# Harvesting Digital Fields:
## Creating a Corpus of Web Scraped Data

Camila Lívio
Department of Research and Computational Data Management

University of Georgia

November 15, 2023

# Table of Contents

# Introduction

## Three Wolf Moon T-Shirt

This shirt has changed my life! (...) with The Mountain Three Wolf Moon Short Sleeve Tee, the SHIRT now draws the eye. One young teen even shyly approached me, and instead of asking for a picture or an autograph, simply smiled conspiratorially and whispered, "Team Jacob, right? Me, too. He's soooooooo dreamy." Yes he is, young lady. Yes. He. Is.

## Tuscan Milk

Shipping was fine, and the product was not damaged in any way, but my husband and I (both of us have college degrees, mind you, his in Engineering) could not figure out how to assemble this. No instructions, no diagrams, not even a lousy cheap allen wrench. So basically we're using it as a one gallon paper weight. I haven't gotten any response from Tuscan. It earns two stars simply because it is heavy and does do a fair job of holding down the stack of newspapers awaiting recycling.

1. What kind of text are these?
2. What are the general characteristics of this text genre?

# Online Reviews

Vásquez, Camilla. "My life has changed forever!" Narrative identities in parodies of Amazon reviews." Narrative Inquiry 27.2 (2017): 217-234.

Today's lecture:
(1) Building a specialized corpus from web scraped data
(2) Intensification
(3) Some elements to consider when building a corpus

# Intensifiers – A quick overview

- Speaker's tendency to innovate and engineer different ways to boost meaning
- Scale up the quality or property of another word
- Go beyond semantic gradability as they increase the expressive strength and subjectivity in discourse
- **'Really cool', 'Absolutely amazing', 'A proper disaster'**

- What are other intensifiers do you hear these days (in any language you speak/work on)?

# Semantic gradability & Subjectivity

‘Absolutely amazing movie’ vs. ‘an amazing movie’

Their illocutionary force varies: statement and an expressive act

# Semantic gradability & Subjectivity

cont.

Intensifiers have also been studied in light of social membership

Martos (2010): *super* has ascended, socially and stylistically, to be perceived as a middle and upper-class female usage in Madrid (2010: 202)

Bradac et al. (1995): 'female register' or 'register of low power'. The use of intensifiers, hedges, and tag questions as linguistic forms that are commonly associated with women

# Inventory

1. Adverbs (*muito/muy, bem/bien*, 'very')
2. Multi-word expressions (*um monte de/montón de*, 'a lot of')
3. Affixes (*Pt. belíssimo/trilegal*, Sp. *riquísimo, superguay, superinteresante*)
4. Absolute adjectives (*Perfeito/perfecto* 'perfect')
5. Pronouns (*Trae unos Raleighs, ¡córrele!*) [1]
6. Certain verbal constructions (*¡Vaya coche tan moderno!*, 'What a modern car!') [2]
7. Que + adjective (*Que lindo!*, 'So + Adjectives')
8. Comparative phrases (Sp. *Más feliz que una lombriz*, Pt. *Mais quebrado que arroz de terceira*, 'Happier than/Cooler than', 'As cool as...')

"intensity is signaled by a large and miscellaneous class of devices, ranging from the most peripheral of ~~prosodic variations to the most central cat~~egories of the grammar" (Labov 1985a: 48).

# Two Studies

(1) Intensifiers in the RCor = Reviews Corpus

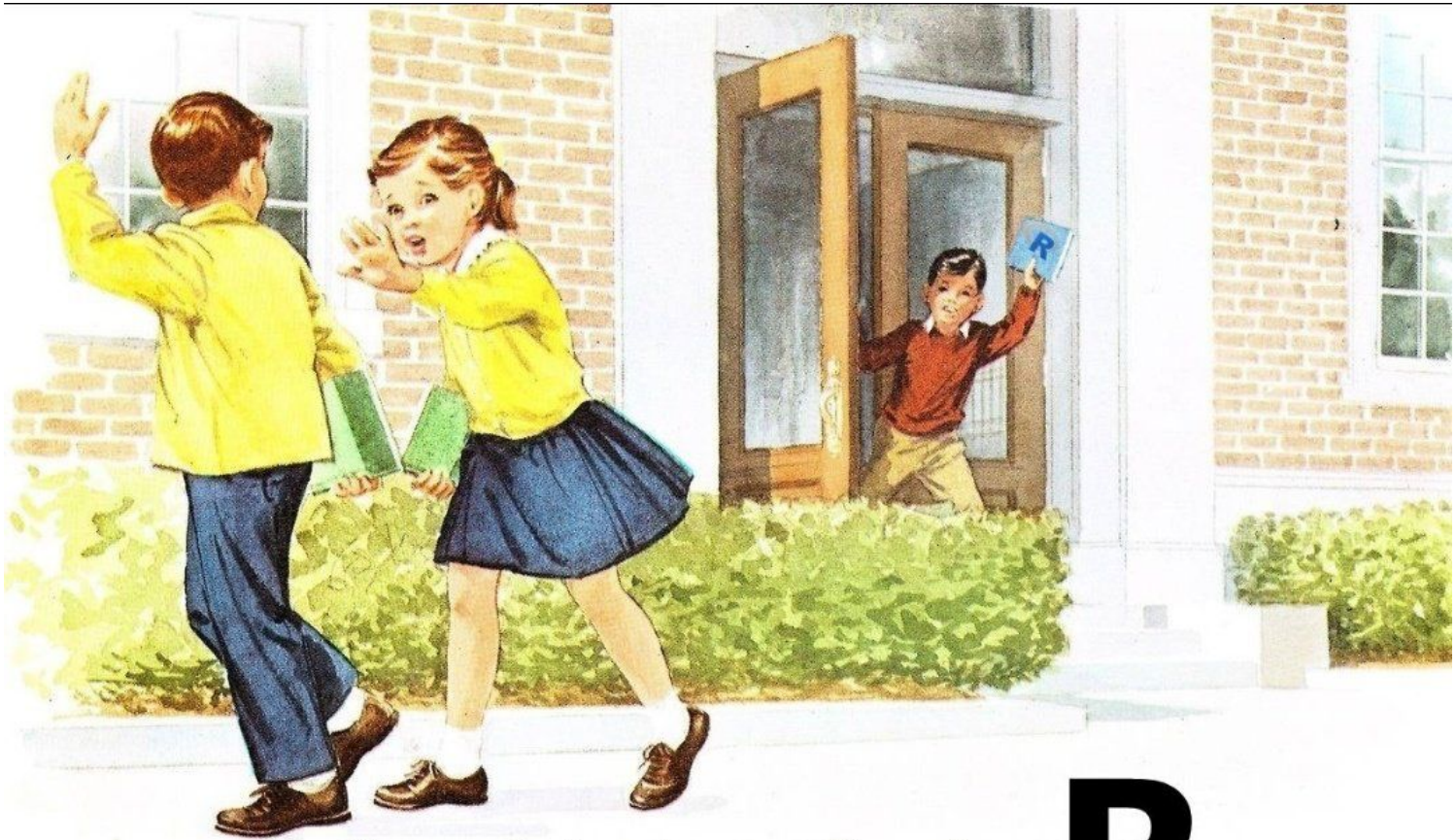
(2) Social Evaluation of Intensifiers

# Why Intensifiers in Online Reviews?

- Intensifiers are very common in evaluative language
- Language that expresses an attitude towards a person, a situation or any entity, being both subjective and situated in a societal value-system (Hunston 2010)
- Evaluative language = subjective
- The scalar nature of evaluative language
- The connection between evaluative language and its scalar nature offers a strong motivation for considering customers' reviews as a relevant linguistic source for the study of intensifiers
- Online reviews as a text genre

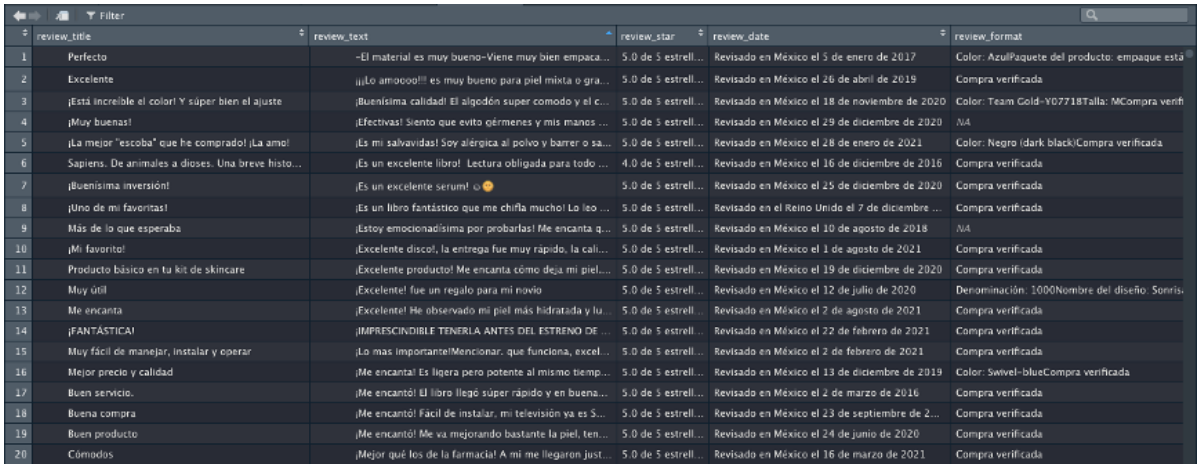
# Building the RCor

- R Core Team (2020)
- Web scraping
- Specialized corpus (Brezina and McEnery 2020: 17-19)
- Brazil and Mexico are the two countries in Latin America with the highest number of visits to Amazon.com
- RCor is balanced across product types
- Filter reviews written in the target language
- Organization of the data in columns and rows
- Data quality check
- Half a million words


R
Run, or he's going to tell us about
again!
R

# What does the RCor look like?

Figure: RCor – MS



| | review_title | review_text | review_star | review_date | review_format |
|---|---|---|---|---|---|
| 1 | Perfecto | -El material es muy bueno-Viene muy bien empaca... | 5.0 de 5 estrell... | Revisado en México el 5 de enero de 2017 | Color: AzulPaquete del producto: empaque está |
| 2 | Excelente | ¡¡¡Lo amoooo!!! es muy bueno para piel mixta o gra... | 5.0 de 5 estrell... | Revisado en México el 26 de abril de 2019 | Compra verificada |
| 3 | ¡Está increíble el color! Y súper bien el ajuste | ¡Buenísima calidad! El algodón super comodo y el c... | 5.0 de 5 estrell... | Revisado en México el 18 de noviembre de 2020 | Color: Team Gold-Y07718Talla: MCompra verifi |
| 4 | ¡Muy buenas! | ¡Efectivas! Siento que evito gérmenes y mis manos ... | 5.0 de 5 estrell... | Revisado en México el 29 de diciembre de 2020 | *NA* |
| 5 | ¡La mejor "escoba" que he comprado! ¡La amo! | ¡Es mi salvavidas! Soy alérgica al polvo y barrer o sa... | 5.0 de 5 estrell... | Revisado en México el 28 de enero de 2021 | Color: Negro (dark black)Compra verificada |
| 6 | Sapiens. De animales a dioses. Una breve histo... | ¡Es un excelente libro! Lectura obligada para todo ... | 4.0 de 5 estrell... | Revisado en México el 16 de diciembre de 2016 | Compra verificada |
| 7 | ¡Buenísima inversión! | ¡Es un excelente serum! | 5.0 de 5 estrell... | Revisado en México el 25 de diciembre de 2020 | Compra verificada |
| 8 | ¡Uno de mi favoritas! | ¡Es un libro fantástico que me chifla mucho! Lo leo ... | 5.0 de 5 estrell... | Revisado en el Reino Unido el 7 de diciembre ... | Compra verificada |
| 9 | Más de lo que esperaba | ¡Estoy emocionadísima por probarlas! Me encanta q... | 5.0 de 5 estrell... | Revisado en México el 10 de agosto de 2018 | *NA* |
| 10 | ¡Mi favorito! | ¡Excelente disco!, la entrega fue muy rápido, la cali... | 5.0 de 5 estrell... | Revisado en México el 1 de agosto de 2021 | Compra verificada |
| 11 | Producto básico en tu kit de skincare | ¡Excelente producto! Me encanta cómo deja mi piel.... | 5.0 de 5 estrell... | Revisado en México el 19 de diciembre de 2020 | Compra verificada |
| 12 | Muy útil | ¡Excelente! fue un regalo para mi novio | 5.0 de 5 estrell... | Revisado en México el 12 de julio de 2020 | Denominación: 1000Nombre del diseño: Sonris |
| 13 | Me encanta | ¡Excelente! He observado mi piel más hidratada y lu... | 5.0 de 5 estrell... | Revisado en México el 2 de agosto de 2021 | Compra verificada |
| 14 | ¡FANTÁSTICA! | ¡IMPRESCINDIBLE TENERLA ANTES DEL ESTRENO DE ... | 5.0 de 5 estrell... | Revisado en México el 22 de febrero de 2021 | Compra verificada |
| 15 | Muy fácil de manejar, instalar y operar | ¡Lo mas importante!Mencionar, que funciona, excel... | 5.0 de 5 estrell... | Revisado en México el 2 de febrero de 2021 | Compra verificada |
| 16 | Mejor precio y calidad | ¡Me encanta! Es ligera pero potente al mismo tiemp... | 5.0 de 5 estrell... | Revisado en México el 13 de diciembre de 2019 | Color: Swivel-blueCompra verificada |
| 17 | Buen servicio. | ¡Me encantó! El libro llegó súper rápido y en buena... | 5.0 de 5 estrell... | Revisado en México el 2 de marzo de 2016 | Compra verificada |
| 18 | Buena compra | ¡Me encantó! Fácil de instalar, mi televisión ya es S... | 5.0 de 5 estrell... | Revisado en México el 23 de septiembre de 2... | Compra verificada |
| 19 | Buen producto | ¡Me encantó! Me va mejorando bastante la piel, ten... | 5.0 de 5 estrell... | Revisado en México el 24 de junio de 2020 | Compra verificada |
| 20 | Cómodos | ¡Mejor qué los de la farmacia! A mí me llegaron just... | 5.0 de 5 estrell... | Revisado en México el 16 de marzo de 2021 | Compra verificada |

- Review title
- Text of the review
- Star rating
- Review date
- Review format

# Scraping the Data

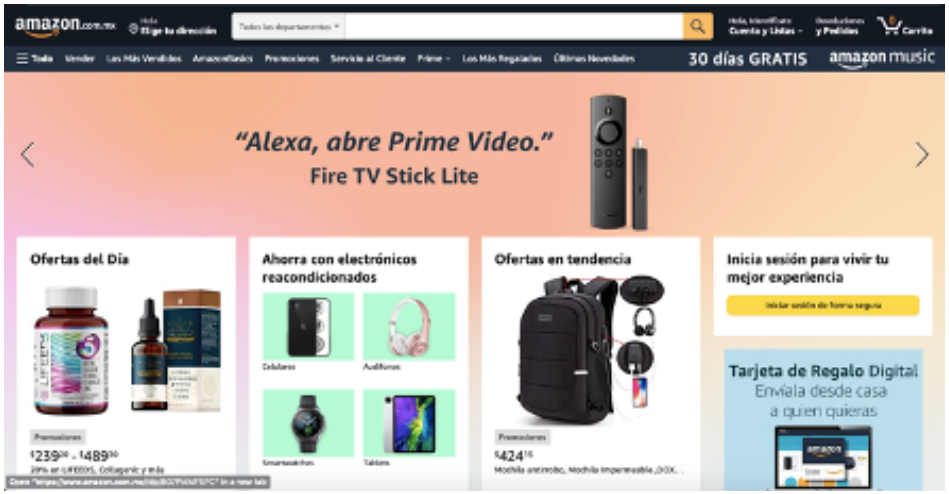



Figure: Screenshot of an Amazon page

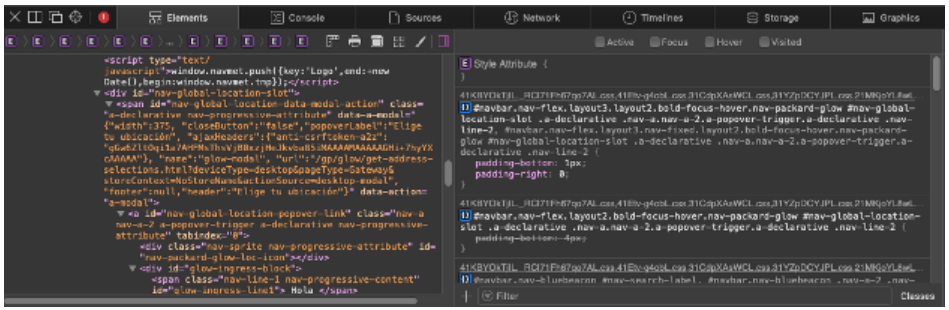



Figure: Screenshot of HTML elements of the same page

# Building a corpus via web scraping

- Data scraping: a technique in which a computer program extracts data from human-readable output coming from another program.
- The scraper was programmed to collect random reviews from 50 pages for each product
- Top-seller of each category (*Mais vendidos*, 'best sellers')
- Benefits: amount of data, and the ability to gather and organize the data set in a way that facilitates looking for patterns through the use of corpus linguistics methods.
- RCor is balanced across product categories

- What would be some alternative ways to balance this type of corpus?

# Ethical Considerations

- Drawbacks: Data set is *not* publicly available so that the research is compliant with ethical considerations, such as personal data and privacy, and copyright and intellectual property
- Do not collect any personally identifiable information
- Use the data only for the purpose for which it was collected
- Use aggregated data rather than individual reviews (helps with maintaining anonymity)
- Check the website's **'robots.txt'**
- Look for information about writing a Data Management Plan (DMP)
- GitHub

# Research Questions – Task 1: Intensifiers in Online Reviews

1. How do intensifiers vary across product categories?
2. Are there any associations that can be made between the use of intensifiers in positive and negative Amazon reviews?
3. What can the study of intensifiers in this environment teach us about computer-mediated communication?

# Intensifiers in the RCor

The analysis is divided in four main parts

(1) Prefixation, Suffixation and Lexical Intensifiers
(2) Intensifiers by Product Categories
(3) Intensifiers by Product Ratings
(4) Intensifiers and the construction of a reliable online persona

# Intensifiers by Product Category

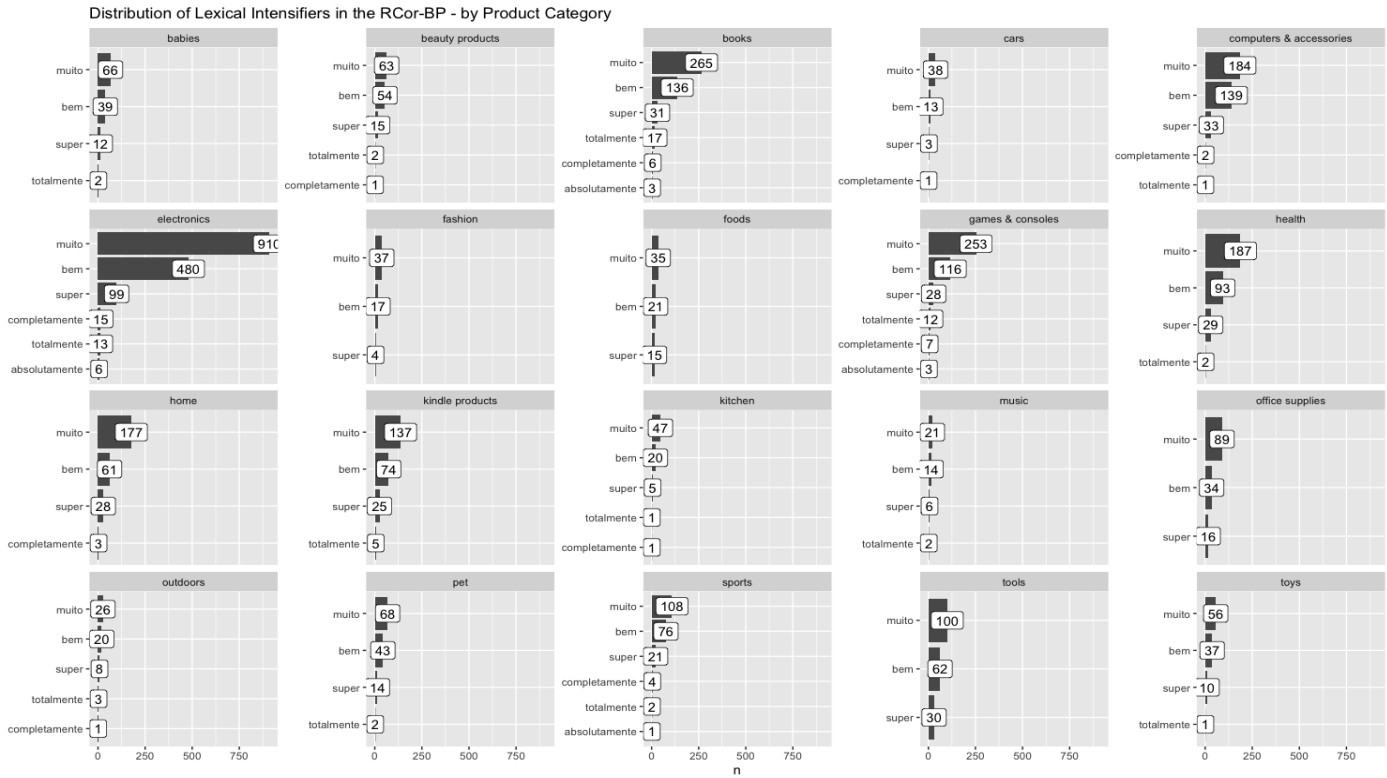
## Question

How is the use of intensifiers different, if at all, across different product categories?

- Some product categories show a higher frequency of intensifiers, as well as a more varied inventory
- Felbermayr and Nanopoulos (2016): where there is language related to play and fun, the emotional spectrum tends to be larger (i.e., a wider set of words that describe emotions)
- BP: books, beauty, computers and electronics, games, home, kindle, kitchen, and music == categories with over 3 different intensifying forms
- MS: amazon devices, baby, books, beauty, home, musical instruments, toys == categories with over 3 different intensifying forms

# Intensifiers by Product Category

Figure: Intensifiers by Product Category



# Intensifiers by Product Rating

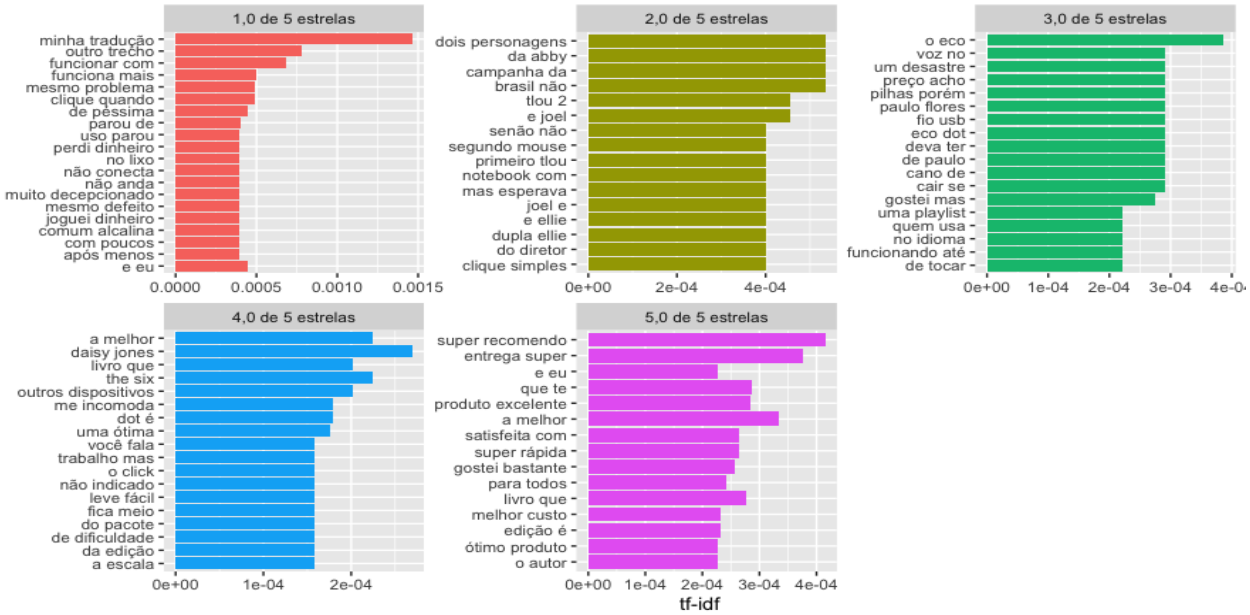
## Question

Does rating matter?

- Plotted the correlation between bigrams extracted from positive (4-5 stars) and negative (1-2 stars)
- Created two tf-idfs
- Combinations that appear more than five times
- Intensifiers are preferred when reviews are positive
- When reviews are negative, users seem to focus on the description of the product (and less on their experience)
- Alternative way to visualize this connection: distribution of bigrams by product rating

# Bigrams by Rating

Figure: Bigrams by Rating in the RCor-BP



# Conclusions of Task 1

- Online reviews can be considered a text genre
- Intensifiers are a crucial feature of such genre
- I connect the compulsory use of intensifiers in online reviews to the user's need to be relevant, and show credibility (i.e., expressing their commitment)
- Canonical intensifiers are most frequent, a fact that I attribute to the user's construction of a reliable persona, trustworthiness, and expertise (Mackiewicz 2008)
- Digital self-expression to state 'entrepreneurial agency' > Weninger and Li (2002) mention the existence of a 'microcelebrity persona' > somebody who is willing to bear the cost of broadcasting to achieve some kind of recognition
- Hu et al. (2009): the overconfident user is the one who is more likely to bear the costs of broadcasting > explain the overwhelming number of positive reviews

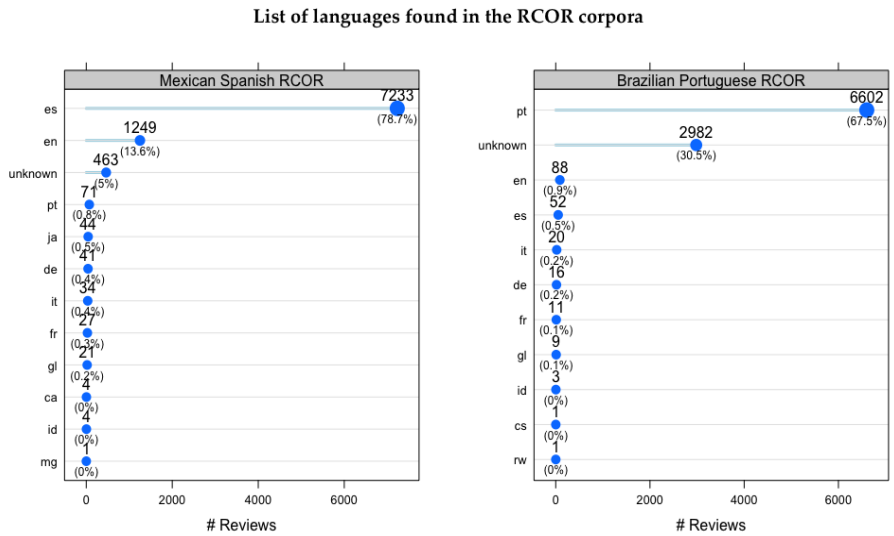
# Future Directions

- Patterns of spelling variation and paralinguistic resources: emojis and caps lock (McCulloch 2019, 2020)
- Quantitative patterns of stacking in the Romance languages (Scheffler et al. 2023)
- Can tense influence the distribution of intensifiers?
- The relevance of intensifiers in the timely subject of fake news (Grieve and Woodfield 2023)

*Super obrigada!*

camila.emidio25@uga.edu

# Languages in the RCor prior to Filtering



# Star Rating Distribution - Note the j-shaped distribution (Hu et al. 2009)