Archetypal Character Analysis on Literary Fiction for Youths

Tim Engelmann
University of California, Berkeley
tim.engelmann@berkeley.edu

Maya Chen
University of California, Berkeley
mayachen@berkeley.edu

Abstract

Leveraging the power of NLP, we analyse the lead characters of literature written for the youth. We investigate if the characters archetype is correlated with its gender, as well as with the overall success and public perception of the book. Furthermore, factors like the authors gender and the publishing year are considered. Through our work, we aim to identify if a gender gap in young adult literature exists.

1 Introduction

When interacting with industries and practices that have been perpetually dominated by certain norms, taking a data-driven approach is a powerful way to expose bias and inequalities.

We chose to focus on the topic of gender bias particularly after watching This Changes Everything (Donahue, 2018), a 2018 documentary about the gender gap in the Hollywood film industry. Many of the most poignant and impactful parts of the film used data from studies done on gendered representations in film and media; some statistics illustrated the gender gap on screen, such as the percentages of male-female speaking roles in G-rated films over the recent decades, while others exposed the disparities behind the scenes with female filmmakers and crew members.

One of the most important points the documentary made was that film and media have an astronomical influence on the ethos of a generation—and it begins with young children. What they see represented on screen directly shapes their beliefs, aspirations, and actions, in both good ways and bad. Taking this idea of early life media influence, we did some further research and found similar sentiments echoed by literary and psychology experts. A study on gender representation in children's books confirmed that "[children's books] are a rich source of information about gender, and that many express gender stereotypes more strongly than adult fiction. These findings suggest that popular children's books may be an underrecognized, inadvertent vehicle for perpetuating gender stereotypes and other gendered associations" (Lewis et al., 2022).

Related NLP work in this area shows that gender biases do exist in literary fiction from the past three centuries (Underwood et al., 2018). Our contribution to this space was to investigate character gender representation in relation with author gender, which could have substantial impact on attitudes towards gender roles during childhood. We elected to correlate character gender with the twelve Jungian archetypes (Mark, 2001), and made some interesting discoveries where character gender representation was concerned. During the annotation process we made changes to this classification framework to improve inter-annotator agreement, further clustering the twelve archetypes into four broader categories.

Over the course of a month, we annotated a selection of 1000 Goodreads children's book summaries from a total corpus of 13782 summaries, classifying labels for name, gender, and perceived archetype. We then built a BERT classification model around the overall dataset that included features such as coreference resolution for gender prediction and word association extraction using BookNLP (Bamman, 2021). We based our gender labeling model off of Lewis et al. (2022) use of pronouns in text to ascertain gender, a generally acknowledged practice in the NLP community (in contrast to using the person's name). As our model's accuracy was not high enough for archetypal prediction (0.340 for archetypes, 0.550 for categories), we used our hand-annotated labels for the subsequent data analysis.

Our findings show a surprising trend in youth literature that contradicted what Underwood et al.

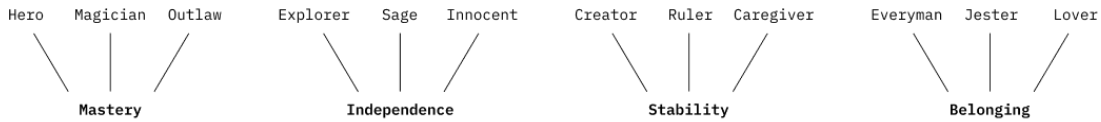


Figure 1: The 12 Archetypes and their 4 cardinal directions

(2018) found in their study — while they found a decline in female authorship and female characterization over the years, we found that females were better represented than expected in both those categories. In the final section of this paper, we briefly go into reasons for this occurrence in youth literature.

1.1 Research Questions

Our paper aims to answer the following questions related to gender and archetype in youth literature: How is gender represented among the different archetypes? Which archetypes are most popular/highly rated? Do certain publishers prefer certain archetypes? Do authors prefer to write about protagonists of the same gender?

2 Related Work

Character classification in NLP: We have found evidence of literary character classification schemes in both recent NLP research and in psychology. Psychologists Berry and Brown (2017) took a multidimensional approach, asking test participants to sort 40 different character types along the Big 5 personality scales (assertiveness, cooperativeness, extraversion, and conscientiousness). Though they had rejected using the Jungian archetypal framework, the 40 types they used were similar in that they revolved around the "role" a character plays in a story—importantly, this approach did not discriminate main characters from side characters. They then calculated cooperativeness and assertiveness among all 40 types. The result was an interesting remapping of plot as an expression of conflict between characters.

Similarly, Flekova and Gurevych (2015) classified characters according to the Big 5 personality dimensions using data collected from Personality Databank, a crowdsourced platform for rating characters by introversion and extraversion. Adjacent to this approach they used vector space semantics, specifically a "pre-trained word vector model created by the GloVe algorithm" to featurize characters' direct speech. The authors focused on verbs uttered by the character, selected the most frequent word sense (using WordNet), and then labeled them according to their semantic fields (e.g. creation, cognition, body, etc.). This yielded a convincing correlation between annotated personality and semantic features (Flekova and Gurevych, 2015).

Use of character archetypes: While both of these papers used personality psychology as a basis for literary character classification, we see potential in using an archetypal approach. Xara-Brasil et al. (2018) state that character archetypes have been deeply rooted in collective human culture for ages, making them compelling for classifying today's most popular brands. In this sense, the 12 archetypes are just as universally experienced as emotions are — applying this to literary characters, we also expect to see a previously hidden reflection of society's attitudes towards gender. While other studies have strayed away from using the 12 archetypes, we believe in their uniqueness as a classification scheme.

Previous findings on gender in literature: A thorough analysis on gender throughout eras of English fiction has already been done and acts as a significant basis for us to compare our own findings to. Underwood et al. (2018) found several contradicting trends in fiction published between the nineteenth and twenty-first centuries (306 years): gender roles have become more fluid, but at the same time books are featuring less female characters than before. They also found that female authorship has been declining. A key feature of this paper is how the authors mapped out occurrences (say, percentage of words used in characterization that described women) along a time axis, contextualizing these trends in history. A steep decline after the mid-1800s then found its way to a spike in the 1970s (Underwood et al., 2018). What does this say about societal events and gender in fiction? We would like to approach our annotation results in a similar way. Additional visualizations we drew inspiration from include Figures 11 through 15 (The difference between a

word's frequency in descriptions of women, and of men.), which used word frequencies to map changing gender portrayals of emotion, space, and physicality (Underwood et al., 2018).

There remains one question: how do these results stack up in children's literature? Lewis et al. (2022) showed that there was a downward trend (as they moved towards more recently published books) in the proportion of male main characters, and "more main characters of indeterminate gender" Lewis et al. (2022). This generally echoes the findings by Underwood et al. (2018) regarding gender roles. But they also found less mixing when it came to character gender and target audience: "children's books more frequently read to girls tend to have both more female content and more female characters, and children's books more frequently read to boys tend to have both more male content and more male characters" (Lewis et al., 2022).

Approaches for related NLP work: We've found that past work in this topic area uses a variety of NLP methods for identifying and extracting features of a character, these methods not being specific to literary character classification. A popular method of identifying characters in a text is through named entity recognition (Flekova and Gurevych, 2015; Adukia et al., 2021; Labatut and Bost, 2020). Labatut and Bost (2020) included in their paper a comprehensive overview of the character extraction process, which starts at detecting occurrences (p.6). For our research, identifying gender will be relevant.

Wan et al. (2019) present a Naive Bayes classifier approach trained on the NLTK name corpus. Most informative features proved to be the ending letters of the name, such that names ending in a, e, and i were predicted to be female and names ending in k, o, r, s, and t to be male. Adukia et al. (2021) uses a similar first name-based approach. We will be using pronouns to infer character gender, as stated in class and shown in (Lewis et al., 2022).

Finally, Pollak et al. (2020) uses word embeddings (FastText) to generate nearest semantic neighbors for given gendered words, yielding relationships that were not further researched upon but were nonetheless intriguing. While this example is not in direct relation with what we hope to do, a possible thread we want to explore is how adjectives used in book reviews can be correlated with character genders or archetypes. Would a specific archetype yield a consistent set of adjectives?

3 Dataset

We decided to base our analysis primarily on the goodreads.com dataset from the year 2017 (Wan and McAuley, 2018; Wan et al., 2019). By applying several filters, we ensured that the considered descriptions would entail enough character information to assign an archetype. For the first phase of the project we used the following filters: Target audience (Youth), Language (English), Description Length (between 500 and 2000 characters) and Character Detection (at least 3 times, for details refer to Section 4.1). This process yielded a total of 35102 books. For the first 1000 observations we had two independent annotators assign an archetype and gender label to the identified character. A third annotator adjudicated these labels, which we will consider as ground truth for the following sections. We reached a Cohen's Kappa of 0.40 for archetype annotations and of 0.93 for gender annotations. When clustering the 12 archetypes by their cardinal direction (refer to Section 2), only 4 labels remain (Stability, Mastery, Independence, Belonging), which yielded in a Cohen's Kappa of 0.48. For the second phase of the project, we also wanted to include author information. We only considered books, that were written by a single author, and for which our algorithm (described in Section 4.2) was able to predict the authors gender. Therefore, our final dataset consisted of 13782 books. The majority of these books were published between 2004 and 2017. An overview of the data model can be found in Appendix 6.

4 Methodology

Our work relies on four distinct NLP models. Firstly, the Character Identification model that detects the first name, last name and number of appearances of a character within a book description. Secondly, the Gender Prediction model that predicts the gender of the character or author based on its name, the book description and reviews where the character/author is mentioned. Thirdly, the Archetype Classification model that predicts the archetype of a character given its name and sentences it appears in. Lastly, the Association Extraction model that extracts related words such as possessive words related to a character given its name and text it appears in. In the following we will outline how we have built these models. And

overview of the entire system can be found in Appendix 7.

4.1 Character Identification

To identify persons within a book description we performed Named Entity Recognition with the "encore-web-sm" model of spaCy (Explosion, 2021). We compared its performance to the NLTK (2022) library, which it outperformed. For each detected person we match different versions of the name (eg. only first name or only last name) to a singular instance. We then check which character was mentioned most often and return its name, split in first and last name.

4.2 Gender Prediction

To predict the characters gender we first performed coreference resolution on text passages where the character is mentioned. We relied on the "neuralcoref" model of spaCy, first proposed by Clark and Manning (2016). We extracted all coreferences related to the previously identified character within the book description and the 10 most popular book reviews, containing the characters name. We then summed over all occurrences of ['she', 'her', 'herself'] and ['he', 'him', 'his', 'hisself', 'himself']. We predicted "female" if the first count out weighted the second and "male" the other way round. If both counts were equal or zero we predicted the gender "other". We used the exact same model to predict the authors gender using the 10 most popular book reviews where the author is mentioned as input. For an evaluation of the models performance please refer to Section 5.1.

4.3 Archetype Classification

We trained and tested several different models for archetype prediction, however were not able to exceed an accuracy of 0.360 for archetype prediction and 0.550 for archetype-cluster prediction. We, therefore, decided to base our analysis in Section 6 related to the characters archetype only on the first 1000 hand-annotated labels. However, for completeness we do want to describe the best performing model here and will compare it to several other models outlined in Section 5.2. Our most promising approach was a BERT model trained and evaluated on character related sentences. Using the coreference resolution algorithm outlined above, we extracted all sentences related to the character (i.e. including its name or a coreference). We then used the first 600 book descriptions (3307 character related sentences) of our hand-annotated data as training data and the next 200 descriptions (1117 character related sentences) as evaluation data. We used the BERT model proposed by Devlin et al. (2018) and provided through the Hugging Face platform. As hyper-parameters we chose the model base-cased with a batch size of 32, max token size of 256 and embedding size of 768. Finally, we predicted the archetype for the remaining 200 observations (1126 character related sentences) and adjudicated the individual sentence predictions in a winner-takes-it-all manner, leading to an accuracy of 0.360.

4.4 Word Association Extraction

We generated additional insights about specific archetypes by extracting word associations related to the identified character. To do so we relied on the BookNLP library by Bamman (2021), which allowed us to extract objects the character posses, verbs where the character acts as agent, patient, or possessive. For each book we adjudicated its description with the 10 most popular reviews mentioning the character and used that as in input for the BookNLP model. We then saved the 10 most common associations for each of the 4 above mentioned categories.

5 Evaluation of models

We performed a detailed performance analysis for our gender prediction and archetype classification models. We used the hand-annotated data as ground truth when calculating performance metrics. In the following we will elaborate on different approaches we took and describe how well these performed.

5.1 Gender Prediction

We initially used a Naive Bayes Classifier, trained on the NLTK name book corpus, to predict the characters gender based on its first name. However, we only reached an accuracy of 0.71 and therefore decided to switch to a pronoun based approach. We first used the build-in functionality of the BookNLP library (Bamman, 2021), which uses coreferences to infer the characters gender. It achieved an accuracy of 0.87. However, for the overall system architecture, we were required to perform coreference resolution separately and it was not feasible to run both BookNLP and neuralcoref on all 13782 books. We, therefore, built a the gender predic-

tion algorithm as outlined in Section 4.2. It also achieves an accuracy of 0.87. However, when looking at the class specific F1 scores, we found that it slightly outperforms the BookNLP algorithm for the classes "male" and "female". As most of our analysis compares those two classes, this represented an additional benefit. Lastly, as a reference we compared the predictions of one individual annotator with the adjudicated labels and determined an "human" accuracy of 0.98. A tabular overview of the mentioned performance metrics can be found in Table 1 and the respective confusion matrices in Appendix 8.

5.2 Archetype Classification

As described in Section 4.3 we trained a BERT model on character related sentences. After adjudicating all sentences related to a single character, we achieved an accuracy of 0.360 with 95% confidence intervals of [0.293, 0.427]. This represented an improvement of 12% compared to the majority class baseline of 0.240. When looking at the confusion matrix (see Figure 9), it became clear that the number of classes might be too high for the relatively little annotated observations that we had and that the dataset was highly imbalanced. To address the first problem we decided to build models which predict the archetype's overarching category instead (Stability, Belonging, Independence, Mastery). Doing so, we reached an accuracy of 0.550 with 95% confidence intervals of [0.481 0.619] (see Figure 2). We then also trained a Logistic regression classifier on all character related sentences and used cross-validation to determine the regularization parameter C. This classifier reached an accuracy of 0.465. We also trained several models on the original book description text instead of individual sentences. For all models this decreased performance (see Table 2). In addition, we tried several other approaches, which all lowered the overall performance. These include: Replacing all mentions of the character with an special [ENTITY] tag, replacing all mentions of the character with the main coreference, using book review data related to the character as additional training/evaluation data and predicting the archetype at first and then clustering them afterwards according to their cardinal direction.

6 Analysis

Our reasonably high accuracy of 0.870 for gender prediction allowed us to run our gender predictor

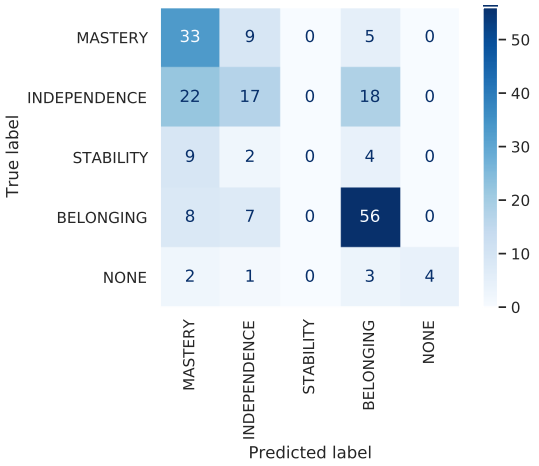


Figure 2: Confusion matrix of BERT archetype cluster prediction model, achieving an accuracy of 0.550

over our entire dataset of book summaries. We used gender prediction to map character gender distributions, author gender distributions, and the relationship between the two. We also visualized character gender distributed by archetypal category.

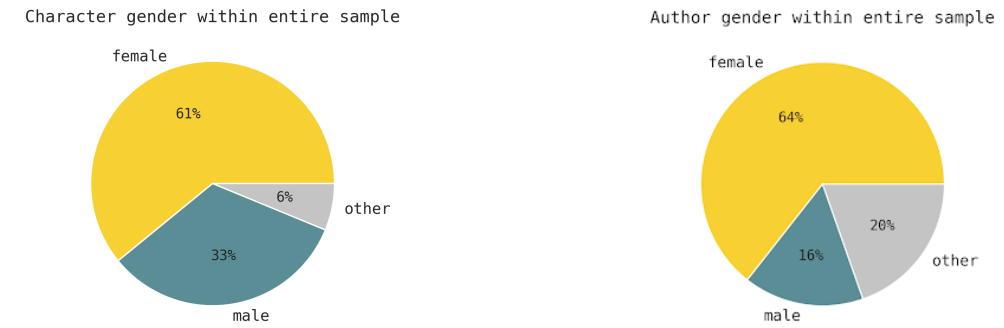
6.1 Overall gender distribution

We found that females made up around 60 percent of all characters and authors, respectively, in our total dataset (Figure 4, Figure 5). This is contrastive to what Underwood, Bamman, and Lee found, which was a "story of decline" in female authorship between 1800 and 1960 (Underwood et al., 2018). Underwood and team go on to create several more distributions using alternative sources and methods to avoid bias—nonetheless, the percentage of books written by women inches close but never surpasses the 50 percent mark, even as we enter the 21st century (Underwood et al., 2018). There's a real possibility that the female majority we're seeing here is due to the Goodreads author demographic being predominantly female (Thelwall, 2019). Albeit using a flawed system of gender classification (determining gender from first name), Thelwall found that a random sample of 500,000 books from the Goodreads website was disproportionately female-authored: "51 percent of books with under 31 reviews had a female author and 62 percent [of books] with more [than 31 reviews had a female author]" (Thelwall, 2019). While this observation in itself warrants further investigation, we conclude here that the numbers for female representation in authorship and characterization marks an overall shift in literary status quo towards a place

where diversity of genre does not constrain female presence.

6.2 Gender - archetype distribution

Distributing character gender by archetype yielded some additional interesting results (see Figure 4). Interestingly, in the hero and explorer categories (two archetypes that we thought would be predominantly male), we found that around 36 percent of all female protagonists fell in these two categories versus 26 percent for male protagonists. Instead, we found males to be predominantly lover archetypes (about 24 percent). None of the related work we found had findings about archetype and character gender, but work done by Underwood et al. (2018) showed that the gender binary has become more blurred over time. We find that our results also fit



(a) Character gender distribution

(b) Author gender distribution

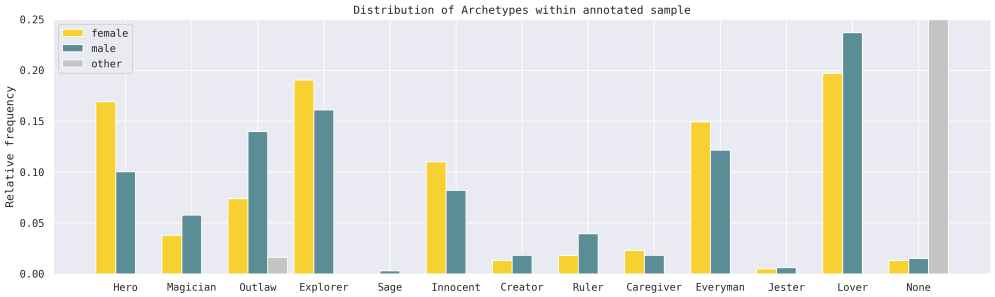


Figure 4: Distribution of archetypes in annotated data, separated by gender.

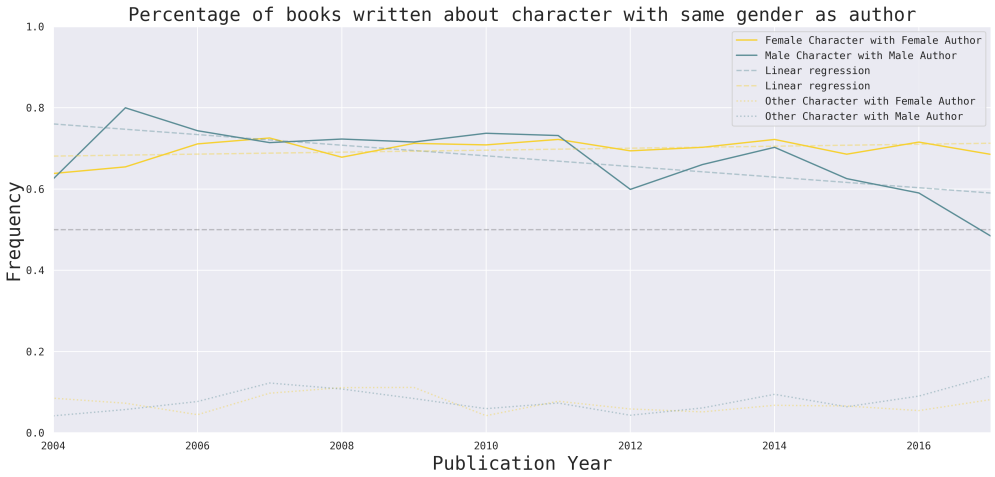


Figure 5: Author to character gender correlation

with this conclusion; protagonists cannot be easily sorted into archetypes based on gender.

6.3 Author - character gender correlation

We already know that authors tend to write about main characters of their own gender (Underwood et al., 2018). From our data, it appears that in the last decade or so, male authors have started to write less about male protagonists (linear regression for male author - male character relation: intercept 27.0%, slope -1.3%, p-value 0.010). This could signal a decline in the polarity of the author-character gender correlation. On the other hand, female author to female main character occurrences have remained steady over time. Paired with Lewis et al.'s analysis of main character gender trends over time Lewis et al. (2022), in which male characters are seen to be growing less represented, we arrive at an interesting inflection point for gender in youth literature.

6.4 Additional findings

Please see the appendix for additional visualizations we created for character archetypes; some analyses we performed included correlating word associations with each archetypal category, evaluating popularity and acclaim, and more.

7 Conclusion

When interacting with industries and practices that have been perpetually dominated by certain norms, taking a data-driven approach is a powerful way to expose bias and inequalities. We posed this question: how can NLP classification methods be used to shed light on gendered associations within children's literature? Our research project aimed to provide radical transparency towards unconscious biases in the creation and perpetuation of common character archetypes, and hoped to shed light on new, more inclusive creative directions for the children's literature of this (and the next) generation.

We found, surprisingly, that many of our initial hypotheses were contradicted. While we thought that there would be significant bias against female authors and characters based on previous studies, we instead found that they were more well represented in youth literature. Though we also happened upon evidence that female authors are disproportionately represented specifically on Goodreads, this tells us that the story is not over for gender representations in literature. If we have stumbled upon a weak signal for the future development of youth media, we leave the door wide open for further work into this specific, critical genre of literary fiction.

Acknowledgments

This project is part of the course INFO 259 - Natural Language Processing taught by David Bamman at the University of California, Berkeley.

References

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2021. What we teach about race and gender: Representation in images and text of children's books.

David Bamman. 2021. Github - booknlp/booknlp: Booknlp, a natural language processing pipeline for books.

Matthew Berry and Steven Brown. 2017. A classification scheme for literary characters. *Psychological Thought*, 10:288–302.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *CoRR*, abs/1609.08667.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tom Donahue. 2018. This changes everything.

Explosion. 2021. spacy models documentation.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal. Association for Computational Linguistics.

Vincent Labatut and Xavier Bost. 2020. Extraction and analysis of fictional character networks. *ACM Computing Surveys*, 52(5):1–40.

Molly Lewis, Matt Cooper Borkenhagen, Ellen Converse, Gary Lupyan, and Mark S. Seidenberg. 2022. What might books be teaching young children about gender? *Psychological science*, 33:33–47.

Margaret. Mark. 2001. *The hero and the outlaw : building extraordinary brands through the power of archetypes*. McGraw-Hill, New York.

NLTK. 2022. Nltk :: Natural language toolkit.

S. Pollak, M. Martinc, and K.M. Poni? 2020. Natural language processing for literary text analysis: Word-embeddings-based analysis of zofka kveder's work. *CEUR Workshop Proceedings*, 2607:33–42.

Mike A Thelwall. 2019. Reader and author gender and genre in goodreads. *Journal of Librarianship and Information Science*, 51:403 – 430.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 3(2).

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Duarte Xara-Brasil, Kavita Miadaira Hamza, and Percy Marquina. 2018. The meaning of a brand? an archetypal approach. *Revista de Gestao*, 25:142–159.

A Figures

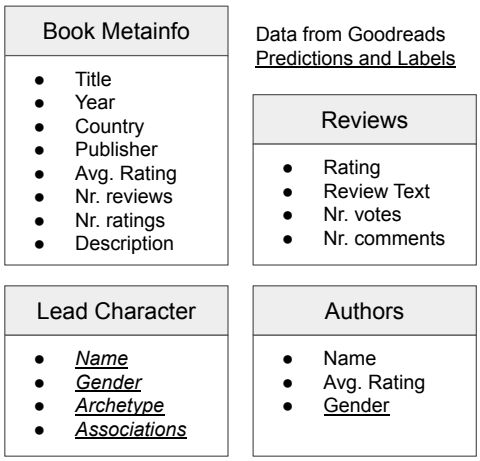


Figure 6: Data model outline

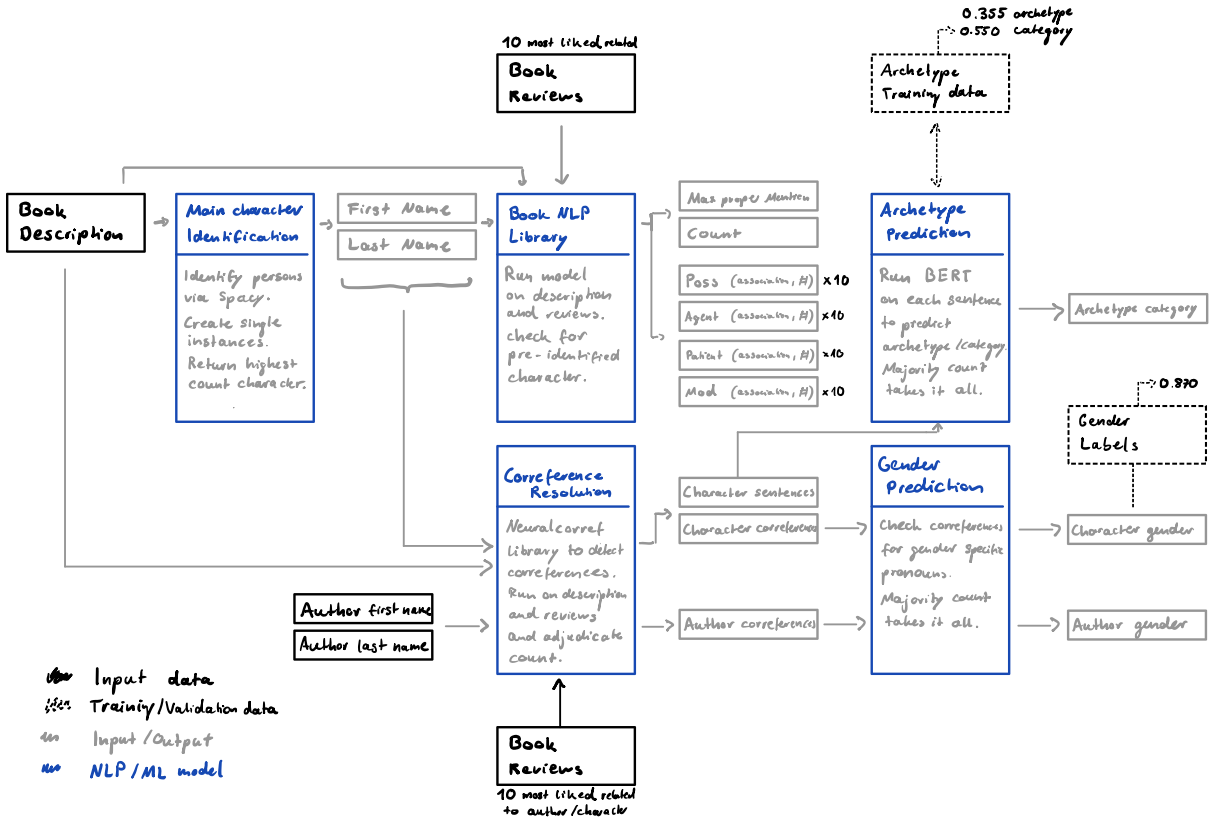
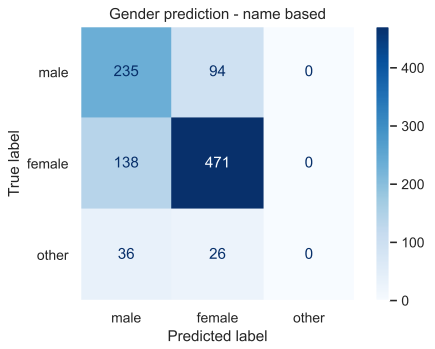
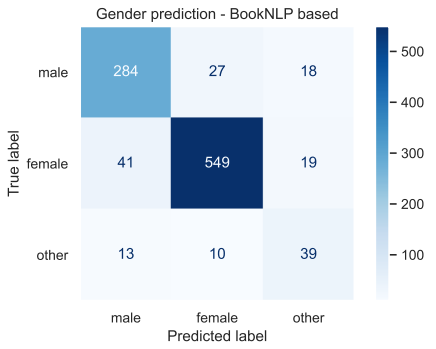


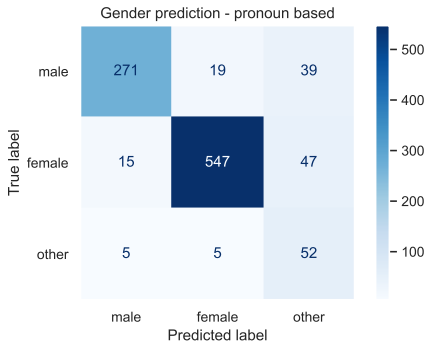
Figure 7: A comprehensive overview of the interaction of our NLP models



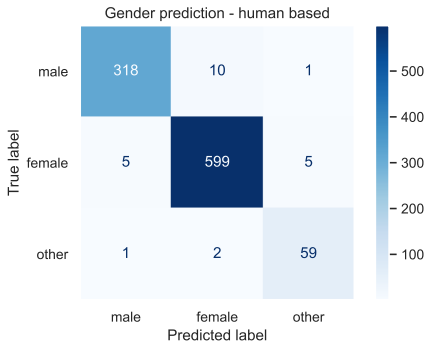
(a) Name based accuracy: 0.71



(b) BookNLP based accuracy: 0.87



(c) Pronoun based accuracy: 0.87



(d) Human based accuracy: 0.98

Figure 8: Confusion matrices and accuracy of several different gender prediction models.

Model	Accuracy	F1 (male)	F1 (female)	F1 (other)
Name based	0.71	0.64	0.79	0.00
BookNLP based	0.87	0.85	0.92	0.57
Pronoun based	0.87	0.87	0.93	0.52
Human based	0.98	0.97	0.98	0.93

Table 1: Overall accuracy and class specific F1 scores for different gender prediction models.

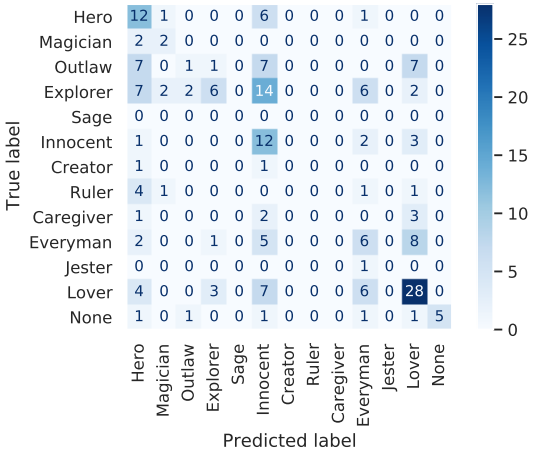


Figure 9: Confusion matrix of BERT archetype prediction model, achieving an accuracy of 0.360

Model	Dataset	Dev Accuracy	Sentence Test Accuracy	Test Accuracy	Parameters
Majority	text			0.355	
BoW	text			0.510	
BERT	text	0.475		0.505 [0.436 0.574]	
BERT	sentence	0.504	0.491 [0.460 0.522]	0.550 [0.481 0.619]	
Logistic	text	0.570		0.515 [0.446 0.584]	C: 10
Logistic	sentence	0.467	0.451 [0.420 0.481]	0.465 [0.396 0.534]	C: 0.4

Table 2: Accuracy of different models used to predict archetype clusters (Stability, Belonging, Mastery, Independence). ”Text” means the model has been trained on the entire book description, while ”Sentence” refers to model trained on character related sentences. Sentence predictions are then adjudicated to yield a single label for each book.

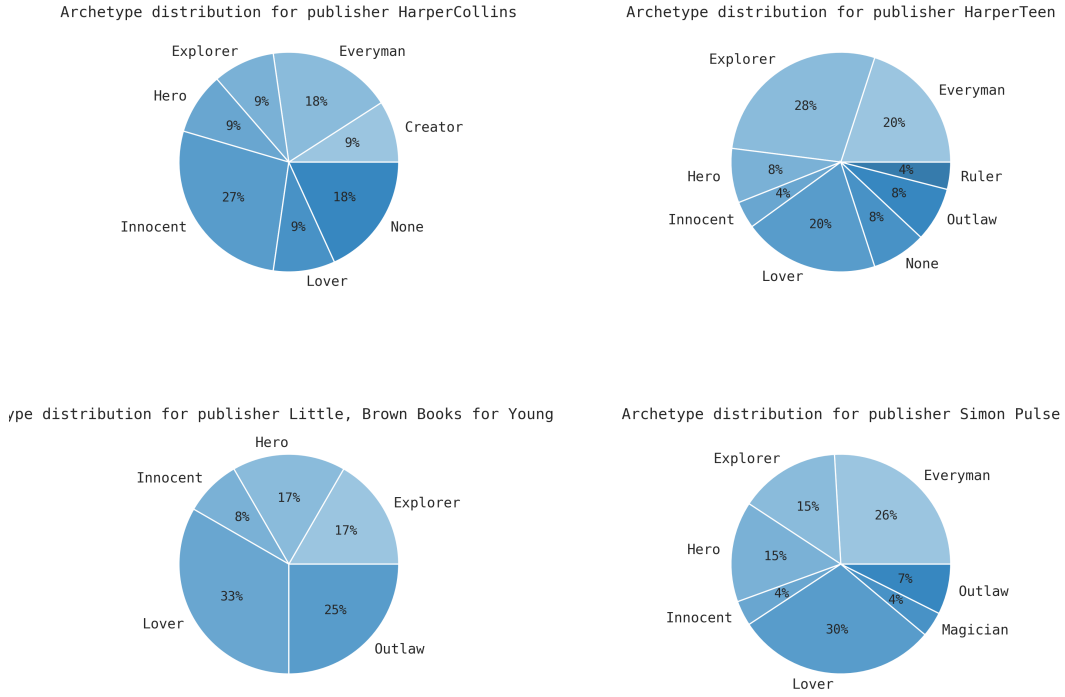


Figure 10: Are there publishers who are more prone to publishing books about a character of a specific archetype?

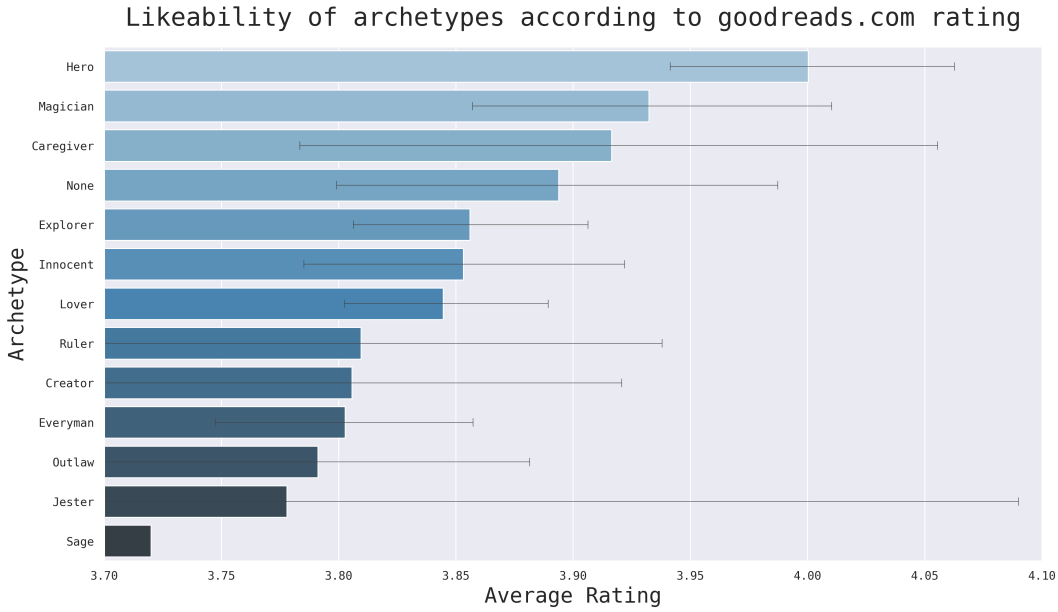


Figure 11: We found that books with Hero archetypes were rated the highest.

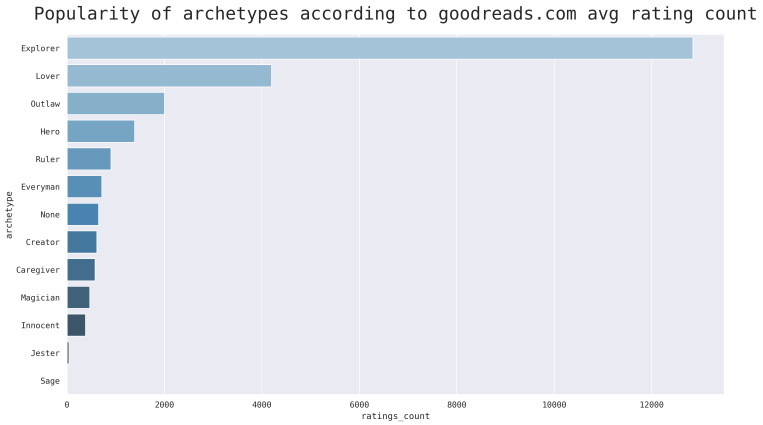
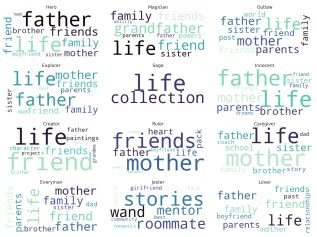


Figure 12: We found that books with Explorer archetypes received the most ratings.

(a) Agentive relationships

(b) Patient relationships



(c) Possessive relationships

Figure 13: Word association (coreference) findings. We couldn't find any outstanding differences here between the different word clouds and archetypes, but it was nevertheless interesting to map out the word associations. It left us with more questions, such as: in the Possessive word cloud, why is father most prominent for Heroes versus Rulers, where mother is dominant?