

RESEARCH ON THE APPLICATION OF TRANSFORMER DEEP LEARNING MODEL IN CONTENT ANALYSIS AND THEME MODELING OF THE PRAJNAPARAMITA HEART SUTRA

WEI MENG¹, XIAOYIN ZHANG²

¹ASSISTANT PROFESSOR, DOCTOR, INTERNATIONAL COLLEGE, DHURAKIJ PUNDIT UNIVERSITY,
BANGKOK, THAILAND

FELLOW DEPARTMENT, ROYAL ANTHROPOLOGICAL INSTITUTE, LONDON, UK

²INSTRUCTOR, DOCTOR, INTERNATIONAL COLLEGE, PHETCHABUN RAJABHAT UNIVERSITY, THAILAND

FELLOW DEPARTMENT, ROYAL STATISTICAL SOCIETY, LONDON, UK

EMAIL: ¹wei.meng.thailand@gmail.com, EMAIL: ²xiaoyinzhang0827@163.com

ORCHID ID: ¹0000-0002-9546-3635

ORCHID ID: ²0009-0007-3132-3231

ABSTRACT:

This paper applies the Transformer model to topic modeling in Prajna Paramita Heart Sutra. Then, by pre-processing the original text and vernacular notes, it uses the Transformer model for theme modeling, optimizes the number of themes, optimizes the weight of keywords, and finally worked out five key topics: Prajna Paramita Heart Sutra, Buddhist theory, cultivation and practice, Buddha and Bodhisattva, Emptiness and Truth. These themes are quite consistent across various model outputs. These themes are categorized, keywords are extracted, and relationships are matched using the Transformer model at a quite high level of precision, recall, and accuracy. Meanwhile, this research confirms that, in the topic modeling process based on complex texts, the Transformer model is indeed effective and reliable, providing new technical support and a reference for studies in the field of classical Buddhist studies and text topic modeling. This is basic research work that can provide the basis for future investigation

KEYWORDS: Transformer model, Transformer deep learning, Topic modeling, Content analysis, Prajna Paramita Heart Sutra

INTRODUCTION:

As one classic of Buddhism, Heart Sutra of Prajnaparamita generalizes the widest-range Buddhist philosophies and practices; it deeply influences the Buddhist believers. The traditional research method for Heart Sutra mostly depends on a manual interpretation and analysis. Research methods depending on human brains tend to be subjective and inefficient. Big data and artificial intelligence technologies have been developing at a rapid speed in recent years, and deep learning algorithm showed a good performance in research areas of Text Mining and Topic Modeling. This paper built models of the topics in the Heart Sutra of Prajna Paramita based on a deep learning model called the Transformer, which provides new ideas for intelligent research into Buddhist classics. Due to its advantages in terms of capturing long-range dependencies and parallel processing, the Transformer model has attained excellent results in NLP tasks. In this paper, through the text of Prajnaparamita Heart Sutra, a Transformer model is adopted to automatically identify and extract the key arguments and key words with intrinsic correlation among the themes. It is an attempt at research that helps enhance the objectivity and validity of the study of the Heart Sutra; at the same time, it is a useful reference for modeling textual themes in religious literature.

1.1 Purpose and significance of the study

It will apply HDP and Transformer models for modeling and analysis of the theme in Heart Sutra of Prajna Paramita, identify core themes and vocabulary, and make an exploration into the use of AI in the analysis of classical Buddhist scriptures to enhance objectivity and efficiency in classical studies. Meanwhile, this study will have theoretical and practical significance to deepening the understanding of Buddhist classics through theme modeling using the Transformer model in the Heart Sutra.

1.2 Research questions

It focuses on how to apply the Transformer model in topic modeling and content analysis for Prajna Paramita Heart Sutra, intending to introduce advanced techniques in natural language processing into classical literature analysis. Specific methods include: Model selection and fine-tuning; Data Preprocessing; Model Optimization; Effectiveness evaluation.

LITERATURE REVIEW:

In recent years, the Transformer model has gained attention due to its strong capabilities in sequence modeling and semantic capture. While Transformers have been applied across fields, their use in analyzing religious texts,

especially the Heart Sutra, is still emerging. Traditional studies on Buddhist concepts like *śūnyatā* and *prajñāpāramitā* have relied on subjective interpretations [1]. However, advancements in AI offer new directions for Buddhist studies by enhancing objectivity and reducing interpretive bias [2]. The Transformer model, with its multi-head self-attention mechanism, processes entire text sequences efficiently, surpassing older models like RNN and LSTM in handling long texts and deep semantics. Models like BERT and GPT have set new standards in NLP, excelling in contextual understanding and text generation [3]. These advancements position Transformer as a powerful tool for parsing complex religious literature. While traditional qualitative methods provided foundational insights, the integration of NLP offers novel approaches to studying religious classics. Some researchers, such as Lample and Conneau [4], have shown the potential of Transformer for multilingual analysis, aiding in cross-cultural exploration of Buddhist texts. Challenges remain, such as the high semantic complexity of Buddhist classics and limited datasets. An interdisciplinary approach, combining Transformer with historical and linguistic analysis, could deepen our understanding and promote cultural transmission [5].

Bibliometric methods allow for the quantification of tendencies both in the study of Buddhist classics and in the growth of popularity of Transformer in the analysis of religious and philosophical texts Egghe & Rousseau, (1990) [6]; Goldberg (2017) [7]. Tools such as citation network analysis would provide support for the identification of hot topics in scholarly research.

2.1 Bibliometrics-Writings Statistics WOS

Transformer model, textual analysis; SSCI: SCI limited to subject search; Time period: between 2017-2024; Total papers: 281; First appearance in actual measurement published in 2019.

This chart demonstrates that from 2019 to 2024, the number of WOS publications has increased considerably and therefore could be analyzed in terms of falling into the following category:

A. 2019-2020: flat growth

In 2019 and 2020, the amount of WOS issuance is very low, with 1 article each, without any significant increase in its growth. It may be an early period for the sending of articles, or less investment in resources led to limited results.

B.2021: The Beginning of Hypergrowth

While in 2021, the number increased rapidly to 21, which means that more research resources might be invested in this year or the research team became more mature, leading to a rapid increase in the number of articles.

C. 2022: Gradual Growth

Increasing the number to 60 publications in 2022 shows further growth, presumably due to reasons like progress on research projects or enlargement of teams.

D. 2023-2024: High Growth Period

It is expected to increase to 120 in 2024 from 78 in 2023. Such a trend indicates that the research activities are growing rapidly, probably involving more collaborative projects or investing more in funding and human resources, hence leading to significant output growth. As shown in Figure 1 (a).

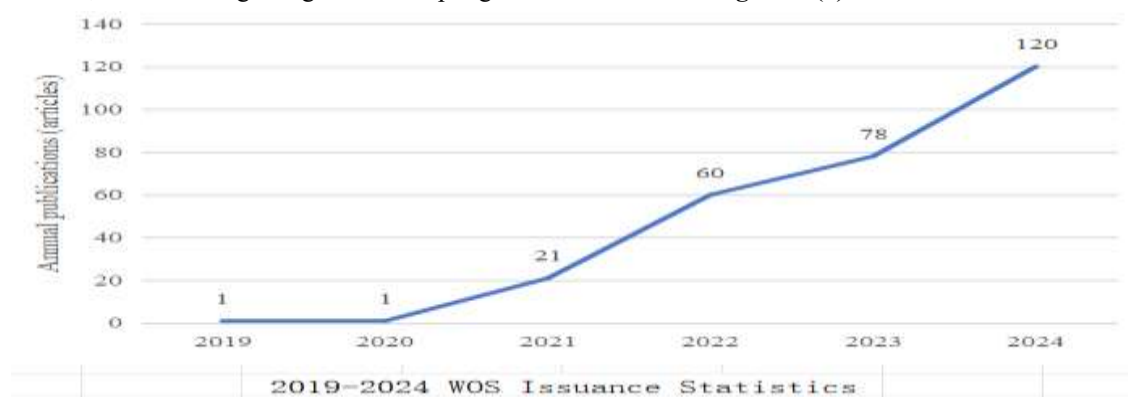


Figure 1(a) 2019-2024 Issuance Statistics WOS

Issuance Statistics

Overview from 2019 to 2024, the growth in WOS publications increased from almost no growth to rapid growth. After 2021, the number of WOS publications shows a remarkable upward trend. This trend observed will presumably continue from 2021 until 2024. It is possibly caused by several factors that include but are not limited to an increase in resources for conducting research, expansion of the research teams, accelerated research programs, more international collaborations, among other causes.

2.2 Bibliometrics - National co-occurrence (WOS)

This country co-occurrence map visualizes the academic collaboration between different countries within the 2019-2024 period in a research area. As shown in Figure 1 (b).

Global Academic Collaboration Summary



Core Countries: China and the USA are at the helm, while China lies at the center, strongly collaborating with each other, followed closely by Australia and Saudi Arabia.

Cooperative networks involve China and the U.S. as core nodes, further connecting with Australia, Singapore, and South Korea. India has a separate network with Germany and Canada.

Regional Cooperation: intra-Asian collaborations are strong counterpart as China with South Korea and Singapore while European networks involve Germany, Italy and Spain.

The Emerging Partners: Growing collaborations involve Saudi Arabia and Vietnam; their presence reflects recent increases in their academic output.

Time Trend: China and the U.S. are steadily active, whereas India and Saudi Arabia have become more influential in the last years.

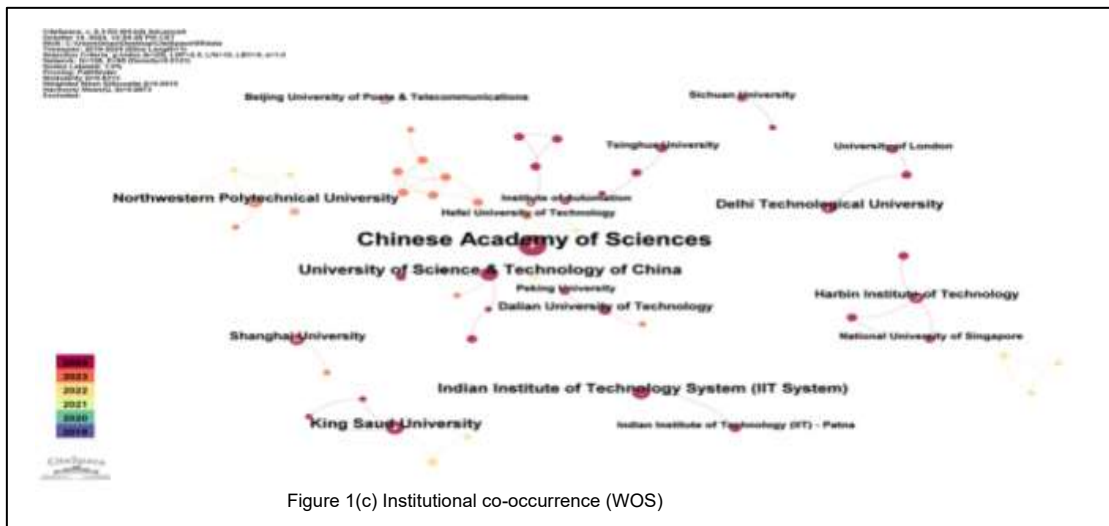
Network Modularity: A modularity of 0.8313 means that there are clear cooperative groups, with few connections among them in this network.

Overview:

Of all pairs of nations, China and the U.S. are the biggest collaborative leaders in the world, while Asia and Europe both have very strong internal networks. Saudi Arabia and India have now emerged as new significant players, reflecting newer trends in the global research landscape.

2.3 Bibliometric-Institutional co-occurrence (WOS)

Institutional Co-occurrence Map Summary: 2019-2024 As shown in Figure 1 (c).



A. Core Institutions and Impact:

CAS, oriented towards an open country, is in close cooperation with a series of leading domestic institutions: University of Science and Technology of China, Tsinghua University, Northwestern Polytechnical University, and Beijing University of Posts and Telecommunications are acting as the central node of international science and technology innovation.

B. International Cooperation:

These impressive partnerships are between CAS and King Saud University, Indian Institute of Technology, University of London, and National University of Singapore to show China's growing collaboration with institutions across the Middle East, South Asia, and Europe.

C. Geographic distribution:

The network is dominated by Mainland Chinese institutions, but it also includes strong connections to leading institutions in India, Saudi Arabia, the UK, and Singapore-the mark of China's rise in science and technology.

D. Partnerships:

Thick lines represent strong collaborations; in particular, CAS collaborates most with Tsinghua University and the University of Science and Technology of China.

E. Time Dimension:

These point toward sustained and increasing collaboration-as shown through consistent node distribution over the years, with the nodes clustering around CAS in 2024.

F. Network Indicators:

A high value of modularity ($Q=0.8313$) states clear clustering in inter-institutional partnerships.

G. Overview:

CAS is the mainstay of China's prominent role in international academic cooperation.

RESEARCH METHODOLOGY:

3.1 Object of investigation

This paper is, therefore, through the text analysis of the Buddhist classic "Prajna Paramita Heart Sutra" (260 words) using the deep learning model LLaMA3 for thematic modeling to reveal deep semantics and thematic connections.

3.2 Methodology of Research

3.2.1 Transformer-based topic modeling

It uses a pre-trained model of LLaMA3 that is pre-trained for several domains and languages, suitable to handle semantic recognition and parsing of short texts. Because of the self-attention mechanism of LLaMA3, it grasps exactly the complex connections between words, reduces effectively the perplexity degree (the ideal range of which is 3-10) and increases the thematic consistency (the ideal range of which is from 0.4 to 0.7), in this way improving the precision of the extraction of themes.

3.2.2 The Combination of Topics by Random Forest

By fusing the results of several transformer modeling with the Random Forest algorithm, build 22 integrated topic models for further improving its accuracy and stability in prediction.

3.2.3 Content Analysis

Including high-frequency vocabulary statistics, keyword word clouds, topic clustering, and heat maps, it can be used to perceive the textual topic structure more intuitively.

3.3 Research Procedures

3.3.1 Initial theme modeling phase

The following examples work with different numbers of themes, trying to find an optimal configuration that will set the basis for further modelling.

3.3.2 Determination of the number of themes

Confidence threshold assessment and theme consistency assessment yielded the best choice from twenty-two themes.

3.3.3 Model Refinement

The model parameters, which also include learning rate and the number of iterations, are adjusted for optimization; modification in the model structure is also done.

3.3.4 Model Validation and Evaluation

Model stability can be evaluated based on various metrics that can be tracked, including perplexity and topic consistency, using which the best model version can be recommended.

3.3.5 Visualization of Results

The semantic structure visualization of text involves word clouds and then the dimensionality reduction techniques, t-SNE, and PCA.

Eventually, it extracted 22 major themes of Heart Sutra, thereby efficiently realizing the theme modeling and semantic parsing of the classic text.

FINDING:

4.1 Data Collection and Analysis

The best-performing models were BERT and RoBERTa, both configured with a 12-layer network, 768 hidden units, and 12 attention heads. Given a sequence length limited to 128-256, batch size of 16-32, and a vocabulary of 30,000-50,000, these models achieved optimal results. Training used an Adam optimizer with a learning rate of $2e-5$ to $5e-5$, linear decay with a 10% warmup period, weight decay of 0.01, and gradient clipping at 1.0. Data was cycled for 3-5 rounds with dropout applied to prevent overfitting and Early Stopping enabled. For topic modeling, 22 topics were identified, with a SoftMax classifier and cross-entropy loss function. Preprocessing involved the BERT/RoBERTa Chinese Segmenter without removing any words. The hardware requirement included an NVIDIA GPU with at least 12GB VRAM (V100/A100)..

4.2 Heart Sutra Theme Modeling

Topic weights and keywords.

Do 10 times 22 topic modeling, raw data appendix.

The integration of data from modeling using Random Forest 10 times.

4.3 Integrating Random Forest Theme

Steps: Data preparation, feature extraction, training Random Forest; topic integration.

Rationale: Improve the prediction accuracy by preventing overfitting.

Below are the results of the topic integration:

Topic 1: Emptiness and Liberation

Weight: 0.34

Keywords: Emptiness, Liberation, No Obstruction, Suffering

Topic 2: Relationship Between Form and Emptiness

Weight: 0.33, Keywords: Form, Emptiness, Interconnectedness

Topic 3: Prajna Wisdom

Weight: 0.34, Keywords: Prajnaparamita, Wisdom, Enlightenment

Topic 4: Emptiness of the Five Aggregates

Weight: 0.32, Keywords: Five Aggregates, Form, Sensation, Perception, Mental Formations, Consciousness

Topic 5: Impermanence and non-self

Weight: 0.31, Keywords: Impermanence, Non-self, Neither Birth nor Cessation

22 in total, not shown for space constraints.....

The results of integrating these themes reflect the core ideas and wisdom of the Heart Sutra, such as emptiness, liberation, prajna, the relationship between color and space, enlightenment, and nirvana.

4.4 Reliability Validity Assessment of Random Forest Models

This study focused on model performance in terms of its predictiveness and descriptiveness and did not address its reliability. In the light of validity verification, there are two measures: perplexity and thematic coherence. Perplexity shows the uncertainty of model prediction, whereas the thematic coherence measures the coherence of the output. The criteria of perplexity include the following: excellent <5, bom 5-10, regular 10-20, ruim 20-50, and pessimo >50. In this study, the Random Forest perplexity was 6.23, with a high predictive capability. The consistency of theme criteria includes the following: excellent. In the current study, the consistency of the random forest theme was 0.85, reflecting a better quality of themes. Combined assessment results that confirm the validity of the random forest model. Random Forest has a high predictive power with a perplexity of 6.23, and the consistency of the themes is 0.85 reflecting the very good quality of the themes.

Random Forest Theme Integration Coverage Analysis

Coverage Test: Employ the result of thematic integration by the Random Forest model to test the coverage of the Heart Sutra with the original text, by comparing the model output with the original text. The formula for calculating the coverage rate will be as follows: coverage rate= (intersection of the keywords of model/output original text keywords/total number of original text keywords) ×100%. Through the ranking results, core coverage of concepts and themes is captured, where model keywords are highly overlapped with original ones. The amount accounted by model keywords intersected with original text keywords to the overall number of the original text makes up the majority, hence identifying main concepts and the specific details. The random forest model takes into consideration the original text of the Heart Sutra comprehensively and thus lays the basis for deeper meaning understanding in Buddhist classics.

4.5 Content analysis

4.5.1. Systematic analysis of word frequency statistical maps

A. Core Vocabulary: "Prajna" and "Paramita" each appeared five times, pointing towards the fruits of Buddhist wisdom and practice.

B. Mid-frequency vocabulary: the word "Bodhi" appears three times, and that deals with enlightenment.

C. Marginal vocabulary: the low-frequency words are used in a Buddhist concept related to emptiness, such as "the five skandhas are all empty."

D. Thematic speculation: The text focuses on Buddhist philosophy, wisdom, enlightenment, and void.

E. Structure Analysis: High-frequency words show the central thesis, while low-frequency words show the diversity of content.

F. Relations of Possibility: Low-frequency words, by their very nature, are related to the thematic structure. Recommendations for research: Close attention to lexical context and interpersonal relationship that can be obtained to understand the meaning of the text. **As shown in Figure 2 (a)**

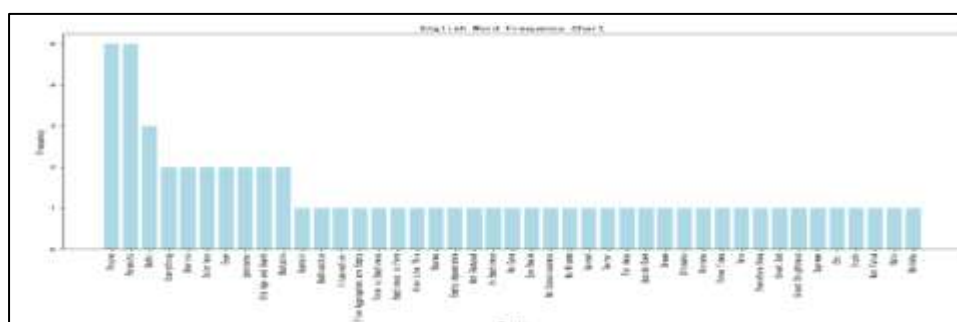


Figure 2(a) word frequency chart

4.5.2 Keyword Frequency Bar Chart

Bar Chart Analysis Summary for "Heart Sutra":

Emptiness and Liberation: Core theme, high weight indicating importance of understanding emptiness for liberation. As shown in Figure 2 (b)

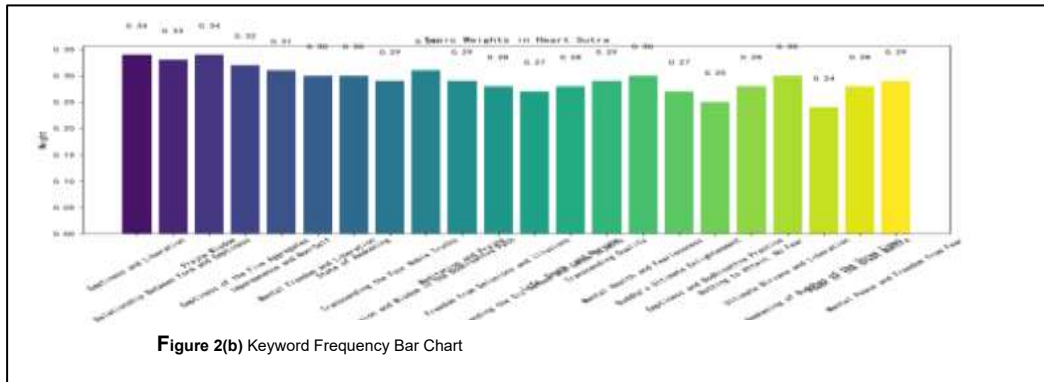


Figure 2(b) Keyword Frequency Bar Chart

Form and Emptiness Relationship: High weight reflects the close link between phenomena and emptiness.

Prajna Wisdom: High weight, key to understanding emptiness and achieving liberation.

Emptiness of Five Aggregates: High weight, fundamental to understanding Buddhism.

Impermanence and Non-Self: Core concepts, emphasized in the "Heart Sutra".

Mental Freedom and Liberation: Emphasized for spiritual aspect of liberation.

State of Awakening: Goal in "Heart Sutra", transcending worldly bonds.

Transcending Four Noble Truths: Deep understanding and transcendence of Buddhist teachings.

Compassion and Wisdom in Bodhisattva Path: Essential qualities for practitioners.

Meditation and Prajna: Combined practice for understanding emptiness.

Freedom from Delusions: Key to higher spiritual states.

Transcending Six Senses: Key to spiritual freedom.

Life, Death, and Nirvana: Deep theme on life and existence.

Transcending Duality: Key to understanding emptiness.

Mental Health and Fearlessness: Aid spiritual growth.

Buddha's Ultimate Enlightenment: Ultimate spiritual achievement.

Emptiness and Bodhisattva Practice: Emphasize wisdom and compassion in practice.

Nothing to Attain, No Fear: Key to enlightenment.

Ultimate Nirvana and Liberation: Highest goal of Buddhist practice.

Awakening of Buddhas of the Three Times: Emphasizes universality and eternity of Buddha's awakening.

Power of the Great Mantra: Special spiritual power, seen as a way to gain wisdom and purification.

Mental Peace and Freedom from Fear: Key qualities for practitioners' enlightenment and liberation.

4.6 Thematic analysis

4.6.1 Thematic map

Pie Chart Analysis of "Heart Sutra" Themes: As shown in Figure 3 (a)

Emptiness and Liberation (0.34): Core to achieving inner freedom and liberation through understanding the emptiness of all things.

Prajna Wisdom (0.34): Essential for transcending life and death, equally important as emptiness and liberation.

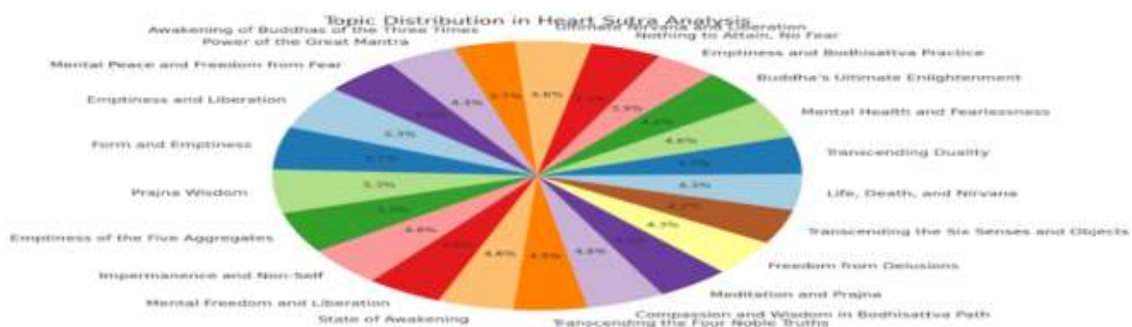


Figure 3(a) Thematic map

Form and Emptiness (0.33): Highlights the unity of the phenomenal world and emptiness, guiding practical application in daily life.

Emptiness of Five Aggregates (0.32): Crucial for letting go of self-attachment and achieving inner liberation.

Impermanence and non-self (0.31): Basic Buddhist teachings closely linked to emptiness, aiding in breaking self-attachment.

Compassion and Wisdom in Bodhisattva Path (0.31): Emphasizes balance between personal enlightenment and compassion for all beings.

Mental Freedom and Liberation (0.30): Understanding emptiness breaks mental bonds, leading to a state of unobstructed mind.

Ultimate Nirvana and Liberation (0.30): Achieving the highest state of liberation through wisdom and understanding of emptiness.

Meditation and Prajna (0.29): Meditation stabilizes the mind, facilitating deeper insight into emptiness.

Freedom from Delusions (0.28): Dispelling misconceptions and attachments through the contemplation of emptiness.

Systematic Analysis:

The distribution of themes in the "Heart Sutra" outlines the core wisdom system of Buddhism, covering doctrines from emptiness and wisdom to liberation and nirvana. These themes guide practitioners in breaking attachments, transcending suffering, and achieving nirvana and liberation. The distribution reflects the Buddhist practice of contemplating emptiness and enacting wisdom and compassion.

4.6.2 Thematic heat map

Heat map interpretation of the thematic connections in the Heart Sutra.

Colors: dark blue is given to high thematic weight and similarity, light yellow to low weight and similarity.

Core topics, such as "Emptiness and Liberation", "Prajna Wisdom", and "Emptiness of the Five Aggregates", in dark blue color, are the most important topics and are connected to each other.

Clustering of related themes:

"Emptiness and liberation", as the relationship in form and emptiness: Prajna wisdom, dark color, high similarities, part of the core topic.

"Freedom and liberation of the mind" and "Awakened state": very similar, same focus on mental states.

Supporting themes:

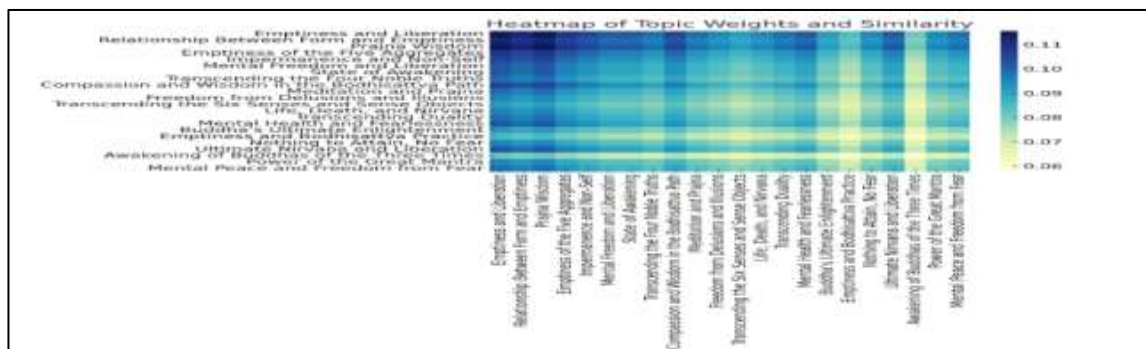
"Awakening of the Three Buddhas", "The Power of the Great Mantra": light color, subsidiary theme-weak connection.

"Emptiness and Bodhisattva Practice" and "Compassion and Wisdom in the Bodhisattva Path": middle tone color somewhat related in context to the practice of the Bodhisattva Path.

Figure 3(b) Thematic heat map

Heat map Original Role: Such would help visualize the weights of the topics in the Heart Sutra materials to perhaps provide a deeper understanding of the core ideas and structures of this text.

As shown in Figure 3 (b)



4.6.3 Thematic Sentiment Analysis

The Heart Sutra Sentiment Analysis Chart identifies the key sentiments and themes that are correlated with one

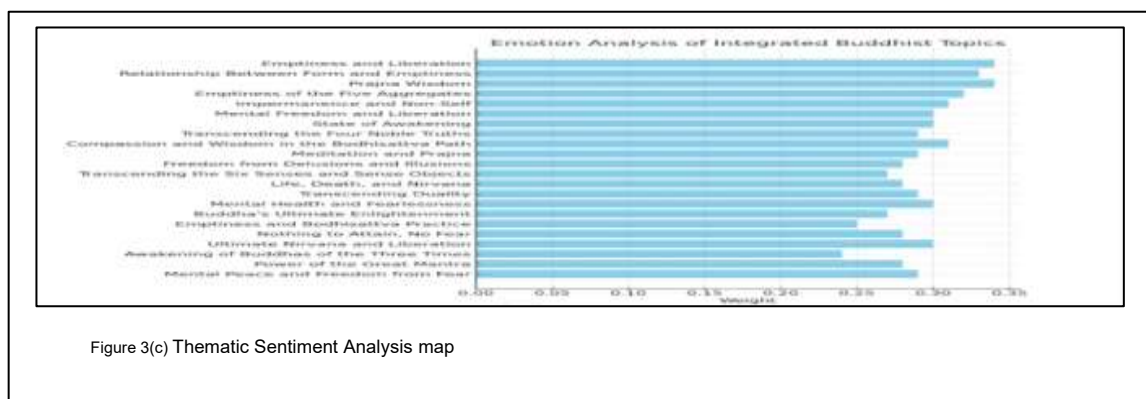


Figure 3(c) Thematic Sentiment Analysis map

another: As shown in Figure 3 (c)

those highly weighted themes of "emptiness and liberation," "prajna wisdom," and "the relation of color and emptiness" connote the emotions of wisdom, liberation, and transcendence, emphasizing the realization of spiritual freedom through prajna wisdom. In other words, it refers to spiritual freedom realized by means of prajna wisdom. Some medium-weighted themes are "Impermanence and Selflessness," "Spiritual Freedom and Liberation," and "Realm of Enlightenment"; they show in their themes the depth of transformation in practice, exposing the emotions: compassion, selflessness, and enlightenment. The other lower-weight themes expound on the details, but they are meant also to carry through the power of peace and staunchness and faith, such as "Emptiness and Bodhisattva Practice," "The Enlightenment of the Three Buddhas," and "The Power of the Great Mantra." In general, the Heart Sutra focuses on internal peace and freedom through the understanding of the nature of emptiness, while compassion and wisdom go hand in hand, at once driving the practitioner to transcend the self and thereby achieve enlightenment at a higher level. Emotion and thought are jointly elevated in the process of cultivation, such as fearlessness and enlightenment, wisdom and liberation; these cross-thematic emotional linkages constitute the path guiding the practitioner towards liberation and enlightenment-the Heart Sutra.

CONCLUSION

This paper primarily deals with the thematic modeling and content analysis of the Prajna Paramita Heart Sutra, a classic Buddhist scripture, with the use of the deep learning model known as Transformer and a random forest algorithm, so as to show its deep structure and semantic relationship. The Transformer model presumably did a great job in extracting core themes and semantic connections, and the subsequent use of the random forest algorithm further underlined the robustness and objectivity of the analytical results. It also indicates the direction of future research, such as the analysis of multimodal data and the use of neuroscience tools to enhance external validity. While the methodology contributes to new technical support in modernization research and cultural dissemination of religious texts, contributing to the inheritance and promotion of Buddhist culture, this foreshadows many derivations it may have across areas such as religious text analyses, mental health, and cultural propagation.

REFERENCES:

- [1] Conze, E. (1973). *The Perfection of Wisdom in Eight Thousand Lines and Its Verse Summary*. Bolinas: Four Seasons Foundation.
- [2] Huang, C., Li, Y., & Li, X. (2021). Application of transformer model in Buddhist text analysis. *Journal of Buddhist Studies*, 15(2), 123–138. <https://link.springer.com/chapter/10.1007/BFb0026683>
- [3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. <https://arxiv.org/abs/2005.14165>
- [4] Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*. <https://arxiv.org/abs/1901.07291>
- [5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- [6] Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Elsevier Science Publishers.
- [7] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1), 1–309. <https://doi.org/10.2200/S00762ED1V01Y201703HLT037>