The Digital Barricade : Computationally Analyzing *Les Misérables*

Yong-Yu Huang

Abstract

This project analyzes Victor Hugo's *Les Misérables* through a computational linguistic lens, studying sentiment and semantic similarity. I focus my project on three main questions: Can we track plot with sentiment analysis? How are revolutionary characters described in relation to each other? What words are most closely linked to questions of class and politics? I utilize Python-based natural language processing tools to analyze plot, characters, and themes important to Hugo's work.

1 Introduction

Computational linguistics provides robust tools in the digital humanities. Options such as word embedding models and sentiment analysis allow scholars to take a more quantitative approach to the rich texts that they study. Additionally, it allows the quantification and, therefore, visualization of the text's qualities. In this project, I approach Victor Hugo's 1862 novel, *Les Misérables*, originally written in the French language, with computational linguistics methods in an attempt to better understand and visualize the relationships within Hugo's novel. I aim to examine characterization and sociopolitical concepts such as class and gender.

Les Misérables is an epic novel following the escaped convict Jean Valjean as he makes a new life for himself under various names, whilst being pursued by Javert, a policeman. Simultaneously, Hugo covers class, politics, and gender as Valjean's path crosses with Fantine, a prostitute trying to care for her daughter Cosette, and the students at the fictionalized version of the June Rebellion of 1832.

I first thought about analyzing plot through sentiment after reading **Goyal, et al, 2010**'s paper [2] on plot unit analysis and projection rules. One of this project's is a much simpler undertaking, given the time and knowledge constraints that I am working with. Although I am not comparing narrative voices, **Brooke, et al, 2015**'s work [3] on narrative voices in T.S. Eliot's "The Wasteland" prompted me to think about the different characterization of Les Amis de l'ABC in *Les Misérables*.

2 Data

2.1 Data Source

I downloaded the text of *Les Misérables* and removed the preface and publishing edition text, preserving the Volume, Book, and Chapter labels to help me keep track of the novel's organization.

2.2 Cleaning Data, Pre-Processing

I then wrote a script to separate the novel into chapters, tracking which volume and book each chapter belongs to. This way, I did not have to manually go through all 365 chapters. I stored each chapter's text into a `pandas` dataframe.

Next, I wrote a function called `process_text` to remove stopwords from the `nltk.stopwords` lexicon and punctiation from each chapter's text. I then stored the processed text into the dataframe.

3 Methods

To conduct this analysis, I used sentiment analysis and semantic simlarity. In terms of tools, I used a Jupyter notebook, which runs a Python environment. I imported the text of *Les Misérables*, with the preface excluded, into `lesmis.txt`. I employed the Python `pandas` library to organize my data.

To process my data, I used the `nltk` library, a natural language processing toolkit. From the `nltk` library, I downloaded `stopwords`, `punkt`, `vader_lexicon`, and `SentimentIntensityAnalyzer`. Additionally, I imported `Word2Vec` from `gensim` and the `GloVe` library.

Finally, I visualized some of the work with `matplotlib`, a library for data visualization.

4 Analysis

4.1 Tracking Plot with Sentiment Analysis

I used nltk's SentimentIntensityAnalyzer to look at the sentiment throughout *Les Misérables*, using the tool on both the original, unprocessed text (stopwords and punctuation included) and the processed text (stopwords and punctuation removed). I wanted to see how the results would change.

I then wanted to use the processed text's sentiment results to visualize changes in sentiment throughout the novel. Specifically, I wanted to answer this question: Given sentiment scores for the novel, does the compound score generally correspond to high points (positive) and low points (negative) throughout the plot of the novel?

I plotted the sentiment analysis values from both the original and the processed text with matplotlib, representing positive, negative, neutral, and compound values on a line graph. Figure 1 visualizes the original text, while Figure 2 visualizes processed text.

The largest compound sentiment score for the unprocessed text was 0.216609708737864 (Volume 4, Book 5, Chapter 5), while the smallest compound sentiment score was -0.04799137931034483 (Volume 1, Book 7, Chapter 5).

Meanwhile, the processed text had much more drastic fluctuations for the compound values, with the largest compound sentiment score for the processed text being 0.9997 (Volume 4, Book 5, Chapter 5), while the smallest compound sentiment score was -0.8957 (Volume 2, Book 6, Chapter 9).

Both versions of the text saw their highest compound sentiment score values occur in Volume 4, Book 5, Chapter 4. Plotwise, this is an emotional high point, with Marius and Cosette finally meeting and confessing their love to each other. However, their lowest compound sentiment scores occurred in very different places.

In the processed text's sentiment analysis graph, one of the longest periods of noticeably negative sentiment ranged from Volume 5, Book 1, Chapter 19 to Chapter 22. This roughly corresponds to the brunt of character deaths occurring during the famous barricade—including Enjolras and Grantaire's joint execution.

Another long stretch of negative sentiment can be observed from Volume 4, Book 12, Chapter 6 to Volume 4, Book 14, Chapter 2. This section of the book sees Marius in despair over Cosette's im-

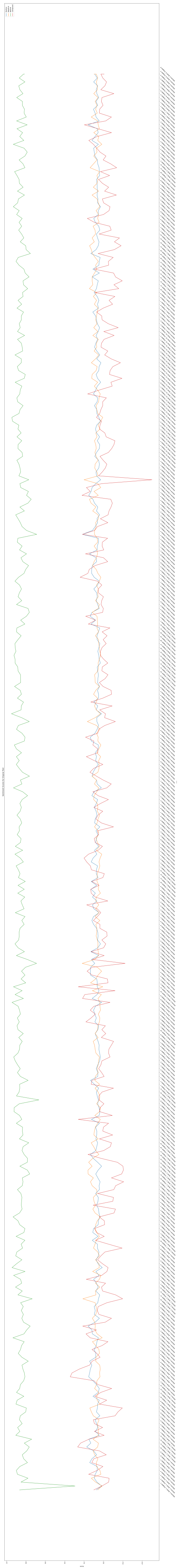

Figure 1: Original Text Sentiment Analysis

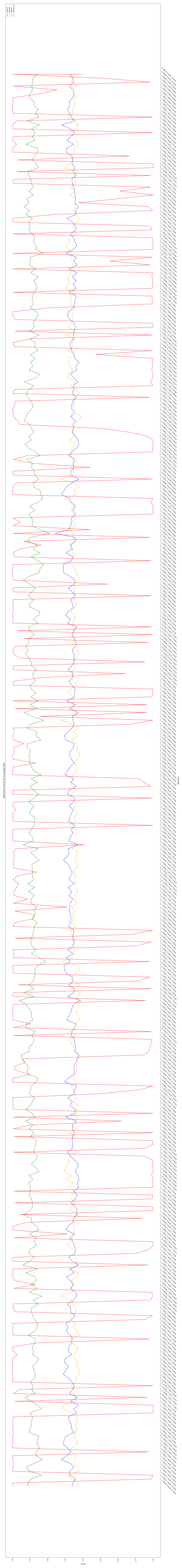

Figure 2: Processed Text Sentiment Analysis

pending departure to England. To that end, I argue that sentiment analysis can be used to roughly track narrative, at least with regard to general mood.

An extended stretch of relatively high sentiment occurs in Volume 3, where Hugo introduces Marius, a student in Paris who falls in love with Cosette from a distance. He details Marius' life, from his struggles to his eventual peace with a more solitary lifestyle. No major, ill-fated events happen in the section of the story, so I argue that this corresponds to a "high point" of the plot.

4.2 Sentiment Analysis on Characters

The most diverse defined group of people in the text would be *Les Amis de l'ABC*, the student revolutionaries that feature so heavily in the Paris Uprising. *Les Amis* includes upper-middle class students as well as the working laborer. Thus, there is the possibility of studying how Hugo portrays class by the descriptions they are assigned.

Table 1 (Character Sentiment Analysis on Members of Les Amis de l'ABC) contains the results of using the `nltk` library to run sentiment analysis. Some key words for Enjolras in the processed text, after stopwords and punctuation were removed, are `"wealthy enjolras charming young man capable terrible angelically handsome savage antinous."`

It is interesting to note that Grantaire's positive sentiment scores are higher than expected. I attribute this to the positive descriptions of Enjolras that are inextricable from Grantaire's corpus, as Hugo centers much of what we know of Grantaire around his relationship with and perception of Enjolras.

An observation on class bias in this model could be Feuilly's noticeably lower sentiment value. This may be due to his character introduction being markedly more political, and he has a more tragic backstory than the other members. The first portion of his processed text includes words like `"feuilly workingman fan-maker orphaned father mother earned difficulty three francs day"`, processed from `"feuilly was a workingman, a fan-maker, orphaned both of father and mother, who earned with difficulty three francs a day...""` This is opposed to the other members' well-off backgrounds as students. There is much more political commentary and Hugo informs the reader of Feuilly's opinions on po-

Character	Sentiment from Original Text (Positive, Negative, Neutral, Compound)	Sentiment for Processed Text (Positive, Negative, Neutral, Compound
Enjolras	(0.15554166666666663, 0.07908333333333334, 0.765375, 0.2032375)	(0.25, 0.091, 0.659, 0.9937)
Combeferre	(0.14402857142857145, 0.05205714285714286, 0.8038857142857142, 0.19259714285714286)	(0.23, 0.124, 0.646, 0.9953)
Courfeyrac	(0.06076923076923077, 0.041692307692307695, 0.8975384615384615, 0.08915384615384617)	(0.146, 0.045, 0.808, 0.9494)
Feuilly	(0.05815625, 0.10534375000000001, 0.8365625, -0.137346875)	(0.118, 0.201, 0.681, -0.9686)
Jehan	(0.11850000000000001, 0.052, 0.8295000000000001, 0.17847857142857146)	(0.247, 0.133, 0.619, 0.976)
Bahorel	(0.04911764705882354, 0.02064705882352941, 0.9302352941176469, 0.09201764705882354)	(0.181, 0.048, 0.771, 0.9758)
Bosseut	(0.05972500000000001, 0.049574999999999994, 0.8906750000000001, 0.015379999999999996)	(0.144, 0.118, 0.738, 0.7559)
Joly	(0.05238461538461538, 0.012384615384615385, 0.9353076923076924, 0.1835076923076923)	(0.157, 0.038, 0.804, 0.9426)
Grantaire	(0.06076923076923077, 0.041692307692307695, 0.8975384615384615, 0.08915384615384617)	(0.146, 0.045, 0.808, 0.9494)

Table 1: Character Sentiment Analysis on members of Les Amis de l'ABC, Word2Vec

Characters	Similarity (vector size=1000, window=100)	Similarity (vector size=100, window=10)
Grantaire, Valjean	0.003599994	-0.008130907
Enjolras, Grantaire	-0.044779196	0.0463875
Valjean, Combeferre	0.03970809	-0.106069386

Table 2: Character Similarity, Word2Vec

Characters	Similarity (vector size=1000, window=100)	Similarity (vector size=100, window=10)
Grantaire, Valjean	0.08144424	0.12389229
Enjolras, Grantaire	0.6490359	0.64294946
Valjean, Combeferre	0.27262577	0.3236971

Table 3: Character Similarity, GloVe

litical violence and certain events, the description of which contains words such as infamous, suppressed, treason, crime, monstrous ambush, prototype, pattern, horrible, suppressions, crimes, origin, outrages, and traitor.

4.3 Semantic Similarity

I used gensim's Word2Vec in the skip-gram model to run semantic similarity analyses on the text as well. I used the skip-gram model because I wanted to make sure it understood rare words. Specifically, I wanted to see what connections might be made with themes of class and authority. To that end, I trained a model with a vector size of 1000 and a window of 100. It has a vocabulary size of 25057.

I then queried it with a series of words that I hoped would shed some light as to how Hugo approaches these themes in his novel. Interestingly enough, church is most often associated with fiscal terms like livres, expenses, and francs. However, this may be attributed to the careful record-keeping of expenses in the church that appears in the novel's early chapters. Bourgeois was most closely associated with paved, which indicates some connection to urban areas, while "working" was closely associated with distress and nails, the latter of which may indicate a more visceral and physical connection to labor.

When looking at similarity between characters in Table 2, however, I found that the model with vector size 1000 and window size 100 was not particularly accurate, given that it claims Enjolras and Grantaire are less related than Grantaire and Valjean, who do not really interact.

Thus, I then trained another model with vector size 100 and window size 10. It has a vocabulary size of 25191. Its scores more accurately reflect characters' relationship. It can be inferred from the novel that Enjolras and Grantaire are characters that are much more closely related, than Grantaire and Valjean or Valjean and Combeferre. The latter pairs do not interact, nor are their roles and contexts in the book similar.

I then tried training the GloVe algorithm on a processed corpus of *Les Misérables*, removing punctuation, stopwords, and standardizing the text to lowercase. I used the same two vector size and window combinations ([vector=1000, window=100] and [vector=100, window=10]). The results can be seen in Table 3.

4.4 Discussion

This project was extremely interesting to explore. I learned that in order to effectively be able to confirm these computational results, I had to already be very familiar with the novel's content, especially when assessing plot. Comparing sentiment scores for both the original and processed texts were helpful, especially when studying individual characters. I found that part of the project to be most interesting, as I was able to better visualize the keywords used to characterize the group of young revolutionaries and note the difference in linguistic connotations.

The model could definitely be further fine-tuned—especially the GloVe vector. This was my first time compiling a corpus to train an algorithm on, and I'm not quite sure I completely understood the impact of threads on the process. The text-preparation and initial data-cleaning could also be further refined. There were words that sometimes ended up joined together, which might have been a result of how the online text I used was formatted originally. I didn't find the semantic similarity queries particularly compelling; in light of that, my dataset or how I trained it is probably flawed in some way.

I found how Hugo frequently references other characters when introducing new ones, such as in the context of Grantaire and Enjolras, especially

Query	10 Most Similar Words
revolution	('distance', 0.9999632239341736), ('blows', 0.999962568283081), ('existence', 0.999960720539093), ('progress', 0.9999470114707947), ('breast', 0.9999411702156067), ('mother', 0.9999338984489441), ('forth', 0.9999300241470337), ('french', 0.9999241232872009), ('woman', 0.9999233484268188), ('alas', 0.9999207854270935)
law	('flesh', 0.9999615550041199), ('seemed', 0.9999436736106873), ('though', 0.999942421913147), ('almost', 0.9999412298202515), ('moreover', 0.9999284744262695), ('avoided', 0.9999257922172546), ('angel', 0.9999216198921204), ('epoch', 0.9999197721481323), ('body', 0.9999184012413025), ('distance', 0.9999173283576965)
god	('things', 0.9999383091926575), ('think', 0.9999316930770874), ('fall', 0.9999259114265442), ('souls', 0.9999130368232727), ('make', 0.9999104738235474), ('future', 0.99990314245224), ('pleases', 0.9998994469642639), ('say', 0.9998980164527893), ('last', 0.9998935461044312), ('may', 0.9998865127563477)
working	('distress', 0.9994498491287231), ('nails', 0.9994300603866577), ('imagine', 0.9994211196899414), ('close', 0.9994202256202698), ('strongly', 0.9994131326675415), ('furnace', 0.9994075894355774), ('workman', 0.9993896484375), ('evening', 0.9993865489959717), ('gallican', 0.9993826150894165), ('hatred', 0.9993798732757568)
bourgeois	('paved', 0.9996536374092102), ('discovered', 0.9996536374092102), ('ugly', 0.9996182918548584), ('magloire', 0.9996134638786316), ('washed', 0.9996126294136047), ('grandchamp', 0.9996075630187988), ('difficult', 0.9995889067649841), ('attended', 0.9995853304862976), ('whitewashed', 0.9995787739753723), ('madame', 0.9995787143707275)
church	('livres', 0.9999460577964783), ('expenses', 0.9999343752861023), ('ten', 0.9999341368675232), ('francs', 0.9999332427978516), ('diocese', 0.9999294281005859), ('found', 0.999925434589386), ('bishops', 0.9999219179153442), ('wrote', 0.9999213814735413), ('hundred', 0.9999212026596069), ('left', 0.9999200701713562)
poor	('president', 0.9984006285667419), ('immediately', 0.9983896613121033), ('required', 0.9983195066452026), ('activity', 0.9983102679252625), ('waited', 0.9982302188873291), ('breathin', 0.9982102513313293), ('ii', 0.9980977773666382), ('chambers', 0.9980403780937195), ('corpulent', 0.9980330467224121), ('story', 0.9980077147483826)]
rich	('behold', 0.9999831318855286), ('service', 0.999981164932251), ('one', 0.9999791979789734), ('priests', 0.9999779462814331), ('whole', 0.9999777674674988), ('men', 0.9999733567237854), ('teach', 0.9999715089797974), ('bread', 0.9999693036079407), ('taken', 0.9999679327011108), ('en', 0.999967634677887)
government	('jeannette', 0.14183488488197327), ('ravaging', 0.12477511912584305), ('heat', 0.1151130348443985), ('conversing', 0.11435714364051819), ('disdains', 0.11013119667768478), ('aphorisms', 0.10968374460935593), ('1861', 0.10950001329183578), ('wicket', 0.10807859897613525), ('vanquishing', 0.1076110303401947), ('thy', 0.10630739480257034)
prison	('tried', 0.9996908903121948), ('refused', 0.9996017217636108), ('entire', 0.9994922280311584), ('tragic', 0.999484658241272), ('writer', 0.9994750618934631), ('worthy', 0.9994733929634094), ('relating', 0.9994693398475647), ('spoke', 0.999467134475708), ('task', 0.9994634389877319), ('invented', 0.9994630813598633)
justice	('forth', 0.9999598860740662), ('silence', 0.9999483823776245), ('cloud', 0.9999323487281799), ('judge', 0.9999301433563232), ('love', 0.9999282360076904), ('crown', 0.9999268054962158), ('indignation', 0.9999260306358337), ('created', 0.9999234676361084), ('wretched', 0.9999202489852905), ('face', 0.9999187588691711)
france	('archbishop', 0.9999614953994751), ('days', 0.9999533891677856), ('paris', 0.9999474883079529), ('gold', 0.9999470114707947), ('charles', 0.9999450445175171), ('little', 0.9999440908432007), ('receive', 0.9999428391456604), ('trees', 0.9999415278434753), ('century', 0.9999412298202515), ('place', 0.9999390840530396)
liberty	('allpowerful', 0.9999784827232361), ('ecclesiastes', 0.9999331831932068), ('quarto', 0.9999311566352844), ('emerged', 0.9999234676361084), ('clinton', 0.9999234080314636), ('germain', 0.9999215006828308), ('generals', 0.9999197125434875), ('whence', 0.999919593334198), ('volume', 0.9999175071716309), ('kings', 0.9999111294746399)

Table 4: Semantic Similarity for Relevant Vocabulary, Word2Vec

interesting because of how that influences their sentiment scores. Additionally, Hugo seems to characterize some characters almost solely by their political views and background, such as Feuilly. Similarly, he associates certain political words with tangible imagery, making class delineations clear to the reader. In the future, I would be interested in trying to look at character relationships through character networks, like in **Labatut and Bost, 2019** [1].

References

[1] Vincent Labatut and Xavier Bost. 2019. Extraction and Analysis of Fictional Character Networks: A Survey. ACM Comput. Surv. 52, 5, Article 89 (September 2020), 40 pages. https://doi.org/10.1145/3344548

[2] Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically Producing Plot Unit Representations for Narrative Text. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 77–86, Cambridge, MA. Association for Computational Linguistics.

[3] Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Distinguishing Voices in The Waste Land using Computational Stylistics. Linguistic Issues in Language Technology, 12.