International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-11
© Cerebration Science Publishing
https://doi.org/10.70917/ijcisim-2026-2363
Article

A Distant Reading Study of Marriage Discourse in Jane Austen's Novels Using Thematic Modeling

Le Jing [1] and Yali Yang [1,*]

[1] International Cultural and Educational College, Northeast Agricultural University, Harbin, Heilongjiang, 150030, China
* Correspondence author: ariel@neau.edu.cn

Abstract: Jane Austen's works primarily explore women's perspectives on marriage and express the pursuit of female independence, making them masterpieces of feminist literary criticism. This study uses textual data from the complete works of Jane Austen's six representative novels as its data source. It employs the TF-IDF algorithm for text feature analysis, utilizes semantic network graphs to visualize high-frequency words in the text, and applies the LDA topic model to identify four major themes of marital discourse in Jane Austen's novels. Finally, it analyzes the feminist themes from the perspective of emotional shifts across the chapters of the novels. The study found that the discourse on marriage in Jane Austen's novels can be categorized into four major themes: the emotional and rational dynamics of marriage; the economic and social class constraints of marriage; the essence of marital happiness and the pursuit of value; and the influence of family and social opinion on marriage. Austin's feminist thought, shaped by emotional structures, manifests both progressive and conservative aspects. The emotional values across the chapters of her novels are influenced by her feminist ideology—which embodies both progressiveness and limitations—and evolve through the interplay of positive and negative emotions.

Keywords: TF-IDF; LDA topic modeling; sentiment analysis; Jane Austen; discourse on marriage

1. Introduction

Jane Austen's six classic works—"Pride and Prejudice", "Sense and Sensibility", "Mansfield Park", "Emma", "Northanger Abbey", and "Persuasion"—focus on depicting the lives of the British middle class at the time, painting a vivid picture of their daily existence [1–4]. Among these, the theme of marriage runs throughout the works, possessing a rich complexity that extends far beyond the mere emotional level [5-6]. During Austen's lifetime, British society was in the midst of a critical transition from an agricultural to an industrial society. Feudal remnants remained deeply entrenched and difficult to dislodge, and the social hierarchy was marked by stark divisions. For women, marriage was not merely a personal emotional choice but a critically important decision closely tied to survival, social status, and the rise and fall of the family [7–9]. Traditional interpretations have simply categorized Austen's novels as romantic love stories; however, upon further analysis of the texts, one discovers that behind her views on marriage lie profound class metaphors, which simultaneously reflect the gradual awakening of female consciousness [10–12].

Thematic modeling is a text mining technique that aids in the understanding and analysis of large-scale textual data [13–14]. In the analysis of marriage discourse within Austen's novels, thematic modeling can automatically identify relevant themes within the text and classify or cluster texts based on thematic distributions. This approach helps identify the core themes of marriage discourse in Austen's works, thereby offering new perspectives for in-depth research and broader interpretations of Austen and her literature [15–17].

This paper takes the complete collection of Jane Austen's six representative novels as its research subject and performs preprocessing on the acquired data, primarily including Chinese word

Copyright: © 2026 by the authors

segmentation and stop-word removal. Subsequently, relevant theories such as TF-IDF, topic modeling, and perplexity are introduced. In the empirical study, text feature analysis is conducted using the TF-IDF algorithm, and the text data is visualized using a semantic network graph. The LDA topic model was used to identify four major themes in the discourse on marriage within Jane Austen's novels. Finally, by integrating feminism, sentiment analysis, and sentiment structure, this study explores the emotional shifts in the main chapters of *Pride and Prejudice* and analyzes the distinctively era-specific feminism present in the novel.

2. Data Sources and Research Methods

2.1. Data and Data Processing

2.1.1. Data Analysis Subjects

The data analyzed in this paper consists of the complete collection of six representative novels by the British realist writer Jane Austen. Following text extraction and cleaning, the dataset is expected to comprise approximately 14,000 lines of text and 13,900 words, covering the main narrative, character dialogue, and key narrative passages. Austin's literary career spanned the Regency period in Britain from the late 18th to the early 19th century. Her works adopt the "rural gentry" as their central perspective, using narratives of marriage to profoundly reflect the economic logic, gender norms, and moral dilemmas of the time.

2.1.2. Dictionary of Emotions

HowNet, provided by CNKI, and NTUSD from National Taiwan University are two of the most widely used sentiment dictionaries in the Chinese language domain. HowNet consists of two parts: evaluative terms and sentiment terms. The sentiment terms include 899 positive sentiment words and 1,536 negative sentiment words, while the evaluative terms include 3,920 positive evaluative terms and 3,688 negative evaluative terms. NTUSD contains 2,574 positive sentiment words and 8,698 negative sentiment words. However, sentiment polarity still requires manual modeling, and different modeling approaches yield different sentiment polarities. Additionally, given the characteristics of the original English texts in the Austin corpus, sentiment dictionaries adapted for the English domain—including ANEW and SentiWordNet—have been introduced.

2.1.3. Chinese Word Segmentation and Stop-Word Removal

Text processing requires word segmentation of text objects. In English, spaces serve as natural delimiters between words, whereas in Chinese, characters, sentences, and paragraphs can be easily demarcated by clear boundaries; however, it is particularly difficult to define word boundaries. Therefore, when processing Chinese text, word segmentation is necessary to represent sentences as individual words. This paper uses the Jieba Chinese word segmentation library in Python for this purpose.

Stop words are a category of words that have no practical meaning but are very common and indispensable in text; they are often articles, prepositions, adverbs, or conjunctions. In information processing, removing stop words helps save storage space and improve the efficiency of search and analysis. This paper uses the Harbin Institute of Technology (HIT) stop word list for stop word removal.

2.2. TF-IDF

TF-IDF is a common weighting technique used in text mining that evaluates the importance of a word in a document within a corpus based on its frequency within that document and its frequency across the entire corpus. In this context, TF stands for term frequency, and IDF stands for inverse document frequency.

In a document, term frequency (TF) refers to the number of times a word appears in the document. To prevent bias toward longer documents, normalization is required by dividing the term frequency by the total number of words in the document, for a given word t in a specific document d, $tf_{t,d}$ can be expressed as:

$$tf_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}} \tag{1}$$

In this context, $n_{t,d}$ represents the number of times the word t appears in document d, and $n_{k,d}$ represents the number of times the word k appears in document d, $\sum_k n_{k,d}$ represents the total number of occurrences of all words in the document d.

Inverse document frequency (IDF) measures the relative importance of a word. The frequency idf_t of a word t in a document d can be expressed as:

$$idf_t = \log \frac{|D|}{df_t + 1} \tag{2}$$

Here, df_t represents the number of documents in the corpus that contain the word t, and $|D|$ represents the total number of documents in the corpus. The denominator is increased by 1 to avoid a denominator of 0, which would occur if none of the documents in the corpus contained that word.

TF-IDF is the product of TF and IDF. The $tfidf_{d,t}$ for a given word t can be expressed as:

$$tfidf_{d,t} = tf_{d,t} \cdot idf_t \tag{3}$$

The core idea behind TF-IDF is that, within a corpus, the importance of a word is often proportional to its frequency of occurrence within a document, but inversely proportional to its frequency of occurrence across the entire corpus. If a word appears frequently in a particular document but rarely in others, we can conclude that it has strong discriminatory power—that is, it is more important than other words.

2.3. Thematic Model

A topic model is a type of clustering model that uses unsupervised learning to cluster the latent semantic structures of semi-structured and unstructured text, thereby generating abstract topics. Topic models are based on the assumption that text is composed of a mixture of multiple topics, and that abstract topics are composed of a mixture of different words. Therefore, a topic model is clearly a Bayesian network structure, divided top-down into three layers: the text layer, the topic layer, and the word layer.

2.3.1. LDA Thematic Model

LDA is a document topic generation model; it is a Bayesian probabilistic model consisting of three layers: words, topics, and documents. In this model, the relationships between documents and topics, and between topics and words, each follow a multinomial distribution. LDA is an unsupervised machine learning algorithm; it does not require manual intervention during standard topic training, but the number of topics to be extracted from the documents must be specified in order to train the document topic model. Its principle is based on the assumption that each document is a probability distribution composed of multiple topics, and each topic is a probability distribution composed of many words. Under this assumption, topics are generated from words, and text is then generated from topics. Therefore, the LDA topic model determines the feature words that compose a topic based on the magnitude of the topic probability distribution for each feature word, making the topics more representative of the text's characteristics. Compared to other topic models, LDA uses a Dirichlet distribution as prior knowledge for the topic distribution and infers meaning by analyzing the context of each word, resulting in superior topic extraction performance. Consequently, the topics generated by the LDA model better represent the semantic meaning of the text. After processing a document with LDA, the document is represented by a specific number of latent topics. The Bayesian composition model of LDA is shown in Figure 1.

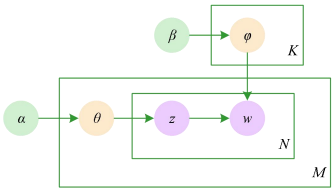

Figure 1. The Bayesian composition model of LDA

The symbols in the figure are defined as follows: K : Number of document topics. M : Total number of documents. N_m : Total number of words in the m th document. $\vec{\beta}$: Dirichlet prior parameter for the multinomial distribution of words under each topic. $\vec{\alpha}$: Dirichlet prior parameter for the multinomial distribution of topics under each document. $Z_{m,n}$: Topic of the n th word in the m th document. $W_{m,n}$: The n th word in the m th document. $\overrightarrow{\theta_m}$: The distribution of topics in the m th document. $\overrightarrow{\varphi_k}$: The distribution of words under the k th topic.

2.3.2. Parameter Estimation for the LDA Thematic Model

Solving the LDA topic model in full is highly complex, and currently common methods do not yield very precise parameters. The primary methods include Gibbs sampling and Bayesian inference. The commonly used Gibbs sampling algorithm is based on the Markov Chain Monte Carlo (MCMC) method. Monte Carlo is a random sampling method that combines the problem to be solved with a probabilistic model, then performs simulated sampling on a computer to obtain an approximate solution. The primary role of Gibbs sampling in the LDA model derivation process is to construct a Markov chain. This chain is formed by concatenating words from the text, with each word assigned to a specific topic. Using the Markov algorithm, random sampling is performed on each word in the chain to determine its topic, and this process is iterated until the topic probability of the word node on the chain converges. When iteratively determining the topic to which a word belongs in the chain, the topics of other words in the chain must first be kept constant, and the topic of the current word is inferred based on the topics corresponding to the other words. The computational expression is shown in (4):

$$p\left(z_i = k \middle| z_{-i}, w\right) = \frac{p\left(w, z\right)}{p\left(w, z_{-i}\right)} \propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{v=1}^{v}\left(n_{k,-i}^{(v)} + \beta_v\right)} * \frac{n_{m,-i}^{(k)} + \beta_t}{\sum_{z=1}^{k}\left(n_{m,-i}^{(z)} + \alpha_v\right)} \tag{4}$$

Regarding the specific meanings of the variables in the formula: z_i represents the topic variable for the i th word, and $-i$ represents the removal of i words. Let $W_i = t$ denote the i th word, which corresponds to the t th entry in the vocabulary list. $n_{k,-i}^{(t)}$ indicates that, in the k th topic, the i th word is excluded, and the number of words in the topic is equal to the t th word in the vocabulary. β_t denotes the value of the t th dimension of the hyperparameter β . $n_{m,-i}^{(k)}$ represents the number of words in the m th text that are excluded from the i th word and have the topic ID k . α_k represents the k -dimensional value of the hyperparameter α . Equation (4) is the algorithmic formula used in Gibbs sampling, which first computes the word-topic probability distribution and then the text-topic probability distribution. After determining the topic to which each word belongs through computation, the parameters of the topic-word probability distribution and the text-topic probability distribution can be approximated using Equations (5) and (6).

$$\varphi_{k,t} = \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{v=1}^{v}\left(n_{k,-i}^{(v)} + \beta_v\right)} \tag{5}$$

$$\theta_{m,k} = \frac{n_{m,-i}^{(k)} + \beta_t}{\sum_{z=1}^{k}\left(n_{m,-i}^{(z)} + \alpha_v\right)} \tag{6}$$

Here, $\varphi_{k,t}$ represents the probability of term t belonging to topic k . $\theta_{m,k}$ represents the probability of topic k appearing in text m . From this, it is clear that the process of constructing the LDA topic model primarily involves determining which topics each word in the text belongs to; once the relationships between words and topics are established, the parameters of the entire model can be estimated.

3. Empirical Research

3.1. Analysis of High-Frequency Word Features

High-frequency words in the text are key indicators of marriage discourse. The preprocessed file "Jane Austen Marriage Discourse Text.txt" was imported into the ROSTCM software for word segmentation and frequency analysis to identify the top 100 most frequent words in the text, thereby analyzing the comprehensive characteristics of marriage discourse in Jane Austen's novels. The top 100 most frequent words in the marriage discourse text are shown in Table 1. In terms of lexical attributes, the feature words in the table can be classified into three categories: nouns, adjectives, and verbs. Among them, nouns primarily include characters ("Elizabeth, Bennett," etc.), marriage resources ("dowry, annual income," etc.), cultural and social concepts ("prejudice, pride," etc.), and behaviors and settings ("walking, proposing," etc.). Adjectives primarily modify the emotional and value-related characteristics of marital relationships, falling into two categories: emotional perception and trait evaluation. Verbs mainly reflect the action logic of marital development, divided into two categories: marital interactions (such as "breaking off an engagement" and "choosing a spouse") and social actions (such as "guardianship"). Among the top 10 most frequent feature words, "Elizabeth" appears most frequently, serving as a typical character sample for the study of "marriage discourse." "Marriage" and "happiness" ranked second and tenth, respectively, corroborating Austen's exploration of the "essence of marriage"; the ultimate value of marriage discourse is happiness. Ranking third is 'Darcy'; as the central male protagonist, his character transformation and choice of marriage embody the conflict between emotional awakening and social discipline from a male perspective within "marriage discourse."

Table 1. A special term of the top 100 in the word frequency of marriage

Serial Number	Vocabulary	Frequency	Serial Number	Vocabulary	Frequency	Serial Number	Vocabulary	Frequency
1	Elizabeth	352	35	Mrs. Gardiner	156	69	Dignity	108
2	Marriage	298	36	William Price	155	70	Respect	107
3	Darcy	276	37	The Fenway family	154	71	Understanding	106
4	Bennett	264	38	Fanny Price	152	72	Dowry	134
5	Love	258	39	Sir Thomas	150	73	Manor	132
6	Prejudice	248	40	Knightley	148	74	Annual income	130
7	Arrogance	242	41	Bath Town	146	75	Social Season	128
8	Rational	238	42	Miss Crawford	144	76	Engagement	126
9	Emotion	232	43	Sir Eliot	142	77	Church	124
10	Happiness	226	44	Miss Smith	140	78	Carriage	122
11	Eleanor	218	45	Harriet	138	79	Dance party	120
12	Marian	212	46	Wickham	136	80	Take a walk	118
13	Edward Ferras	206	47	Bentley	166	81	Take a walk	116
14	Mr. Darcy	200	48	Colonel Brandon	164	82	Female	114
15	Emma	196	49	Mr. Woodhouse	162	83	Widow	112
16	Mr. Knightley	192	50	John Knightley	160	84	Primogeniture e	110
17	Captain Wentworth	188	51	Mr. Claude Ford	158	85	Dowry	108
18	Annie	184	52	Louisa Masgrave	156	86	Family status	106
19	Wentworth	180	53	Charles Hite	154	87	Marriage alliance	104
20	Fanny Price	176	54	Old Masgrave	152	88	Separation	102
21	Mrs. Bertram	172	55	Mrs. Russell	150	89	Refuse to marry	100
22	Sir Bertram	170	56	Mrs. Clay	148	90	Elopement	98
23	Edmund Bertram	168	57	Captain Smith	146	91	Break off the engagement	96
24	Mary Claude-ford	166	58	The Palmers	144	92	Guardianship	94
25	Henry Claude Ford	165	59	Mrs. Middleton	142	93	Harmful production	132
26	Catherine Moran	162	60	Colonel Brandon's sister	140	94	Annuity	130
27	General Tinier	160	61	Mrs. Allen	138	95	Parish	128
28	Mr. Alan	158	62	Mrs. Hargreaves	136	96	The priest's residence	126
29	Norsangjue Monastery	156	63	Mrs. Ball	134	97	Lady	124
30	Mansfield Park	154	64	Mrs. Jennings	132	98	Gentleman	122
31	Hatfield	152	65	Mr. Palmer	130	99	Marriageable age	120
32	Highbury Village	150	66	Mrs. Ferras	128	100	Choosing a spouse	118
33	Bath Hot Spring	148	67	Mr. Gardiner	126			
34	Pemberley Estate	157	68	Mr. Collins	110			

The 100 characteristic words in the table were classified according to the established categories of cognitive imagery analysis. The number of characteristic words in each main category, as well as the proportion of their cumulative word frequency relative to the total word frequency, were calculated. The distribution of the number of characteristic words and the proportion of cumulative word frequency across the main categories is shown in Figure 2. As shown in the figure, the "Emotion and Morality" category has the highest number of feature words (27), with a cumulative frequency proportion of 29.21%, ranking first. This indicates that Austen's discourse on marriage places significant emphasis on emotional rationality and moral judgment within marriage; core terms include "love," "pride," and "prejudice." The number of feature words in the "Society and Institutions" and "Culture and Aesthetics" dimensions is similar, at around 19, but there is a significant disparity in their cumulative frequency shares. The cumulative frequency of "Society and Institutions" accounts for 11.83% of the top 100 feature words. Upon examining specific terms, core words include "marriage," "engagement," and "refusal of marriage." Compared to other dimensions, the "Behavior and Scenes" and "Characters and Relationships" dimensions show weaker representation, with only 8 terms each, accounting for 5.28% and 5.09% of the cumulative word frequency, respectively.

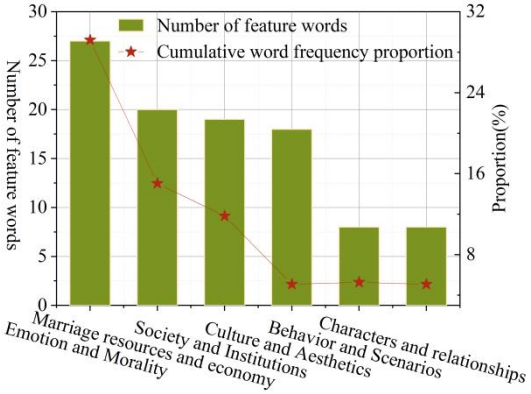

Figure 2. The number of feature words in the main category and the proportion

3.2. Semantic Network Analysis

Semantic network analysis can visually deconstruct the relational patterns between words as well as the semantic pathways linking syntax and concepts within the text, thereby further identifying the associations and meanings between words. By constructing a semantic network diagram of the vocabulary using ROSTCM's Netdraw software, the analysis revealed that the semantic network exhibits a "core-periphery-edge" structure. The semantic network, constructed based on high-frequency feature words, is shown in Figure 3. The core layer comprises "person-related" and "theme-related" categories, serving as the "conceptual hub" of marital discourse. Co-occurring terms in this layer exhibit broad coverage and close relationships, primarily reflecting how family power structures constrain marital choices and the need to balance emotional impulses with rational judgment in marriage. The peripheral layer serves as an extension of the core layer and primarily includes three categories of terms: "economic resources," "social actions," and "institutional norms." The peripheral layer is primarily composed of feature terms from two aspects: "Scenario" and "Emotion." Through weak associations, it enhances the realism of the narrative and enriches the contextual and emotional details of marital discourse.

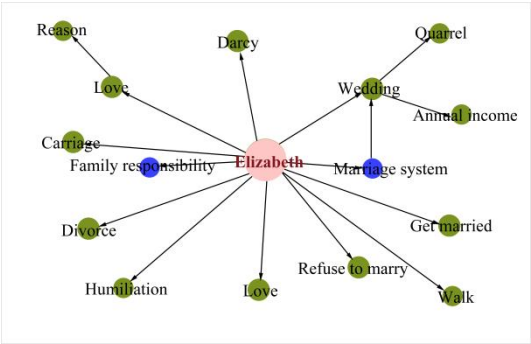

Figure 3. Build a semantic network based on high-frequency feature words

3.3. A Thematic Analysis of Discourses on Marriage in Jane Austen's Novels

3.3.1. Confusion Matrix Calculation

Using Python software, we calculated the perplexity of the LDA model for different numbers of topics. The calculation of LDA topic perplexity is shown in Figure 4. Generally, the perplexity value decreases as the number of latent topics increases; the smaller the perplexity value, the stronger the generative capability of the topic model. Local minima of the perplexity are distributed between 4 and 5 topics. LDAvis is a core technology topic extraction method based on an improved LDA topic model. It presents core technology topics through a dynamic, interactive visual graph, where different noun clusters represent distinct topics. The size of each noun cluster is proportional to the interpretability of the topic. By comparing the sizes of these clusters and their interconnections, the core technology topics of a document can be identified. By using Py-LDAvis to visualize the number of topics with low perplexity, the document's core technical topics can be identified from the visual graph.

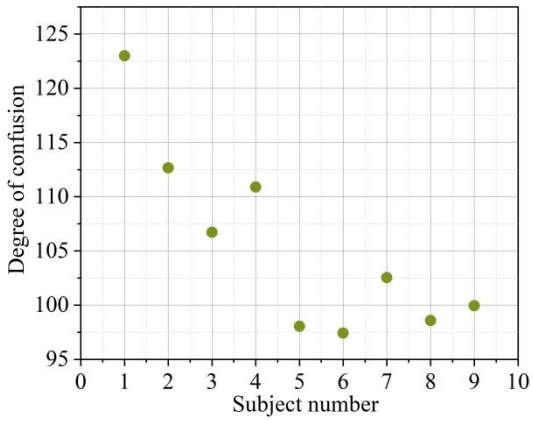

Figure 4. LDA topic confusion

When the number of LDA topics is set to 4, the resulting LDAvis visualization is shown in Figure 5. All four topics consist of distinct clusters of noun phrases, and these clusters are relatively dispersed, indicating that all four topics exhibit strong clustering capabilities. All four topics represent core technical themes within the documents.

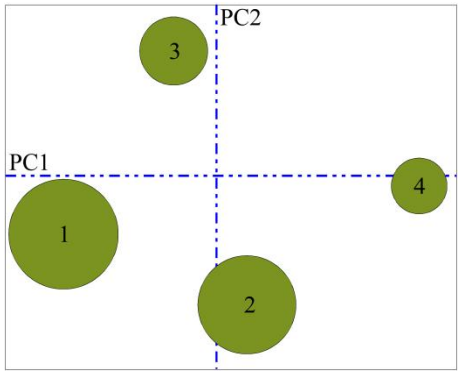

Figure 5. Visualizing the LDAvis map for female

3.3.2. Topic Extraction

Using Python, we identified the terms with the highest fit for each topic; we refer to these as "high-frequency topic terms." These terms can be used to further explain and summarize the various topics. The topic terms and their weights are shown in Table 2. Analyzing the high-frequency words and their weights listed in the table, the high-frequency words for Theme 1 are "marriage," "love," "reason," "Darcy," "Elizabeth," "prejudice," "estate," "proposal," "rejection," and "reconciliation." The high-frequency words for Theme 2 are "property," "dowry," "status," "sister," "younger sister," "older brother," "sister's marriage," "family expectations," and "social class." The high-frequency terms for Theme 3 are "happiness," "dignity," "choice," "autonomy," "the cost of marriage," "the test of love," and "the triumph of reason"; those for Theme 4 are "mother," "nagging," "father's expectations,"

"sibling rivalry," "family pressure," "social scrutiny," and "marital choices." By synthesizing these high-frequency terms, the following themes are identified:

Theme 1: The Emotional and Rational Dynamics of Marriage, which focuses on the conflict between "emotional conflict" and "rational considerations" in marriage choices, with the interactions of the central characters serving as the narrative backbone.

Theme 2: Economic and Social Class Constraints on Marriage, which reveals the logic of "economic foundations" and "social class" underlying marriage choices, with the marital dilemmas of secondary characters serving as typical examples. This specifically includes the Bennett sisters' motivations for marriage, the implicit constraints of Darcy's "aristocratic status" on marriage, and societal norms regarding "marriages of equal standing."

Theme 3: The Essence of Happiness and the Pursuit of Value in Marriage, which explores the core meaning of a "happy marriage" through a comparison of the characters' marital outcomes.

Theme 4: The Influence of Family and Public Opinion on Marriage, analyzing the dual pressures on marriage choices from within the family and from society at large. This specifically includes Mrs. Bennett's eager desire for her daughters to marry into wealth, the competition among the sisters regarding marriage, and societal restrictions on "women's autonomy in marriage."

In summary, the four major themes influencing the discourse on marriage in Jane Austen's novels are: the interplay of emotion and reason in marriage; the economic and social class constraints on marriage; the nature of marital happiness and the pursuit of value; and the influence of family and social opinion on marriage.

Table 2. Key words and weights

Theme 1		**Theme 2**		**Theme 3**		**Theme 4**	
Feature Words	Weight	Feature Words	Weight	Feature Words	Weight	Feature Words	Weight
Marriage	0.27	Property	0.259	Happiness	0.114	Mother	0.103
Love	0.108	Dowry	0.108	Dignity	0.108	Nagging	0.09
Rational	0.071	Status	0.058	Choose	0.098	Father's expectations	0.08
Darcy	0.05	Sister	0.053	Autonomy	0.097	The competition among sisters	0.073
Elizabeth	0.046	Sister	0.052	The price of marriage	0.091	Family pressure	0.05
Prejudice	0.04	Elder brother	0.041	The test of love	0.09	The perspective of society	0.044
Manor	0.028	My sister's marriage	0.036	The victory of reason	0.068	The choice of marriage	0.041
Proposal	0.027	Family Expectations	0.036				
Refuse	0.024	Social class	0.029				
Reconciliation	0.02						

3.4. An Emotional Analysis of Discourses on Marriage in Jane Austen's Novels

3.4.1. The Characteristics and Limitations of Austin's Feminist Thought

Figure 6 illustrates the comparison of positive and negative emotions across the chapters of the novel. As shown in the figure, the solid line representing positive emotions generally lies above the dashed line representing negative emotions, reflecting the novel's overall positive emotional tone. The significant variation in the scatter plot reflects the novel's dramatic plot twists and the protagonist's fluctuating emotional experiences.

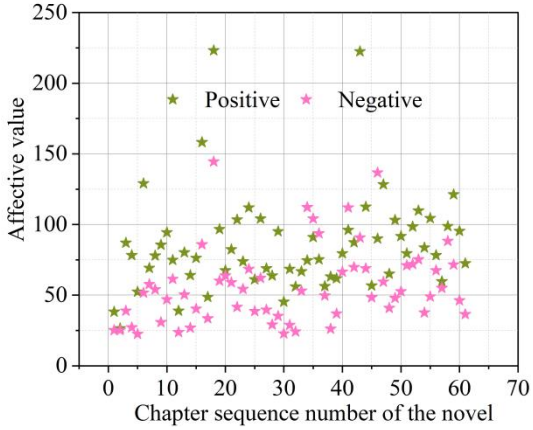

Figure 6. The contrast of positive and negative emotions in each chapter of the novel

The emotional shifts across chapters are shown in Figure 7. Chapters with higher overall sentiment scores include Chapter 6, Chapter 16, Chapter 18, and Chapter 43, all of which have scores greater than 70. Chapters with lower overall sentiment scores include Chapter 34, Chapter 36, and Chapter 46, all of which have scores lower than -12. A high overall sentiment score for a chapter does not mean that the chapter is entirely free of negative emotions. For example, in Chapter 18, the positive sentiment score reached 223, while the negative sentiment score was as high as 144, suggesting the protagonist's complex emotional conflicts within the chapter. Similarly, chapters with low overall sentiment scores do not imply the absence of positive emotions. For instance, in Chapter 46, while the negative sentiment score reached 137, the positive sentiment score was 90. This suggests that although the female protagonist's mood plummeted in this chapter, there was still room for a turnaround—a hypothesis that requires further analysis to verify. Chapter 6 marks the novel's first peak of positive emotion, with a value of 129. In this chapter, Austen satirizes the prevailing marriage ideology of the time—which prioritized wealth over the emotional compatibility of the couple—and emphasizes that a woman's wisdom can counteract a man's arrogance and earn respect. This serves as the cornerstone of equality between the protagonists and lays the foundation for the blossoming of their romance.

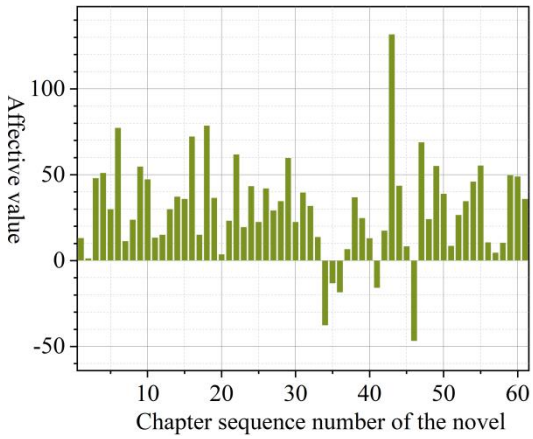

Figure 7. Emotional changes in each chapter

The protagonist's negative emotions are closely linked to ingrained notions of social standing and class. In Chapter 18, while the positive emotion score is quite high (223), the negative emotion score is also significant (144). At the ball, Mrs. Bennett's exaggerated remarks—such as her belief that she would become the mistress of Mr. Bingley's house through marriage—embarrassed Elizabeth and led others to view the Bennet family as of low character. Elizabeth was also mocked by Miss Bingley and Darcy as a result, which is a major reason for the high negative sentiment score in this chapter. The reason Elizabeth cares so deeply about others' opinions of her family is that she has not yet been able to shake off the influence of social class distinctions. Even a woman with such a strong personality as Elizabeth feels inferior because of her humble origins, demonstrating that Austen's feminist ideas had not entirely escaped the class-based notions of her time. As the plot progresses, the negative sentiment scores in Chapters 34 through 36 remain consistently high, at 112, 104, and 93, respectively. Through

the depictions in these three chapters, the author reveals Elizabeth's vulnerability: her repeated bouts of low spirits are linked to her fear that Darcy might look down on her family background. Yet women should not place undue importance on men's opinions of them, nor should they belittle or devalue themselves.

3.4.2. Feminism, Affect Analysis, and Affective Structures

This paper examines Jane Austen's "Pride and Prejudice" by integrating feminism, affect analysis, and affective structure. It finds that chapters with high positive affect values exhibit relatively clear manifestations of progressive feminist ideas, whereas chapters with high negative affect values contain relatively more manifestations of the limitations of feminist thought. Furthermore, chapters with higher positive emotional values consistently demonstrate a tendency to challenge male-dominated notions and express views on gender equality, whereas chapters with higher negative emotional values generally depict Elizabeth's feelings of inferiority and vulnerability within a patriarchal society. It is worth noting that the heroine's unconventional behavior and her distinctive, independent personality traits reveal Austen's tendency to criticize patriarchal traditions. Chapters featuring these elements consistently exhibit high positive sentiment scores. This suggests that the ability to transcend mainstream culture while remaining grounded in it determines the richness of feminist ideas in each chapter, and the depth of these ideas is one of the factors influencing the positive sentiment scores.

In summary, while Austen opposes the gender disparities stemming from class differences, she simultaneously presents marriage as the sole solution to bridging class conflicts. Essentially, she remains unable to escape the traditional notion that women exist in dependence on men. This coexistence of progressiveness and limitations precisely reflects the complex emotional structure of the social groups of that era.

4. Conclusion

This article examines the complete collection of Jane Austen's six representative novels, employing TF-IDF algorithms, word clouds, semantic network association analysis, and LDA topic modeling to conduct quantitative textual analysis and explore the feminist themes within the novels. The study reached the following conclusions:

(1) The top 100 most frequent words in the marriage discourse can be categorized into three types: nouns, adjectives, and verbs. Based on the distribution of the number of words in each main category and their cumulative frequency share, the categories of "emotion" and "morality" have the highest number of words (27) and account for 29.21% of the cumulative frequency. This indicates that Austen's marriage discourse places significant emphasis on emotional rationality and moral judgment within marriage.

(2) The four major themes of the discourse on marriage in Jane Austen's novels are: the interplay between emotion and reason in marriage; the economic and social constraints of marriage; the nature of marital happiness and the pursuit of value; and the influence of family and public opinion on marriage.

(3) An analysis of feminism in Jane Austen's novels reveals that her feminist thought, influenced by emotional structures, manifests in both progressive and conservative aspects. The emotional tone of each chapter is shaped by the author's feminist ideology—which combines both progressive and limited perspectives—and fluctuates amid the interplay of positive and negative emotions. Constrained by the era, Austen's feminist thought could not escape the traditional view of women within the institution of marriage; she continued to regard marriage as a woman's ultimate destination.

References

1. Mammadova, I. (2025). Class and Character in Emma: Jane Austen's Subtle Critique of Social Hierarchies. Journal of Azerbaijan Language and Education Studies, 2(3), 125-139.
2. Sadeq, A. E. (2017). The Presentation of the Gentleman in Jane Austen's Novel Emma (1816). Asian Social Science, 13(7), 181.
3. Mochar, A. (2025). Social Change, Morality, and Dress in the Novels of Jane Austen and France Burney. Wiener Digitale Revue, (6).
4. Suyunova, M. (2026). THE REPRESENTATION OF LOVE AND SOCIAL CLASS IN THE NOVELS OF JANE AUSTEN: A LITERARY ANALYSIS. BG pulse journal, 2(5), 1617-1622.

5. Bailey, M. (2015). The Marriage Law of Jane Austen's World. Persuasions: The Jane Austen Journal On-Line, 36(1).
6. Akman, F. B. (2018). AN INVESTIGATION OF SOCIO-ECONOMIC INCENTIVES AND IMPLICATIONS OF MATRIMONY ON WOMEN'S LIVES IN JANE AUSTEN'S NOVELS. Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi, 58(1), 1101-1127.
7. Mohammed, A. A. (2016). Marriage in Jane Austen's Pride and Prejudice. International Journal of Media Culture and Literature, 2(4), 59-73.
8. Muir, R. (2024). Love and marriage in the age of Jane Austen. Yale University Press.
9. Maeda, Y. (2023). The Marriage Plot in Jane Austen. Reading, 43, 19-46.
10. Harper, B. S. (2020). The Influence of Jane Austen's Works on Societal Attitudes Regarding Women and Marriage, Education, and Slavery from the Early Nineteenth to Twentieth Centuries. California State Universit.
11. Sundari, S. G. (2025). The Theme of Love and Marriage in Jane Austen's Novels: Pride and Prejudice and Sense and Sensibility. Language In India, 15(1), 107-117.
12. Yavuz, M. E. (2019). JANE AUSTEN'S ROMANTIC WORLD AS REFLECTED IN HER NOVELS. Euromentor Journal-Studies about education, 10(4), 122-127.
13. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. Foundations and Trends® in Information Retrieval, 11(2-3), 143-296.
14. De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. Scientometrics, 103(2), 413-433.
15. Li, D., Wu, K., & Lei, V. L. (2024). Applying topic modeling to literary analysis: A review. Digital Studies in Language and Literature, 1(1-2), 113-141.
16. Chu, K. E., Keikhosrokiani, P., & Asl, M. P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. Pertanika Journal of Science & Technology, 30(4), 2535-2561.
17. Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. Frontiers in artificial intelligence, 3, 42.