# Multilingual Sentiment Analysis and Translation: Spanish and English Story Arcs in Juan Rulfo's *Pedro Páramo*

Eva Donahue

IPHS 200 Programming Humanity (Fall 2022) Prof Elkins and Chun, Kenyon College

## Abstract

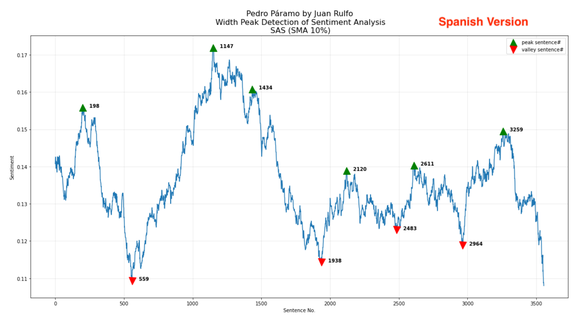
Sentiment analysis technology can help create data regarding works of literature that were once only thought about in the abstract. This project conducts a sentiment analysis of the novel *Pedro Páramo* by Juan Rulfo, once in English and again in Spanish. Through the comparison of the story arcs and the high and low crux points of the story found by each of the sentiment analysis programs, the impact of translation was revealed. Not only were the ensemble graphs and crux plots looked at, but a deep dive into the context surrounding each of the crux points was conducted. Though both stories have similar narrative arcs overall, the points of emotional impact differed between the original and the translation. This reveals that the experience of the reader may be different depending on the language they read a novel in, even if the takeaway of the book is the same. Further research must be conducted on the efficacy of the sentiment analysis models across languages, the accuracy of translation, and more tests with different, larger novels.

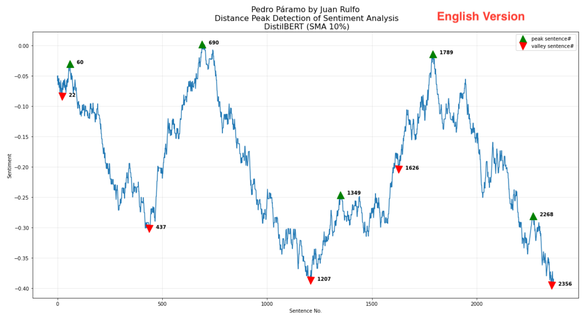
## Introduction

Sentiment analysis arcs are useful tools for tracking the course of a narrative, discovering how we shape stories, and applying that knowledge to the overall study of literature and language today. However, the emotional impact of a story is shaped not only by the course of action, but also by the language used to describe it. For example, a story in the language it was written in compared to a version translated into English can often feel different to the multilingual reader, as if the story itself is propped up by centuries of language tradition. To understand the impact of translation on storytelling more, this project will look at the sentiment analysis of a classic Mexican novel, *Pedro Páramo* by Juan Rulfo, both in its native language, Spanish, and translated into English. *Pedro Páramo* is well -known, and has been translated many times into many different languages, and is full of a depth of emotion as the reader is taken through an allegory for lost identity during the Mexican revolution. The novel follows Juan Preciado, acting on a dying wish of his mother, as he sets out to the mysterious ghost town of Comala to discover his father, Pedro Páramo. The story weaves through a few different generations of characters throughout their time in Comala, told mostly from Juan Preciado's coffin shared with a woman named Dorotea, where he learns the tales of the acts of exploitation committed by his father and his grandfather. Looking into the sentiment analysis arcs of *Pedro Páramo* can tell us how a computer built upon language exposure understands feeling across languages, and in turn how translation may impact the reader's experience.

## Methodology

To conduct this study, I ran a sentiment analysis program of the same novel, *Pedro Páramo* by Juan Rulfo, once in English and then again in Spanish. Most "sentiment analysis models are based on some form of supervised learning" (Chun 2021). Therefore it is not possible to use the same exact model/code for the two different languages. In turn, the project must be split into two parts: first running the code of each independently to evaluate if the sentiment analysis is even accurate enough, then, once the two models have proven to be fairly reliable, comparing the results of the graphs in the two languages. In both versions of the sentiment analysis code, I ran the novel through as a txt file. Each analysis used more than one program. For example, in the English version this included VADER, Textblob, DistilBERT, and RoBERTa. The Spanish version included Pysentimiento, MultiBERT, and more. Each code cleaned and sliced the sentence strings before running the sentiment analysis. Then, I looked at the ensemble graph for each and chose the median version, assuming here that based on a group mentality, the line that seemed most average would be the most accurate. Then, I singled out that graph line and put it through crux detection. The code gave me four options with a different amount of cruxes in each, and again I chose the median result. After that, the crux graphs, along with the sentiment arcs and the text corresponding to each crux were downloaded to my computer for analysis.




## Results

In a broad sense, both analyses depict fairly accurate arcs of the novel in both English and Spanish. Something to note is that the two crux plots have different y -axis scales. In Spanish the change in emotion is plotted to be more subtle, while in English the entire sentiment of the novel is plotted on a negative scale. However, it will be assumed that this is due to the difference of the model, not a significant difference in the meaning of the graphs. Each model understands the general high and low points of the novel, such as feelings of peace or resolution among the characters, and ending with the death of all of the characters as a valley. The first crux I looked at was peak one, the high points of both novels. In the Spanish version, sentence 1147 describes characters coming to terms with the death of their town, Comala, and of themselves. They describe ghosts to the narrator and how they came to live among them. In the English version, sentence 690 represents Dolores, Juan's mother, being told she must marry Pedro so he can prevent paying his debts. This is a discussion with Don Fulgor, who works for Pedro around manipulating a female character, a job reserved for Dorotea. Then I looked at the low point of the novel, crux valley two. In Spanish, sentence 1938 describes the death of Pedro's father and features a lot of language of grief. In English, sentence 1207 discusses Dorotea and the longing for her baby. Both are in similar places of the novel that surround the decline of the town, Comala. Lastly, I wanted to observe why the crux peaks at the end of the novel appear different on the graph, but found that they are not that dissimilar after all. In Spanish, sentence 3259 is Susana San Juan talking to Padre Rentería about her marriage to Pedro, and her impending death due to her misery in her situation. In English, sentence 1789 is also a discussion between Susana San Juan and Padre Rentería surrounding marriage, death, and Don Fulgor, though a different portion of the conversation than the Spanish version.

```
Peak #1 at Sentence #690:

No, it would be better to send someone to tell her, but anyway I
wouldn't be ready before the April eighth.
It's now the first, so it's too soon.
Tell him to wait just a few days."
"He would like to have it right away.
As for the trousseau, we can take care of that.
Before she died, Pedro's mother wanted you to have her clothes.
That is a custom in the family."
"But there's a reason I have to wait.
It's a woman's thing, you know.
Oh, how embarrassed I am to have to say this, Don Fulgor.
You're making me blush again.
IT'S THE TIME FOR MY... OH, I AM ASHAMED TO SAY IT."
```

```
Peak #1 at Sentence #1147:

Me acerqué para ver el mitote aquel y vi esto: lo que estamos viendo ahora.
Nada.
Nadie.
Las calles tan solas como ahora.
»Luego dejé de oírla.
Y es que la alegría cansa.
Por eso no me extrañó que aquello terminara.
»Sí —volvió a decir Damiana Cisneros—.
Este pueblo está lleno de ecos.
YO YA NO ME ESPANTO.
Oigo el aullido de los perros y dejo que aúllen.
```

## Conclusion

Overall, the results of the sentiment analysis found that while the emotional journey and shape of the sentiment analysis align, the exact point of impact changes between the English and Spanish versions. In the English version, the emotional crux in the beginning often surrounds the character Dorotea and her struggle. In contrast, in the beginning of the arc of the Spanish version, the emotional cruxes mention Pedro Páramo or Comala more. According to these models, the language in which a book is written changes the experience of the reader, emphasizing different moments, sometimes with different contexts all together, as the most poignant. However, all readers will ultimately come out of the experience with a similar picture of the story as a whole, as demonstrated through the shape of the two sentiment arcs, which are fairly in agreement. Additionally, each of the cruxes surround similar themes in the story, in congruence with the rise and fall of Comala.This brings up interesting questions about how stories are translated and told across cultures. If the computer forms learns and forms sentiment analysis based on a wealth of data from a particular language, the cultural elements that shape a language may carry over into the computer's learning. This could be an explanation for the difference in cruxes between languages.

## Future and Ethics Statement

This project was completed within an extremely limited time span, and in turn has many different elements that could be expanded upon in future research. Firstly, Pedro Páramo is a very short novella, giving the model a limited amount of data to work with. Looking at more novels, especially longer ones, will give a larger range of data to support the findings that sentiment analysis is different between languages. It is also possible that there are confounding variables impacting the results of this study. For example, some of the Spanish models are trained off of tweets, would be interesting to look at more models that are trained off of novels. Additionally, the difference between results could be impacted by the work of the translators, and in the future more time could be put into making sure all translations are extremely accurate. Lastly, it would be interesting to run the sentiment analysis over many languages, to see if this difference persists.

## References/Acknowledgements

Thank you to Professor Elkins and Professor Chun for their assistance in this project.

Chun, Jon. *Sentiment Arcs: A Novel Method for Self -Supervised Sentiment Analysis of Time Series Shows SORA Transformers Can Struggle Finding Narrative Arcs* . October, 2021, ArXiv.