

# Can a book make you **happy**?

Predicting emotional links between genre,  
plot, and reader response

**Srishti Sharma**

**Dr. Federico Piazola**

Registered Report on OSF (<https://osf.io/xg6d4/>)





**Emotions** elicited by stories are  
vital to the relationship a reader  
shares with a book

# Sentiment Analysis

---

# Emotion Detection

Dumbledore **loves** karaoke. | **POSITIVE**

Snape **hates** roller skating. | **NEGATIVE**

Ron **enjoyed** a pumpkin pie. | **JOY**

Draco **sneered** at Potter. | **ANGER**

# Hypotheses

“framing effect” (Tversky and Kahneman 1981; Levin and Gaeth 1988)

- 01 The sentiment of a book has a positive linear relationship with the sentiment of its reviews.
- 02 The variation of sentiment in the story's ending has a stronger linear relation with the reviews' sentiment
- 03 The total amplitude difference of the emotion arc of a story has a positive linear relation with the sentiment of the reviews
- 04 Stories with a rhythmic-beat sentiment elicit a more positive sentiment in reviews

# Methodology

## Data

~450 books

9 different genres

800K+ Goodreads reviews

## Analyses

### Features

- avg. book sentiment
- avg. review sentiment
- emotions exhibited in book
- emotions exhibited in reviews

### Story arc

- sentence-level sentiment scores of books (as performed by M. Jockers 2014)
- k-means clustering with DBA

# Dataset Curation

**Genres:** Bestseller (Jockers); children; history and biography; fantasy and paranormal; mystery, thriller, and crime; romance; young adult; scifi ; Classics (M. Walsh and Antoniak 2021a)

## Books

### Properties:

- Only unique fiction novels in English are included
- 50 most reviewed books selected per genre
- 2 new genres introduced based on “popular shelves”
- 40% of books written by female authors (gender neutral)
- Unique authors for each genre (to exclude book series)

## Meta-Data

- Author’s gender (manually tagged)
- List of genre tags (“shelves”) most frequently assigned to the book and their corresponding number of users



# Dataset Curation

(contd.)

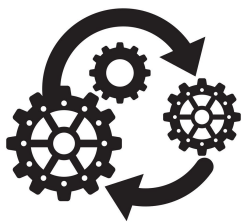
## Reviews

### Source:

- [UCSD Dataset](#): Romance, thriller, history, fantasy, children and young adult, sci-fi
- GoodReads scraping: Classics, Bestsellers

### Properties:

- Reviews are extracted only in English language
- Reviews for each book collated into one long text
- Scraped reviews are more recent than the rest
- Character names to be masked



# Data Preprocessing

## Books:

- Manual removal of unwanted text
- Sentence tokenization
- Masking of character names

## Reviews:

- Collated into one text per book
- Sentence tokenization
- Masking of author and character names

## Masking:

- Using “spacy” for named entity recognition
- Replacing all entities tagged as **PERSON**



# Masking Names

(model used: cardiffnlp/  
twitter-roberta-base-sentiment)

Dumbledore loves karaoke.

Compute

Computation time on cpu: 0.044 s

NEGATIVE

0.025

NEUTRAL

0.301

POSITIVE

0.674

Voldemort loves karaoke.

Compute

Computation time on cpu: 0.037 s

NEGATIVE

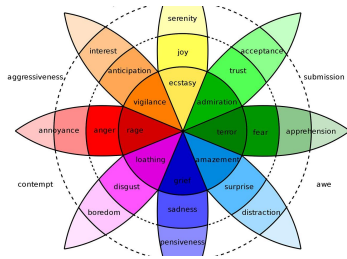
0.154

NEUTRAL

0.511

POSITIVE

0.335



# Analyses

## Sentiment analysis & emotion detection:

- Evaluation based on genre
- Transformers based approach
  - fine-tuned for emotion detection (Plutchik's wheel)
  - base model to be used - XLM-RoBERTa
- Dictionaries based approach
  - Using Syuzhet where “surprise and anticipation” are to be excluded due to ambiguity

## Story arc:

- Moving window score calculation
- K means clustering with DBA as distance metric

# Transformer based SA

(currently working)

## Models tested:

- DISTILBERT-BASE-EMOTION: 'f1': 0.45
- J-HARTMANN/EMOTION-ENGLISH-ROBERTA-LARGE: 'f1': 0.475
- \*CARDIFFNLP TWITTER-XLM-EMOTION: 'f1': 0.50

## Books:

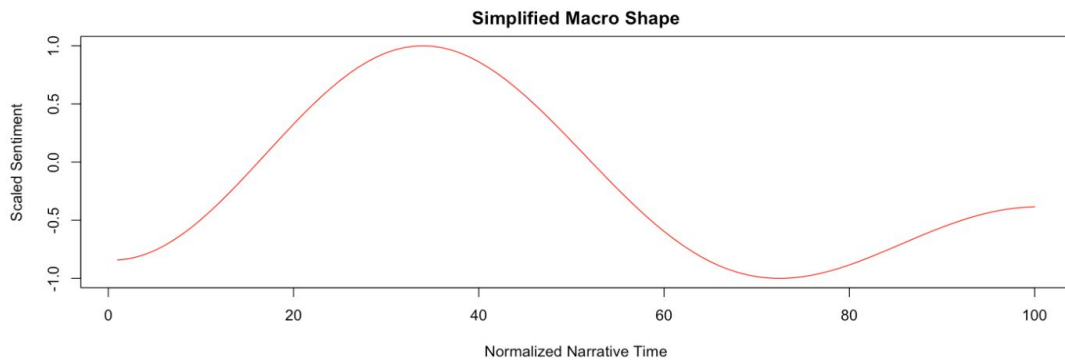
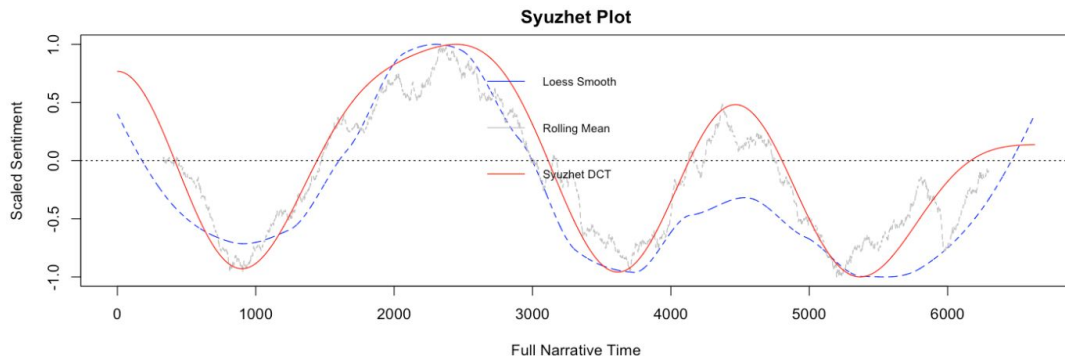
- Fine-tuned XLM-RoBERTa for SA and emotions
- Annotated dataset used - [DENS](#)

## Reviews:

- Fine-tuned XLM-RoBERTa over movie reviews (provided by Alessandro Fossati)
- Annotated datasets used - GoEmotions, CARER, Movie Reviews

# Sentiment story arc

(next steps)



Story arc of Harry Potter and the Sorcerer's Stone

# Conclusion

## Thoughts

- Working with **literary texts** is complex and requires better frameworks that account for the vast variety of styles and languages in literature
- **Online Reviews** as a means of reader response has its limitations which should be kept in mind while performing any form of analysis
- **Emotion detection** as a technique still needs refining and can be improved with the availability of, as always, more data!

## Acknowledgements

- Dr. Federico Pianzola
- Entire team at Ghent Centre for Digital Humanities
- Els Lefever & Pranaydeep Singh, LT3 Team
- H2020 CLS Infra Project

**Thank you!**