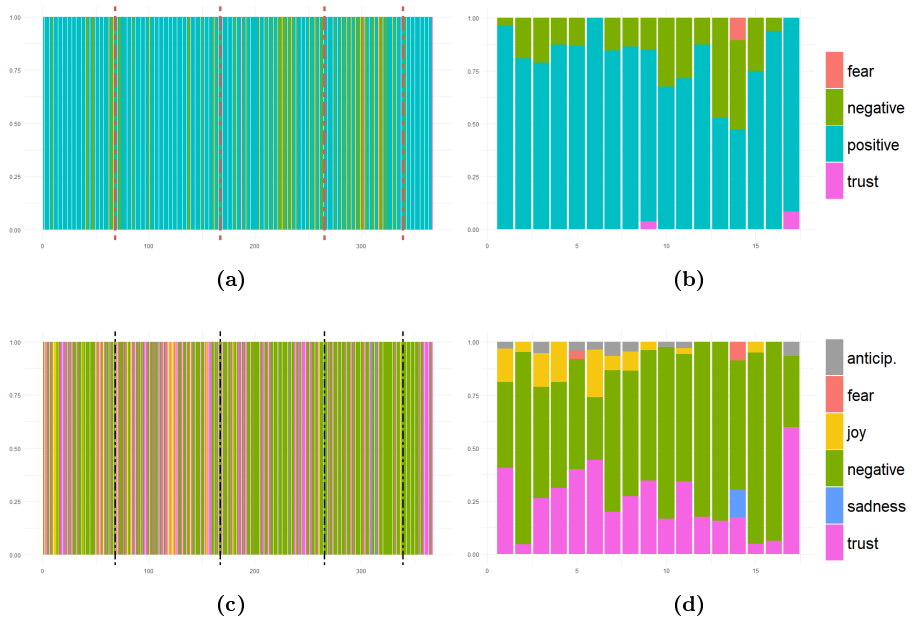


Figure 7: Sentiment distribution with 'positive' terms (a and b) and without (c and d)

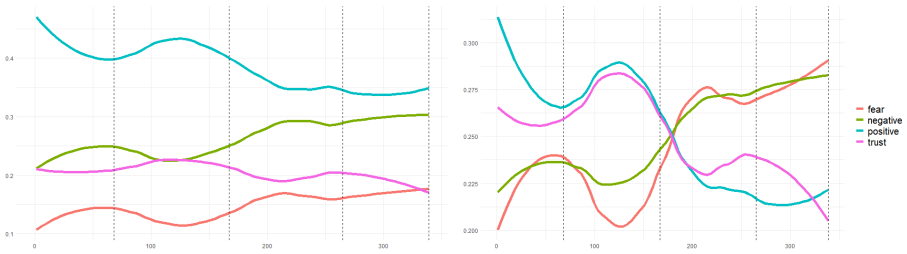
anti-war tone.

In a similar way we can divide the novel into two parts: the first two volumes have higher occurrences of trust and joy, while the last two become generally more negative; there are even three of predominantly 'sad' chapters right before the end of the volumes, as Andrej is dying and the novel takes a tragic turn before resulting in a happier ending.

If we take a closer look at the count trend as the novel progresses, we can see the "sentimental shift" that the novel undergoes at the end of the first two volumes. In Figure 8a, in which the epilogues have been removed, it is clear how much the lexicon composition influences the results: following the same color-coding as before, the chapters seem to be made up of the most prevalent sentiments that are in Table 3, but by dividing the count by the total number of appearances of a sentiment, we can look at the composition also in terms of how much a chapter is important for that feeling in Figure 8b.

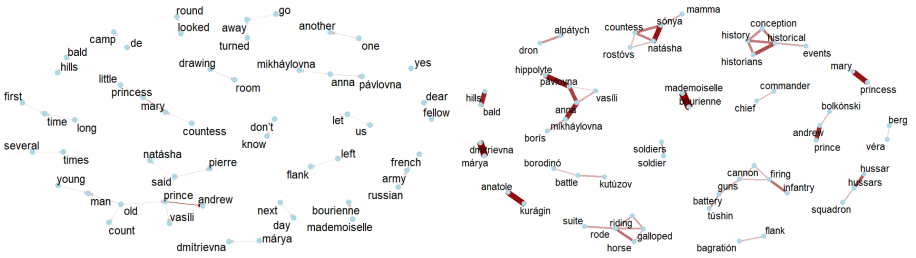
Initially, Russian Society is at ease, and positive sentiments prevail; but not for long, because the Battle of Austerliz is about to begin. The second volume focuses more on characters. Then, Andrej is brought back into the army; the 1812 War begins and lasts for two volumes, with small, more charming chapters, Volume 3 techniques finishes and, one can glance an increase in positive terms at the end of the volumes, signalling a happy ending.

Before moving on, we'll briefly look at words networks created with both `pairwise_count()` and `pairwise_count()` from the `widyr` package.



(a) (b)

Figure 8: Sentiment trend by chapter. Not weighted (a) and weigthed (b)



(a) `pairwise_count()` (b) `pairwise_cor()`

Figure 9: Words network

In Figure 9a it's plotted the network by counting each time two words are paired together. In Figure 9b the network is created by correlation, computed as "co-presence" or "co-assence" of the same word in the same chapter. The word *prince* is highly linked to Andrej, since it's the title by which Andrej is most often referred to. Many names are linked to their charachters' surnames, such as Anatole Kuragin, Natasha Rostov and Anna Pavlovna. Terms are also linked by argument: the ones used in the epilogue (history, historians, etc) are all grouped together, as are the words regarding battles or horse-riding. In Figure 9b we can see the small but significant cluster made of Vasili, Hyppolyte (his son) and Anna Pavlovna, a lady who often hosts a salon in which Vasili, Hyppolyte and Boris are frequent guests.

4.2 Overview of the Main Characters

Before proceeding to the topic analysis, we will outline some statistics concerning the main characters, in the hopes of gaining an even better overlook of the novel's structure. Firstly, it is useful to outline the trend as shown in Figure 10, in which the epilogues have been excluded.

As we can see the main three characters are present throughout the whole novel; interesting enough, each character is more prevalent in some volume and less in others; for example, the three members of the Rostov family are rather

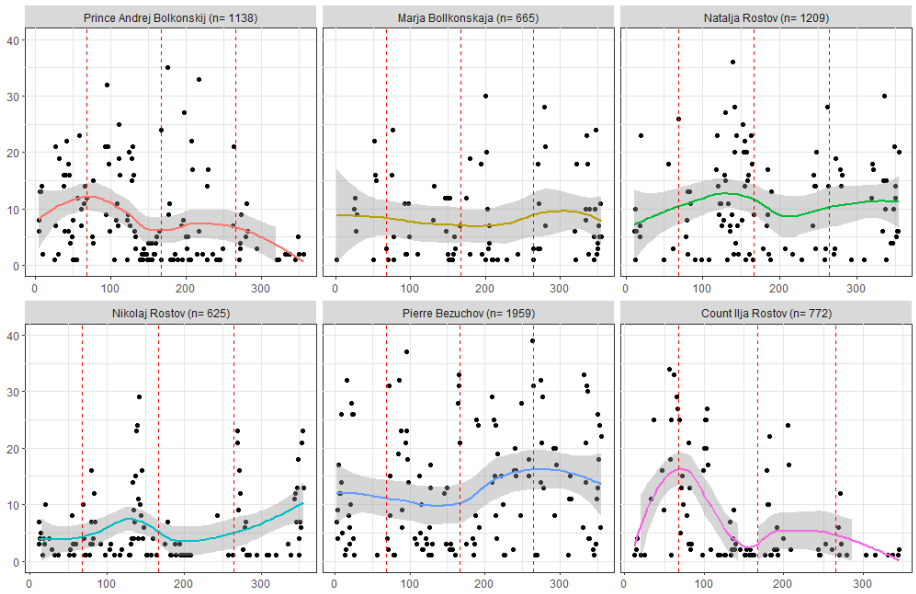


Figure 10: Trend by count of the top six characters

absent in volume 3. Seemingly, Natasha is mainly present in Volume 2, where more than 50% of her total appearances occurs, while almost 70% of Andrej's total appearances occurs in the first two volumes. From this, and from the previous graphs, we can acknowledge that second volume is mostly protagonist-based, focused on more positive and trust-worthy events and dialogues. An interesting result from these graphs is that Tolstoj's two alter-egos, Andrej and Pierre, are equally divided throughout the novel: Andrej is highly present in the first two volumes and Pierre in the last two. Since Tolstoj didn't want to fall under the cliche according to which an author concentrates all of his personality into one character, he attributed his former personality to Andrej, and his new, "reborn", personality to Pierre, which is the one of the two that survives and reaches the end of the novel, with Andrej's death symbolizing the death of his former personality: this is one of many majestic literary tricks that Tolstoj pulls in War and Peace.

We will now focuses on which sentiments are most often associated with each character by analyzing the sentiment of the most frequently paired words with their names. By doing so, and dividing again for the total occurrences of the sentiment within the novel and then dividing again for the total number of sentiment count in each chapter, we will have an outline of the composition of each chapter.

Andrej is associated with more positive sentiments only in the second volume, but as the volume ends the sadness and fear components become increasingly predominant; the 'joy' sentiment has a sharp decrease in the chapters regarding

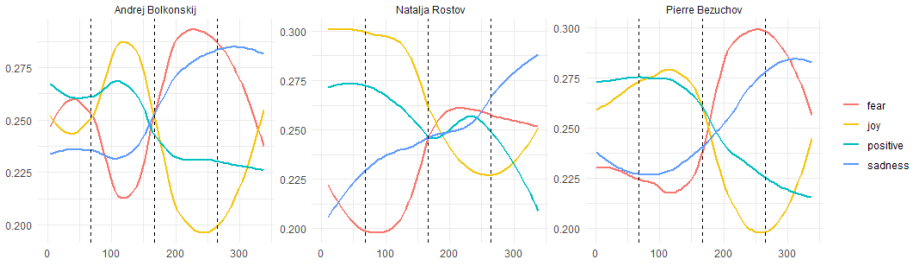
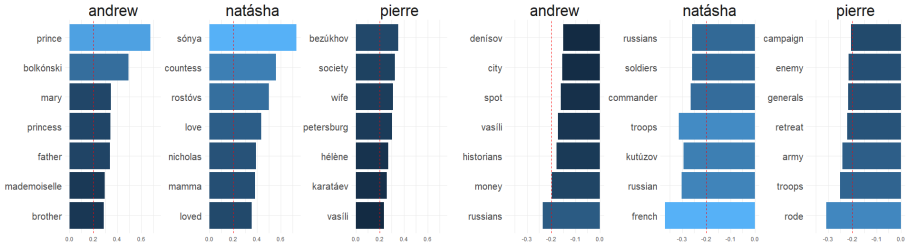


Figure 11: Sentiment trend by characters

the Russians' losses in the war, and the 'sadness' sentiment is the strongest as he dies from his wounds. As for Natasha and Pierre the first two volumes are mostly positive; the most interesting fact is that the predominance of the sentiment 'fear' drops for both protagonist as Volume 4 ends, but there is still a rather high percentage of sadness words, suggesting that even if the war ends and fear slowly decreases, a sad mood still lingers in the last chapter, despite an increase in joy as many of the charachters manages to solve their tragedies.

If we take a closer look at the most correlated words with the main characters, it's possibile to see which words are usually associated with them: Natasha is significantly correlated with the members of hers family meanwhile Pierre is correlated with Saint Petersburg society, especially the members who have an important role to play in his story.



(a) Positive correlated words **(b)** Negative correlated words

Figure 12

Looking at Figure 12b we can see how both Natasha and Pierre have a high quantity of negative correlated war-related terms. From this we can evince that Natasha is a protagonist who, during the novel, stays close to her family and isn't involved in the war; similarly, Pierre is close to the Russian society and also far removed from the battle events. No such clear conclusions can be derived in Andrej's case.

5 Result and discussion

The last section in this project will try to complete the idea stated in the overview, by supporting it with more robust results provided by topic analysis. Topic analysis, as already mentioned, is a type of unsupervised learning technique which tries to generate topics by the word co-occurrences; it assigns one or multiple topics to each document - in our case, chapters.

5.1 Topic Analysis

Using a useful guide by MNidhi in Rpubs that evaluates the optimal number of topic using the empirical method, I determined an optimal number of topics: looking at the charts in Figure 13, 8 was a rather cautious choice.

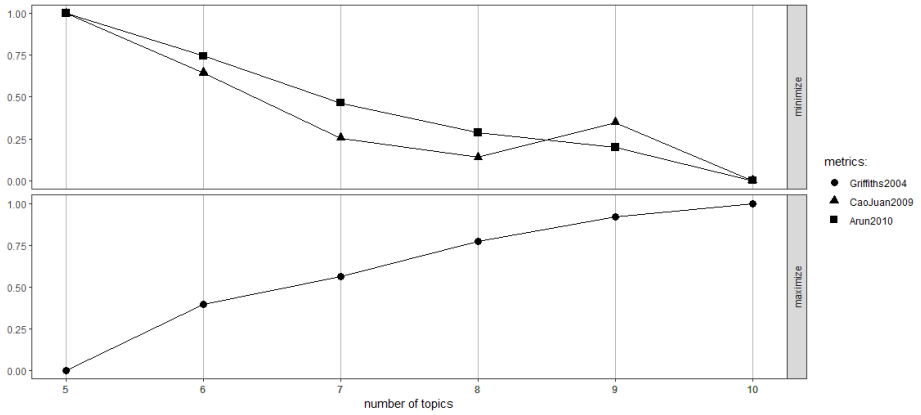


Figure 13

Thanks to the function `tidy()` it was possibile to extract the *gamma* and *beta* matrix probabilities for the various topics. In Figures 14 and 15, we can see both what the topics are made of (which words defines each topic) and the presence trend (when they occur the most throughout the novel).

Before analyzing the word composition of the topics, it is important to note that I personally came up with the names found inthe figures, as they are just a quicker and easier way to refer to a topic with some better context. Interestingly enough, there are some topics with very high-beta probability, such as the "Pierre story", which mostly concers what Pierre's actions; in a similar way, Andrej's and Rostov Story are topics which mainly concern those characters. In general, there are 3 topics about war, one about general battle elements (officer, soldier, horse), one with more specific elements (French Invasion) and one where the topic is more focused on the characters experience in war. With 3 topics about war and 3 about charachters, the novel appears to be is quite balanced. Finally, there are two more abstract topic, one related to the epilogues and one about the Russian saloon.

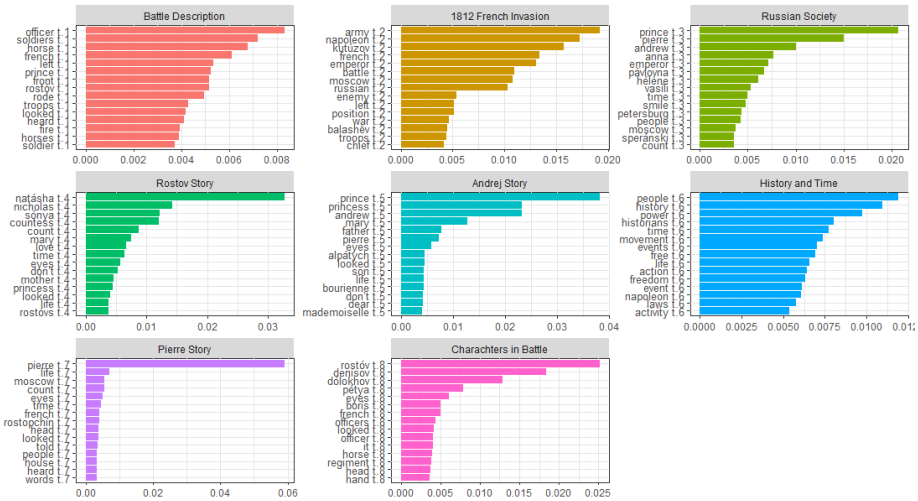


Figure 14

By looking at where in the novel the topic is more prevalent, I managed to come up with the different titles.

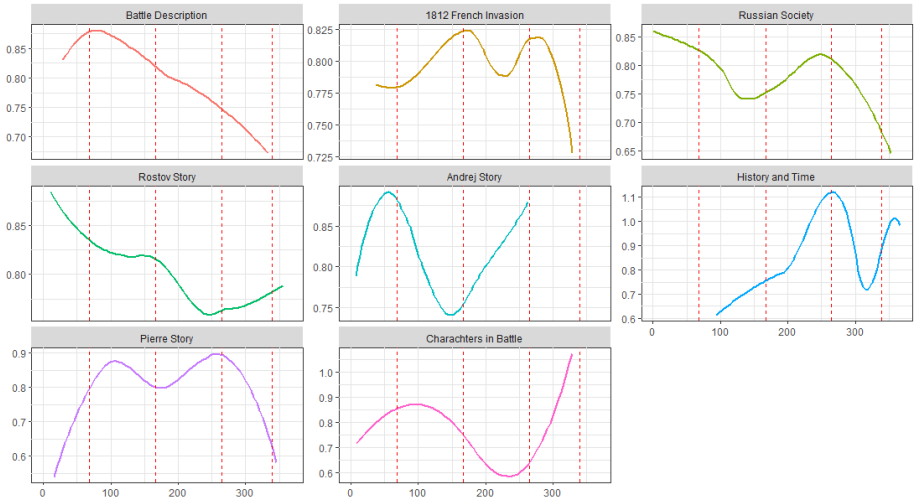


Figure 15

Some obvious findings are that Andrej's Story stops at the third volume, as he dies early on in Volume 4. The epilogue mainly contains the topic "History and time" and the "Rostov's story" as Tolstoj both reflects upon History and how the Rostov's story ended. A surprising result is that there is also a profound reflection at the start of volume 4, an aspect of the book clearly highlighted by

the analysis, which – on a personal note – I did not recall.

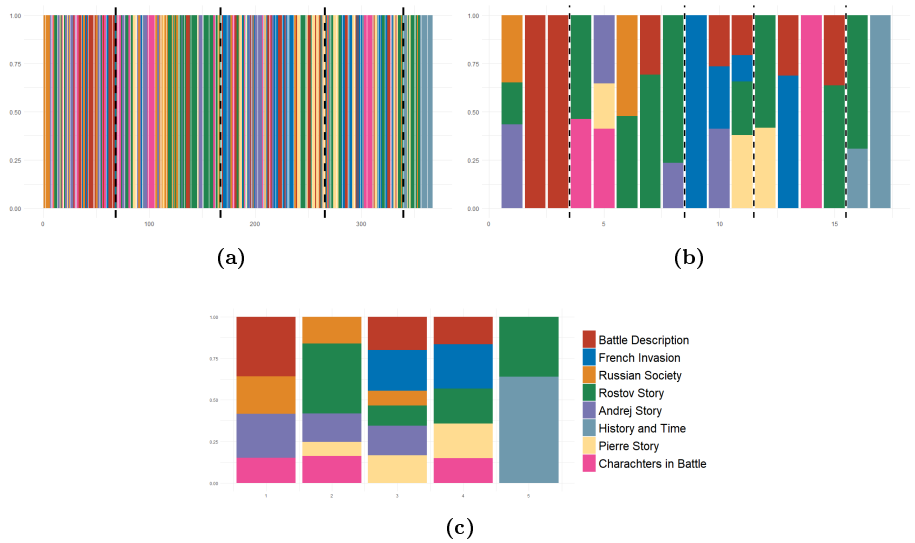


Figure 16: Topic composition by chapter (a), by book (b) and by volume (c)

As Figures 16a and 16b shows, the novel starts off by talking about the Russian society (with its infamous first French line about the Italian situation, *"Eh Bien..."*) and introducing the Rostov's and Andrej's story. Later in the first volume, the author starts talking about battles; the second volume is mainly about Rostov's story, and from previous analysis it's possible to add that is a positive and joyful volume, mainly focused on Natasha. In Volume 3, more attention is paid to the French invasion and Pierre's part in the story, as he is imprisoned while back in Moscow.

In the next table, we can see the number of chapter for which a single topic has a *gamma* probability higher than 0.15 and in which volume they are mostly present. We can see how the how the Rostov Story is widely spread and covers almost 30% of the novel even though has almost 50% of its presence in the second volume (49 chapters with that topic), which is a character volume, as we have already established. The second to last prevalent topic is the Battle one, with only 55 chapter out of 365 having at least 0.15 gamma probability.

5.2 Cluster Analysis

In this last section we'll briefly have a look at some cluster analysis implementation before presenting our the conclusion. My first concern is clustering characters and then the chapters themself. As mentioned previously, the sole purpose of cluster analysis is to create groups of our original data (in these cases characters or chapters) in such a way that they are similar within each cluster, and heterogeneous between them.

topic	total	coverage	max	volume
Rostov Story	101	27.7 %	49	2
1812 French Invasion	81	22.2 %	37	3
Andrej Story	75	20.5 %	26	2
Battle Description	71	19.5 %	27	3
Pierre Story	69	18.9 %	22	3
Russian Society	63	17.3 %	24	2
Charachters in Battle	55	15.1 %	21	2
History and Time	39	10.7 %	17	5

Table 4

5.2.1 Hierarchical Clustering of the Charachters

Firstly, I've chose 17 of the most relevant characters, which features some of the most common characters and some personally added ones. The model for clustering evaluates the "dissimilarity" of each character by calculating Euclidean distances in a n-dimensional space defined by the words associated with the 17 charachters more than 10 times, where characthers with similar sets of associated words are considered "similar"; this was done with `pairwise_count`. Using the `NbClust` function, it evaluated the `dunn` and the `silhouette` indexes to determine the optimal number of clusters, as shown in Figure 17.

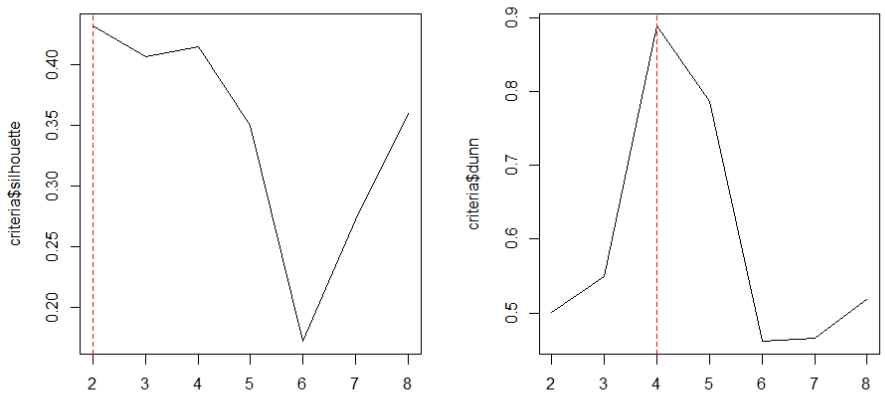


Figure 17: Optimal number of cluster. Red line indicates the maximum

Both graphs suggest that four is a reasonable number of cluster. Using the *Ward method* for clustering and the `hclust` function in `R`, I was returned with

this dendogramm already cut with four clusters of similar charachters.

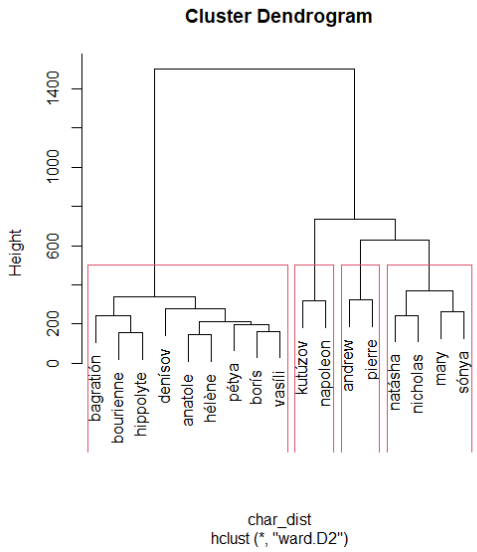


Figure 18: Hierarchical clustering dendogram

Figure 18 shows some interesting results; firstly, there is one cluster with all minor character; in this cluster Helene and Anatole (who are relatives) are grouped together early on. The second cluster features the real-life generals Kutuzov and Napoleon, as both act in the same context; the fourth cluster is made of the Rostov's family members. The third cluster is made of Andrej and Pierre, which is fairly surprising, considering they're part of different stories. However, they are closely linked in terms of their context and the words they are more often associated with.

5.2.2 K-Means Clustering for Chapter

Finally, partitive cluster analysis for the chapter was computed, again confronting the words that appear in the chapters. Considering the high number of chapter in the book and taking into account the limitation that the hierarchical clustering implies, I used the K-Means algorithm to partition the chapters into 7 clusters (optimal number returned by the majority rule using `NbClust` function) and evaluated the top words for each cluster and returned it in the wordcloud in Figure 19.

cluster	chapter
1	134
4	60
3	56
5	44
7	41
6	22
2	10

Table 5: Cluster numerosity

Figure 19: Wordcloud for cluster

As we can see we have a big cluster (1) with much diversity and no predominant words; than we have multiple cluster regarding chapters with protagonist certain characters (as they are the most frequent words): Pierre (7), Andrej (4), Natasha (2) and Napoleon (5). This behaviour was also captured in the topic analysis distribution. If we take a look at Figure 20 there is a clear picture of how the clusters were made. We can look at the gamma probabilities for the topic inside the clustered chapters:

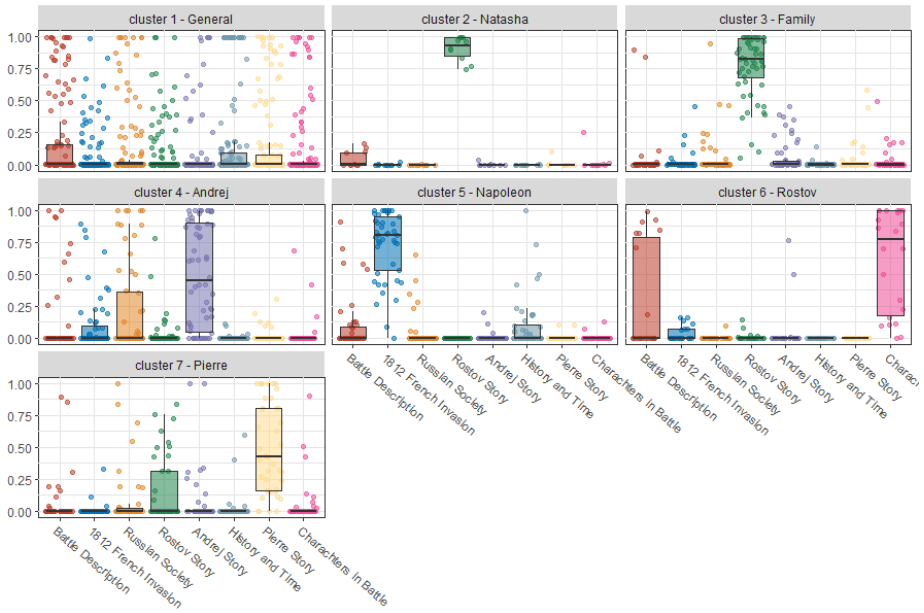


Figure 20

At first, we can see an absence of a cluster with solely the epilogues chapter, as they were grouped in cluster 1 in which were classified all chapters with more vague topic, as the majority of the chapters with gamma probability close to 1 are grouped here. Cluster 2 and 3 are rather similar, but can be interpreted as

the latter being purely a Natasha focused chapter (even if it's a small cluster of size 10, the algorithm decided to divide the two), while the former being more generally situated in the family atmosphere (Andrej is cited and so is Russian Society). Cluster 4 signals that Andrej-focused chapters also includes the Russian Society, French Invasion and Battles; while for cluster 5 Napoleon is also related to the epilogues' argument, as he is often cited as he plays a big role in History.

Interestingly enough, Table 6 shows that the epilogues' chapters, when not assigned to the first cluster, are assigned either to the Family or Pierre's cluster, as the future of both Natasha and Pierre is taken into consideration as the novel ends with them getting engaged; this could mean that the clustering algorithm mistakenly assigned the epilogues as still a part of the novel in which their story unveils. This trend was also captured by the topic analysis, as Figure 16c shows that the epilogues were made up by a small portion of Rostov's Story.

volume	clust 1	clust 2	clust 3	clust 4	clust 5	clust 6	clust 7
1	22 %	3 %	4 %	43 %	3 %	15 %	10 %
2	30 %	5 %	29 %	17 %	1 %	8 %	9 %
3	37 %	2 %	6 %	14 %	24 %	4 %	12 %
4	49 %	1 %	15 %	0 %	23 %	0 %	12 %
5	61 %	0 %	25 %	0 %	0 %	0 %	14 %
total	134	10	56	60	44	22	41

Table 6: Volume's chapters distribution across the clusters. Rows sum up to 100%.

Again, Andrej's chapters stops at the fourth volume, and the Napoleon cluster has significantly more chapters from the last two volumes, meanwhile the first two volumes are clustered more in the sixth cluster. This suggests the recurring result that the first encounters with battles in Volume 1 and 2 are more related to the characters' personal points of view than to a military one. Later on in the novel, how're, the focus shifts to the generals' points of view, as real-life events take up more space.

5.3 Conclusion

The conjoint use of both topic and cluster analysis results in a distinctive way of determining the chapters composition throughout the novel. It has become clear how the war related topics and strongly entangled with the characters story. These two techniques combined managed to delimitate the boundaries in which the main charachters act; they also delimitate some "regions" of chapters in which there are recurrent themes. Although this project only scratched the surface of both sentiment and topic analysis, it managed to indicate some trend in both techniques; the two combined helped to uncover a general view of a tragic novel that begins as a introspective and personal narration about Russian Society, then turns into a more general reflection about the consequences of the war and how it changes everyone's life and every society.

6 Summary

The main focus of this project was to understand how much of a complex text as War and Peace the text mining techniques were able to capture only with statistics. With a little help from my memory the result were quite surprising, both helping me unveil the underlying structure of the novel and discover a well-designed palace that host the events of War and Peace; while the novel flows smoothly and the intricate plot presents more than 600 characters.

What the sentiment analysis managed to make highlight is how emotions evolve as the topics defined in topic analysis changes. Cluster analysis helped us to understand how the characters were constructed in terms of similarities. I, personally, was rather satistfied by the results, as they they exceeded my expectations and managed to link the right charachters and produced interesting vocabulary linked to the war theme.

Before ending, again, I'd like to point out the importance of reading the entire novel, rather than interpreting this analysis (or even a more sophisticated one): interesting as the results may be, they cannot substitute the intimate experience that this reading represents.

Contents