36-668: Final Project

Structural Topic Modeling as a Means of Comparing Love-Related Literature

Andrew Shih

2024-12-11

Table of contents

Abstract

It is a commonly held notion that love is a universal concept, especially in literature, but there is a lack of research that studies differences in how it is portrayed across various dimensions. Existing studies characterize love geographically but do not apply their methods to other levels of analysis (Baumard et al. 2022). Using corpora comprised of love poems and love letters, I used the `stm` package to perform structural topic modelling to detect clusters of topics that distinguish love-related literature along three selected dimensions: **era**, **genre**, and **authorship**. Running T-tests on these models showed that love is largely expressed similarly along different points of these dimensions, with minor exceptions. Future iterations

of these methods applied onto other types of literature can give insight into whether love is truly universal in literature compared to other subjects.

1 Introduction

While love is a universal human experience that appears as a literary theme across all eras and cultures, studies on how it varies across these stratifications are limited. More can be understood about what topics and motifs can distinguish love-related literature in **different eras**, **writing styles**, and **authorship types**. This study will compare pieces of love-related literature and identify the thematic and linguistic features that distinguish them across the dimensions mentioned above. As literature reflects societal values, examining how love is portrayed in relation to time, author gender, and literary genres can give a preliminary view into how broader historical contexts have influenced human notions of love.

2 Data

This experiment has two corpora: the Love Poems corpus—containing **318** poems—and the Love Letters corpus—containing **90** letters.

2.1 The Love Poems corpus

A user on Kaggle by the username Ishnoor assembled a dataset (Ishnoor 2017) containing poems from poetryfoundation.org covering a wide variety of genres and associated metadata. After downloading the dataset as a CSV file, I filtered for poems in the love category to form the corpus I am using for this experiment. The corpus comprises metadata such that each text corresponds to a well-known author and the period in which it was written. poetryfoundation.org designated each poem as being from either the Renaissance or the Modern era, which disambiguates era classification.

Table 1: Breakdown of Poems and Tokens by Era

	Era	**No. of Poems**	**No. of Tokens**
	Modern	75	8,554
	Renaissance	243	32,954
Total	—	318	41,508

Table 1 shows the breakdown of poem and token volume across the two eras after preprocessing:

The corpus lacks diversity in certain areas. First, the poems are primarily written by English-language authors, so the corpus fails to capture variances that may arise due to differences in authors' country of origin. Also, authorship skews heavily male: only 9% of the corpus is female-authored. In a 2010 report published by the organization, VIDA: Women in Literary Arts, male-authored poetry outnumbers female-authored poetry 2-to-1 in the magazine, *Poetry* (O'Rourke 2011). The corpus exacerbates this gender imbalance. Because the gender of a poem's author may be a confounder of the era it was written, discounting this imbalance could introduce unaccounted bias into the study. Lastly, the corpus contains multiple works from some authors while containing as little as one poem from other authors, which could lead to an overrepresentation of a particular writing style that could skew the results of this analysis.

2.2 The Love Letters corpus

Kaggle user Sreeram Venkitesh created a dataset (Venkitesh 201AD) that scraped letters from theromantic.com and countryliving.com. To assemble the corpus, for each letter, I extracted its contents, added metadata about the author (as designated by the aforementioned websites), and included the author's gender (from external research).

Table 2: Breakdown of Letters and Tokens by Gender

	Gender	**No. of Letters**	**No. of Tokens**
	Female	26	4,200
	Male	64	9,626
Total	—	90	13,826

Table 2 shows the breakdown of letter and token volume across the male and female genders after pre-processing:

The corpus lacks diversity in certain areas. Again, authorship skews heavily male. In contrast to the Love Poems corpus, however, this corpus contains no more than 3 works from any one author, which is a more even distribution of individual authors and reduces the influence of any one writing style on the outcome of this analysis.

3 Methods

This study examined how sentiments about love differ along three dimensions: across periods, among author gender, and between literary genres. The first dimension was analyzed with the Love Poems corpus, the second dimension was analyzed with the Love Letters corpus, and the third dimension incorporated both corpora.

I used structural topic modeling (STM) for this analysis because it allowed me to incorporate document-level metadata (i.e. information about the era a poem was written) into the model, which is particularly beneficial for this study because there are already relatively few texts in the corpora. In particular, I used the `stm` package in R, which offers many customization options when modeling on topics. Unlike count- or frequency-based cluster analysis, STM recognizes that a word can have multiple meanings depending on its context. Thus, while a typical clustering algorithm (i.e. K-means clustering) could be used for the task at hand, STM improves upon it by providing more data to the model in a way that provides a more informed answer to the research question. Furthermore, topic models are mixture models, which assign probabilities to each text belonging to a particular "topic" identified by the model. STM uses Bayesian techniques to iteratively update the initially random probabilities, ensuring the results have as low bias as possible.

My methods also inform what variables I chose to include in my dataset. The `stm()` method, which builds the topic model, has a `prevalence` argument that takes in a formula where I can specify the causal relationship between poem topical content and any metadata. For example, I included an `era` column to the Love Poems corpus so that all relevant data are consolidated in one dataset, making the workstream more efficient. A 2019 paper published in the *Journal of Statistical Software* provides a general walkthrough of how to use the `stm` package, as well as an application of it to the CMU 2008 Political Blog Corpus (Roberts, Stewart, and Tingley 2019). The study contrasted topical content prevalence between Liberal- and Conservative-aligned blogs, which has a similar mechanical premise as my research question and supports my usage of the `stm` package on my dataset.

I primarily used the `tm` package, which has text mining functionalities suitable for text preprocessing. I decided not to lemmatize or stem tokens because verb tense and word forms hypothetically shift over time (for example, infinitives are used during one era, and present participles are used in another era), and I want to capture this linguistic effect in my results if it is true. For all 3 analyses, the stopwords I filtered out of the pool of tokens included common stopwords from the SMART stopword list. For the Love Poems corpus, there was additionally a set of unique stopwords I created to account for bibliographic information irrelevant to poem content or words too archaic for the `tm` package to recognize as being synonymous with their modern-day definitions (i.e. "hath", "doth", thou"). No custom stopword list was needed for the Love Letters corpus. I also filtered out words that did not appear in more than **15** poems, a relatively high threshold, to compensate for the relatively small dataset size.

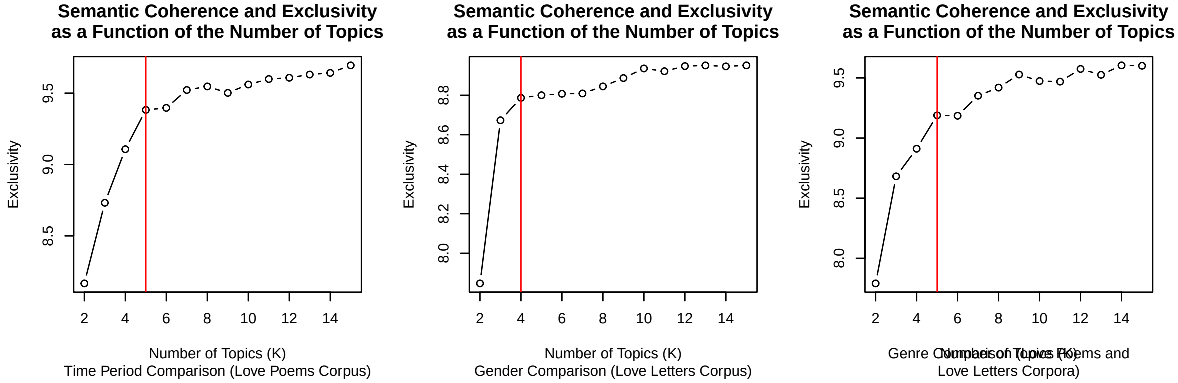

Figure 1: Exclusivity Increases as K Increases

The `stm` package contains several methods I used to aid model selection, hyperparameter tuning, and model evaluation. I determined the optimal number of topics for my model using exclusivity as a metric. While semantic coherence and exclusivity are both valid metrics to evaluate a topic model, I chose to prioritize semantic coherence as it provides a more interpretable conclusion and better answers the research question. However, because semantic coherence usually always decreases as the number of topics in a model increases, I resorted to using exclusivity, which usually always increases, as a metric to decide the optimal number of topics to model with. I used the elbow method on the plots in Figure 1 to pick the optimal number of topics for each analysis. For the period and genre comparisons, the optimal number was **5**. For the gender comparison, the optimal number was **4**.

Next, using the optimal number of topics found in the previous step, I used `stm` functionalities to select a model based on semantic coherence. The modeling process involved 20 runs using Latent Dirichlet Allocation (the default process), where each run was given a maximum of 75 iterations to reach convergence. These numbers were arbitrarily chosen at first, but I believed them to be appropriate as—for the period comparison study—16 of the 20 runs converged before 75 iterations, ensuring stability and robustness. For each analysis, I plotted exclusivity as a function of semantic coherence for the 4 models with the highest likelihood in Figure 2 to pick the "best" model, prioritizing semantic coherence.

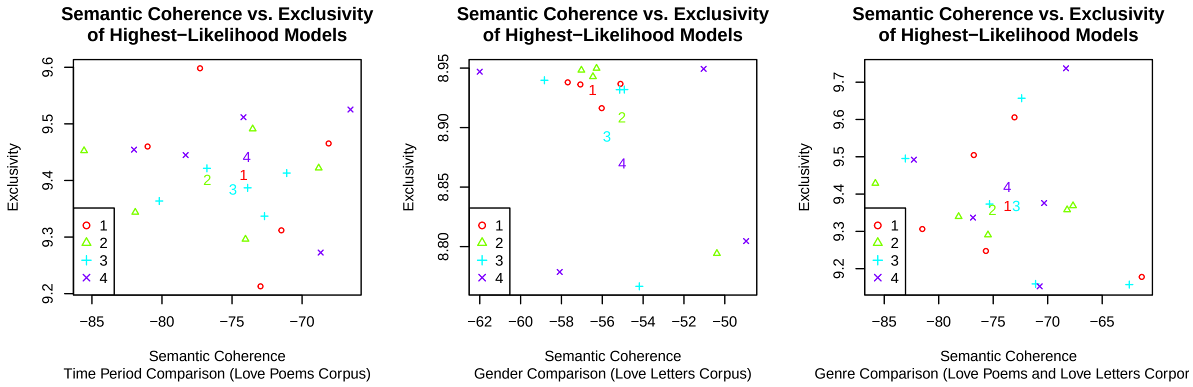

Figure 2: Selecting Among Candidate Models Based on Highest Semantic Coherence

For example, **Model 3** was chosen as the "best" model for the period comparison analysis. For each model selected, I then estimated a regression to determine whether the relationship between an era and the topics identified in the model is statistically significant.

4 Results

The topics chosen by the "best" topic models for each study are exemplified in Table 3, Table 4, and Table 5.

Table 3: Time Comparison Study - Topics Chosen by the Optimal Models and Exemplary Words

Topic	Top Words
1 - Action and Agency	one, yet, let, make, might, give, face
2 - Related to Time	now, like, see, time, still, whose, since
3 - Idealization and Admiration	will, shall, eyes, can, may, fair, beauty
4 - Conviction and Persistence	love, never, must, though, true, loves, well
5 - Possession and Movement	heart, come, mine, say, thus, world, leave

Table 4: Gender Comparison Study - Topics Chosen by the Optimal Models and Exemplary Words

Topic	Top Words
1 - Emotion and Existentialism	heart, can, much, ever, life, now, god
2 - Determination	will, shall, know, may, write, give, every
3 - Contemplation and Negation	one, never, think, day, letter, feel, nothing
4 - Affection	love, see, like, dear, little, long, without

Table 5: Genre Comparison Study - Topics Chosen by the Optimal Models and Exemplary Words

Topic	Top Words
1 - Contemplation and Action	will, know, much, come, see, tell, think
2 - Individuality and Agency	one, can, now, shall, mine, well, long
3 - Commitment and Persistence	love, let, never, must, since, day, true
4 - Aesthetic and Sensory Descriptions	yet, eyes, like, still, make, sweet, beauty
5 - Reflection on the Inner Self	heart, may, fair, light, night, soul, mind

The following results assume the 3 regression models are correct and we use an α of 0.05 for all of them.

While there is a statistically significant relationship between the prevalences of the respective Topics 1-4 and the **era** in which a poem was written, there is a statistically insignificant relationship between the prevalence of Topic 5 and the **era**. For example, for Topic 2, there is a 95% CI $[-0.153, -0.081]$, $p = 6.76 \times 10^{-12}$ increase in topic proportion if a poem is a Renaissance-era one. A negative coefficient, in this case, represents higher prevalence in Modern-era poems, and vice versa. There is no statistically significant relationship between the prevalences of any of the respective topics and the **gender** of a letter's author. The prevalences of the respective Topics 1, 4, and 5 have a statistically significant relationship to the **genre** of the love-related literature, while the relationship is statistically insignificant for Topics 2 and 3.

Figure 3 shows alternative views of these results. Taking the period comparison analysis as an example, Topics 1, 3, and 4 tend to appear more in Renaissance-era poems, while Topics 2 tends to appear more in Modern-era poems. Topic 5 straddles both eras.

5 Discussion

The topics identified and the way that the exemplary words were grouped by the model are plausible in terms of relating them to love as a concept. For the **time period** comparison study, despite the t-tests showing statistically significant relationships between some of the topic proportions and the era in which a poem was written, the actual magnitudes of the topic proportion differences may not be large enough to have a practical impact. The most distinguishing topic (Topic 2 - Related to Time) can be explained by how Modern poetry is more concerned with describing the present moment and reflecting on the human experience than Renaissance poetry. Topics 1, 4, and 5 might lean more toward the Renaissance side to reflect the more ardent and classical aspects of longing and pursuing someone. One can surmise from the relatively small differences in topic proportions that the differences between Modern- and Renaissance-era poems are not substantial.

The results of the **gender** comparison study imply that male and female authors of love letters do not write substantially differently as all topics span both the Male and Female space in Figure 3.

In the **genre** comparison study, Topic 4 (Aesthetic and Sensory Descriptions) stands out as being more prevalent in poems, which reflect their more artistic and stylistic nature compared to letters. In contrast, Topic 1 (Contemplation and Action) stands out as being more prevalent in letters, which reflect their communicative and practical purposes. Across all three analyses, the differences in topic proportions for these two topics are the largest. While only two genres were compared in this analysis, it seems that love-related literature is the most discernible along the lines of genre. However, the relatively small mean topic proportion differences across all prominent topics reflect that love will generally transcend contexts and histories. Performing this same study on the literature of other subjects can provide a means of comparison to see if those other subjects are literarily universal.

In a previous iteration of the **time period** comparison study, there was a topic that solely contained archaic words and leaned more toward the Renaissance side. Including archaic words in the custom stopword list has eliminated this topic, which leads me to believe that linguistic features will take precedence over topical features in these kinds of sentiment analyses.

A limitation of this study, however, is the narrowed focus of the **gender** study to just the Love Letters corpus (the study could have also included the Love Poems corpus). The usage of the `selectModel()` function in the `stm` package requires specifying a regression model. I lack the domain knowledge to truly know if there are interaction terms or confounding variables between gender and writing style (which is a variable that I would have to control on if I used

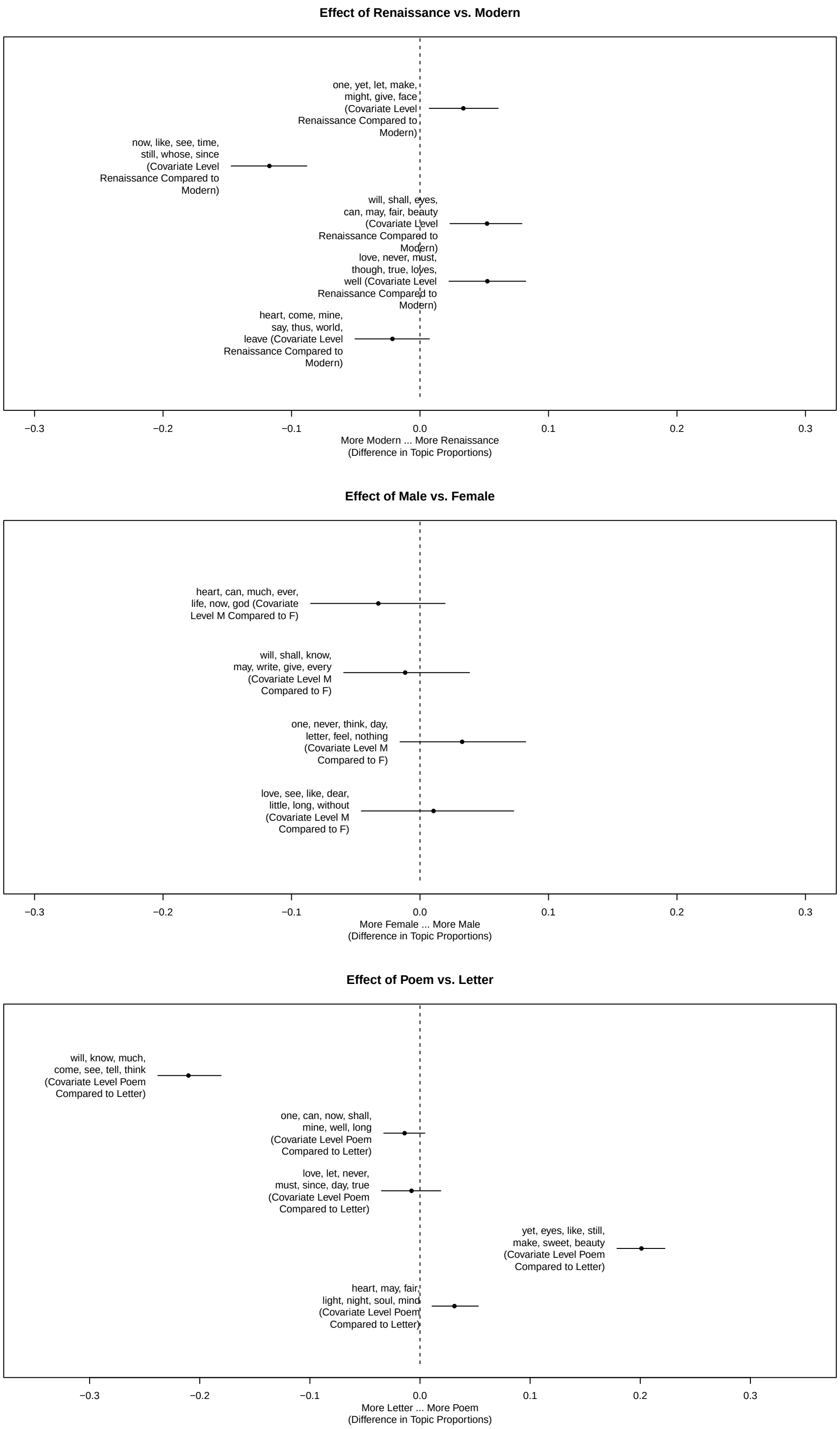

Figure 3: Contrasting Topical Prevalence on Different Dimensions

both letters and poems in this analysis). While the results may not reflect differences in topic prevalence between genders across multiple genres, I achieved clearer findings for letter-style writing. To resolve this issue, future ANOVA tests can be used to compare regression models that differ by an interaction term to determine if the interaction term significantly contributes to explaining differences in topic prevalence. Collaboration with domain experts can also help determine the presence of such interaction terms and confounding variables.

6 Acknowledgments

I want to acknowledge the paper "Academic phraseology; a key ingredient in successful L2 academic literacy" by Sylviane Granger for inspiring the nature of the research question in this experiment (Granger 2017).

I also want to thank Dr. David Brown for introducing me to the `stm` package as a viable way to conduct the analysis above and providing general advice about the direction of the broader project.

I want to acknowledge ChatGPT for helping me ideate topic names.

Works Cited

Baumard, Nicolas, Elise Huillery, Alexandre Hyafil, and Lou Safra. 2022. "The Cultural Evolution of Love in Literary History." Journal Article. *Nature Human Behaviour* 6 (4): 506–22. https://doi.org/https://doi.org/10.1038/s41562-022-01292-z.

Granger, Sylviane. 2017. "Academic Phraseology: A Key Ingredient in Successful L2 Academic Literacy." Journal Article. *Oslo Studies in Language* 9 (3): 1–20. https://doi.org/https://doi.org/10.5617/osla.5844.

Ishnoor. 2017. "Poetry Analysis with Machine Learning." https://www.kaggle.com/datasets/ishnoor/poetry-analysis-with-machine-learning?phase=FinishSSORegistration&returnUrl=%2Fdatasets%2Fishnoor%2Fpoetry-analysis-with-machine-learning%2Fversions%2F1%3Fresource%3Ddownload&SSORegistrationToken=CfDJ8CXYA35d3CRDujxBNSrCTMsLYG5jaKglRW2m
c08Yhkaslwt1HrU_rZJv2IUycUGY8VQIupJfh9CaKAZCVdx5k90P4f0tR8KS8tbZVhUSMdi3nIuT-u7Ppg4DIAlPg7X9DnAUKsjS6IyBfJEPW6PTqRIiW0eznhHHrGZX0dFyXEDJX4X4jv9OPd4rQijvqz0bfIC
9H6XaFKbOlR2sWyZC0ZOkYZJkElE4XjkP8Rc9JaBZfL_iaJP-K3uKusdREV1HmaHcDxwtjlz-z49x5y0eb84DrG2w&DisplayName=Andrew+Shih.

O'Rourke, Meghan. 2011. "A New Tally by Vida Shows How Few Female Writers Appear in Magazines." https://slate.com/human-interest/2011/02/a-new-tally-by-vida-shows-how-few-female-writers-appear-in-magazines.html.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "Stm: R Package for Structural Topic Models." Journal Article. *Journal of Statistical Software* 91 (2): 1–19. https://doi.org/10.18637/jss.v091.i02.

Venkitesh, Sreeram. 201AD. "Love Letters." https://www.kaggle.com/datasets/fillerink/love-letters.