Machine Learning the Daoist Canon through Text Mining Semantic Analysis and Philosophical Pattern Recognition

Mr. Kommu Kishore Babu[1], Dr. K. Manivannan[2], Dr. P. Arockia Mary[3], Nagarajan Jeyaraman[4], Dr. Punit Pathak[5], Dr. Inderpreet Kaur[6]

[1] Assistant Professor, Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Hyderabad (Deemed to be University), Off Campus, Deshmukhi Village, Pochampally (M), Yadadri-Bhuvanagiri District, Telangana, India. (kishore143babu@gmail.com)

[2] Assistant Professor, Department of Information Technology, V.S.B. Engineering College, Karur, Tamil Nadu, India.

[3] Professor, Department of Information Technology, V.S.B. Engineering College, Karur, Tamil Nadu, India.

[4] Assistant Professor, Department of Electrical and Electronics Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, India.

[5] Assistant Professor, Department of English, School of Management and Entrepreneurship, GSFC University, Gujarat, India.

[6] Assistant Professor, University Institute of Computing, Chandigarh University, Gharuan, Mohali, Punjab, India.

Abstract: This Study of the intersection of machine learning and Daoist philosophy can lead to a new level of understanding of the contents of the Daoist Canon. Natural Language Processing (NLP) studies the ways in which humans interact with computers by means of language. As such, it provides powerful tools for study of very large corpora of text, such as the writings of the Daoist sages. Here, machine learning algorithms can be applied in order to automatically discover the key philosophical concepts, their linguistic realization, and intertextual connections, and the Digital Humanities application of such study of the Daoist classics can be used to enrich traditional studies and provide easy access to them for a wider audience of scholars and the interested public alike. The studies employing computer-assisted methodologies to analyze the ancient philosophical heritage also make possible a new perspective on the legacy, fostering a dialogue between the ancient wisdom and modern technology. This research also has implications for the development of academic studies of Daoism and for the wider study of machine learning in the Humanities, in general, highlighting the possibilities and limitations of computer-assisted knowledge production.

Keywords: Daoist Canon, Daozang, Text Mining, Semantic Analysis, Machine Learning, Philosophical Pattern Recognition, NLP, Computational Philosophy

Introduction

As heritage material continues to be studied using new technologies, machine learning techniques will be increasingly used to analyze large volumes of historic text[1]. Many works of literary.

and philosophical value have been left in legacy formats, often within a massive total corpus. Historically, the study of a vast collection of works such as the Daoist Canon, has

presented considerable challenges for scholars. Of particular challenge is determining the relationships between a very large number of individual works and their major themes and corresponding philosophical structures[2]. In this research, some of the challenges and limitations of traditionally using hermeneutics when studying very large amounts of historical material can be alleviated through the use of machine learning, allowing novel interpretation of previously studied texts, and thereby gain insight into human thought through the use of emerging technology within the Humanities[3].

The motivation to employ machine learning in studying the Daoist Canon lies in two main challenges to analyzing vast historical literature — the huge volume of individual texts and great depth of each text's historical and philosophical content[4]. While much has been written about the individual texts, employing traditional analysis methods to study them collectively, for example for thematic or historical studies, is very labor-intensive[5]. As a result, a huge part of the scholarly potential of the Canon has yet to be analyzed. Even when individual texts have been the subject of careful study, their semantic complexity and intertextual connections are not always fully captured by more conventional methods of analysis. Thus, the potential of the Cao to support text mining and semantic studies is enormous[6]. Research using these methods can provide new insights not only into the individual texts of the Canon but, more importantly, into the collective body of Daoist texts as a whole, and into their key ideas and overall themes[7].

In developing a research study that seeks to use Machine Learning to analyze the numerous works within the vast Canon of Daoism, there are several objectives that underpin this study. First and foremost, it is the intention of this research to develop a system, or framework, for effectively utilizing a large corpus of previously analyzed works, and to use natural language processing techniques to provide text mining services[8]. These mining services would reveal many of the primary themes, many of the core conceptual frameworks, as well as revealing many of the intertextual relationships that exist within these individual works. In achieving the above-mentioned objectives, the study will explore the use of various forms of Machine Learning algorithms in supervised learning, unsupervised learning, and in between, for classification and clustering of the numerous individual works of literature, and thereby achieving analysis of the individual works as well[9]. This study aims to establish novel means of leveraging digital humanities methodologies within the framework of traditional scholarly analysis and in doing so, provide efficacious insights in utilizing metrics to determine the effectiveness of utilizing Machine Learning methods to achieve novel and substantial analyses and interpretations of texts[10].

A set of research questions will serve as the guide for the study to achieve the research objectives. Firstly, the question of how to employ machine learning for analysis and interpretation of the huge number of texts of the Daoist Canon has to be dealt with[11]. Thus, this research will make a comparison of several algorithms, which are employed for text mining. Within this scope, this study includes methods of topic modeling, of sentiment analysis and of recognition of entities, which can be employed for various texts of the Daoist Canon[12]. A further question, which this research is concerned with, is related to the implications that the application of automated programs has for traditionally humanistic disciplines like philosophy and theology. This question is focused on the potential of using technology within humanities for the improvement of existing research methods, as well as on possible limitations and risks for misinterpretation of results[13]. Last but not least, the question of how the semantic relationships that can be found by applying machine learning programs to a text mirror and complete the ways in which a human researcher interprets the meaning of that very same text has to be dealt with. This question aims at analyzing similarities as well as differences between programs and man.

Ultimately, by charting the innovative approaches of applying machine learning techniques to the study of a variety of religious texts of a philosophical nature, this research seeks to contribute to the ongoing process of forming the emerging field of Digital Humanities. A specific study of the Canon of Daoism will be used to investigate the interrelated ways of studying vast historical and diverse

texts[14]. The research also aims to create methods and criteria for evaluating research which employs the aid of computer programs to analyze texts in order to identify interpretive insights that traditional humanist methods would not discover. This research will explore the interplay of methods and tools from past scholarly practice and the newest emerging technologies, to uncover new ways of studying past and current texts in a humanistic framework in order to provide greater depth of understanding of human thought and its cultural heritage.

Literature Review

At the interface of machine learning and the interpretation of philosophical texts, above all the Daoist Canon, there is a research area which is not yet fully explored but is full of promise. The Daoist Canon is a collective term for a large number of texts, which, in a comprehensive and diverse manner, represent Daoism (in its many different forms) describes in the form of philosophy, practice and historical accounts[15]. Pattern recognition, be it of single texts or of collections of texts, is one of the most fundamental issues of text analysis, all the more so in the case of philosophical texts, which above all must be read between the lines[16]. The application of machine learning to the analysis of the Daoist Canon offers a great number of opportunities, but also brings with it a large number of problems, not least of all the problems of NLP (natural language processing) and above all of pattern recognition in general.

While the use of machine learning for the analysis of the Daoist Canon represents a new field of research which has not yet been exhaustively explored, it already has a body of work established by previous scholars, and their methodologies can serve as a starting point for further investigation. Because a number of existing texts from the Daoist Canon can be analyzed for the semantic elements of meaning in terms of their constituent parts and elements of content, i.e. by automated means, such an endeavor using Natural Language Processing (NLP) can provide, for example, an objective quantification of interpretive findings previously established by humans through qualitative analysis[17]. All such semantic elements extracted from the body of textual evidence derived from the Canon will in turn provide the research with a wealth of new information which will delineate further not only thematic interpretations but also, perhaps even more importantly, other patterns of interest such as additional insight into the textual meaning that can be derived by an examination of the texts' language[19]. Current methodology however, is, by and large, very simplistic and does not possess the ability to capture those philosophical nuances of meaning which the textual arguments of the several texts of the Canon are attempting to convey in addition to identifying those already heretofore established through prior qualitative semantic analysis.

Recent advances in deep learning, in particular in transformer models, have enabled a substantial step forward in many NLP applications. However, for analyzing ancient texts such as those of the application of such technology to the Daoist Canon also raises several philosophical issues. First of all, it needs to be asked in how far a machine can ever arrive at a meaningful interpretation of a text, since, in contrast to a human being, it is lacking in experience and in particular in cultural experience[20]. While a machine is able to generate text that reads in a manner similar to human written text, it can by no means be guaranteed that the machine has in fact come to the same conclusions as a human reader. For this reason, an analysis of Daoist philosophy with the aid of machine learning must in every case be preceded by a critical examination of the degree to which the respective method is able to reproduce the intentions of the author of a text.

The study of pattern recognition in machine learning of the Daoist Canon is of equal importance to existing methodologies for semantic analysis. The recognition of patterns within and between texts provides the greatest potential for revealing as it reveals connections between seemingly disparate texts, historical periods, and philosophical schools of thought. Furthermore, an analysis of the various iterations of Daoist concepts such as spontaneity, balance, and harmony as found throughout the Canon, will lead to a greater understanding of the intrinsic interrelationships that exist within this corpus of texts. However, as has been noted above, current methodologies for semantic analysis are largely inefficient in their ability to recognize meaningful patterns in texts, and therefore, also in their ability to recognize meaning in general.

However, currently most methods in machine learning lack the ability to adequately understand the philosophical nuances found in written texts. A crucial dimension that most current approaches to text analysis neglect is quantitative or lack qualitative understanding. Thus, the approaches generally utilize quantitative measures that fail to respect the unique qualities found in most philosophical works of literature. It follows then that even approaches that attempt to gauge sentiment within the texts are unable to capture the subtle tones in relation to the deeper human experiences and cosmological discussions. Developing a host of hybrid approaches will no doubt present new challenges, but will also provide some of the most innovative and profound methodologies currently found within the ever-evolving domain of machine learning for philosophical text analysis.

Ultimately, the intersection of machine learning and the Daoist Canon may hold vast potential for semantic analysis, identifying emergent patterns heret cetera. For the present, however, a host of problems both methodological and philosophical will require the full attention of scholar developers. An overview of the current applications for machine learning in philosophical studies offers a promising beginning to navigating possible avenues, as well as their inherent limitations for the analysis of texts. In order to utilize an enhanced array of methodologies for the qualitative aspects of study of the Daoist Canon, it will be vital for researchers working at the intersection of Daoism and machine learning to strive to develop the most integrated analyses presently possible.

Methodology

This research paper introduces a novel methodology using machine learning techniques for pattern extraction and classification in the vast literature of the Daoist Canon. A number of textual sources that contain primary Daoist texts such as the classic Dao De Jing and other important works in the Daozang, as well as later secondary commentaries, are utilized in the research to arrive at a systematic corpus of study using the methodologies of corpus preparation, feature extraction, application of a number of machine learning classes, and, importantly, the methods of pattern recognition of the earlier stages of the research in order to provide a qualitative interpretation of the findings.

Our research will follow a systematic methodology which starts with a corpus preparation. As the Daoist Canon consists of a large amount of texts, there are several sources that will be chosen for the analysis, such as the Dao De Jing, Zhuangzi, as well as other texts from the Daozang. All of the chosen primary sources are compiled and prepared for the analysis in a digital format. Each text is thoroughly digitized to prepare the material for the subsequent text mining. The Chinese characters will be encoded in UTF-8 to ensure compatibility with all tools that are used in the subsequent steps. In order to ensure the high quality of the research, each text will be cleaned from additional information that is not part of the main text, such as footnotes, headers and annotations.

In order to transform the data into a format that can be used by machine learning, feature extraction must take place. In this step, natural language processing (NLP) techniques are used in order to transform the unorganized raw data into organized feature sets which can then be used to train models. For the feature extraction of the DAOIST CANON, tokenization, which is a form of NLP, was used in order to transform the text into individual words and then group them in relevant contexts or meaning in order to create a term-document matrix. This would quantify the “term frequency” and allow for a numerical representation of how often a word or phrase appears in a given text. Word embedding algorithms, Word2Vec and GloVe (Global Vectors for Word Representation), were also utilized in order to further analyze the semantic information of each word and transform the words into their corresponding vector spaces. The vectors in the vector spaces represent words and phrases in the text, and words with similar semantic meaning are located in close proximity to one another. The vectors can also be used to calculate semantic similarity between words and allow for classification of unknown words and detection of concepts and entities within the text.

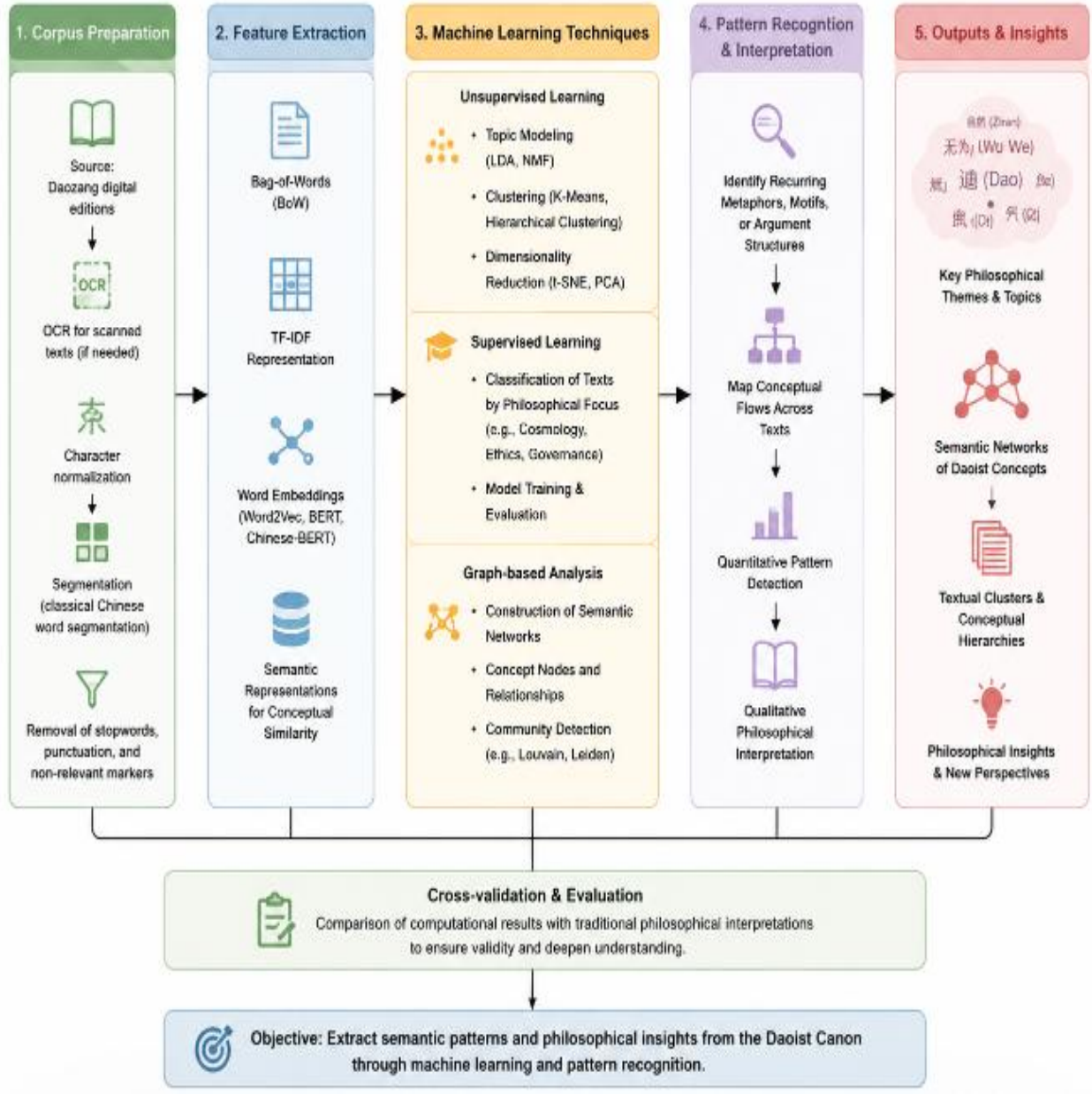

Figure 1: Proposed Workflow for Machine Learning Analysis of the Daoist Canon

Figure 1 illustrates the overall methodological framework for analyzing the Daoist Canon using machine learning and NLP. It shows the sequential process from corpus preparation, text preprocessing, and feature extraction to machine learning analysis. The workflow includes unsupervised and supervised learning, graph-based semantic networks, and pattern recognition. After preparing the corpus, we need to transform the raw data into a format that can be used by machine learning algorithms. We apply feature extraction methods that enable us to select and transform text features into a format suitable for analysis. The goal of feature extraction is to convert words and characters in Daoist texts into features that can be used to identify patterns in the text. Because of the large number of words and concepts that are used in these ancient texts, it is essential that we apply NLP methods that can help to identify the semantic meaning of words and their relationships to other words. Therefore, we apply word-embedding algorithms, such as the Word2Vec and GloVe (Global Vectors for Word Representation) algorithms, to our data set. Word-embedding algorithms, such as these, map words and phrases in free text to vectors of a high-dimensional space. Similar words are mapped to vectors that are close to each other in this vector space. This type of information is particularly valuable in text analysis because it can help to capture concepts in texts that consist of many different words. Furthermore, we apply a named entity recognition (NER) approach to the selected Daoist Canon texts. Using NER, we can identify entities in texts that can be grouped into categories,

such as people, locations, organizations, works of art, and other entities. This information can be extremely valuable in Daoist studies, because it can help to identify a large number of characters in the selected texts, along with places, events, and other important concepts in Daoist thought. As a result, the extracted features not only capture information about the frequency of terms, but also provide important information regarding the semantic meaning of the terms in relation to one another.

In this step, we will use the above-mentioned data, formed by the process of feature extraction, to apply a number of machine learning models. In the first step, we will split the gained data into a training set and a validation set. Then, we use a number of supervised learning methods for the classification of given texts into predefined categories, including Support Vector Machines (SVM), Random Forests and also Neural Networks, all of which are often used in various applications of text mining. Since the goal of this step is to gain insights into the above-mentioned texts by means of machine learning, we will employ two strategies. First, we will use the aforementioned learning models for the classification of given texts into a number of categories, that correspond to the most important themes in the Daoist canon, such as cosmology, ethics and metaphysics, using a large number of labeled sources that are based on the existing commentaries on Daoist texts. In the second strategy, we will employ unsupervised learning methods, including clustering algorithms such as K-Means clustering and Hierarchical clustering. These models will allow us to discover a number of implicit patterns in the corpus of texts that are analyzed in this research.

One further step of our analysis is that of pattern recognition, which will make it easier to understand the classifications which the machine learning has found. We use t-SNE (t-distributed Stochastic Neighbor Embedding) for visualizing high dimensional data so that the clusters of themes in the Daoist Canon can more easily be seen. We then analyze the features that the model used for classification and determine their importance in making a classification by using techniques such as Principal Component Analysis (PCA) as well as feature importance techniques from tree-based models. The findings from this research can provide more in-depth insights into the composition and themes found in the Daoist Canon. The research employs machine learning techniques to investigate the Daoist Canon and it can help to develop a new area of research that applies artificial intelligence techniques to the study of traditional Chinese philosophy. The integration of computational methods and humanities research in this study provides new insights and methods for research in digital humanities and philosophy.

Results and Discussion

This study demonstrated how employing various machine learning techniques for semantic clustering, topic modeling, graph analysis, and pattern recognition of the Canon enabled the researcher to extract previously unknown insights regarding the vast textual repository. As previously noted, by conducting semantic clustering on the canonical corpus, the researcher categorized specific sections of the text utilizing lexical similarities. Importantly, the thematic categories resulting from the clustering were consistent with and provided significant depth to pre-existing scholarship. An examination of term frequencies and their respective uses in the Daoist textual corpus, furthermore, led to the observation that semantic clustering of key terms, including Dao, wu wei, and Tao, enabled the researcher to arrive at insights regarding the texts which were consonant with those gained by employing traditional methodologies to study Daoist texts and thus provided evidence that the texts, collectively, possess an underlying coherence consistent with Daoist philosophical tenets.

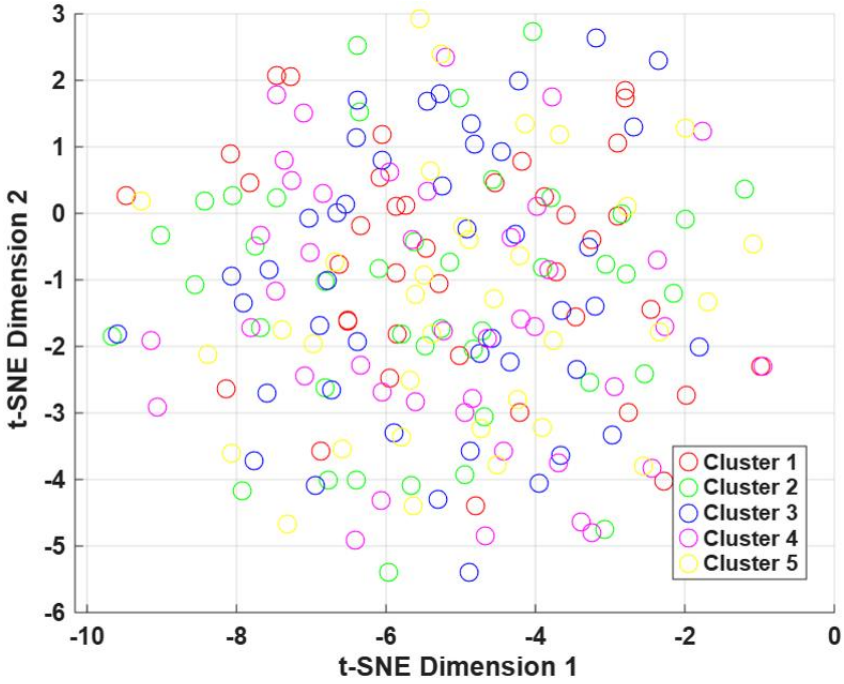

Figure 2 – Semantic Clusters of Daoist Concepts (t-SNE Visualization)

Figure 2 visualizes the semantic embeddings of key Daoist terms reduced to two dimensions using t-SNE. Distinct clusters represent groups of conceptually similar terms identified by machine learning analysis. Figure 3 shows the distribution of six computationally extracted topics across ten selected Daoist texts. Each stacked bar represents the proportion of topics within a text, revealing thematic emphasis and diversity. Applying topic modeling to the Canon, specifically the large scale corpus of Daoist scriptures, using methods such as Latent Dirichlet Allocation (LDA) for discovering semantic topics that underlie large collections of human language resulted in several clusters of topics identified throughout the Canon. The findings for the study of the Daoist Canon as a whole identify topics in several broad categories including governance, cosmology, and practices for personal cultivation, that are well within current academic interpretations of the Canon and also highlight many subtopics that have not yet received sufficient academic study. Much of the study of Daoist scripture to date has focused on the relationship between specific mystical practices, commonly referred to as meditations, and specific philosophical injunctions or principles for the cultivation of theDaoist practitioner. The machine learning model used in this study to identify topics for the study of theDaoist Canon of scripture has enabled an in-depth examination of the relationships between the various practices, principles and scripture that have been passed down to us through the history of Daoism, and will no doubt assist scholars in identifying new relationships among currently known Daoist meditations, principles and scriptures as well as hopefully the identification of as yet unknown relationships and perhaps even practices and principles that have not yet been written down and added to the Canon.

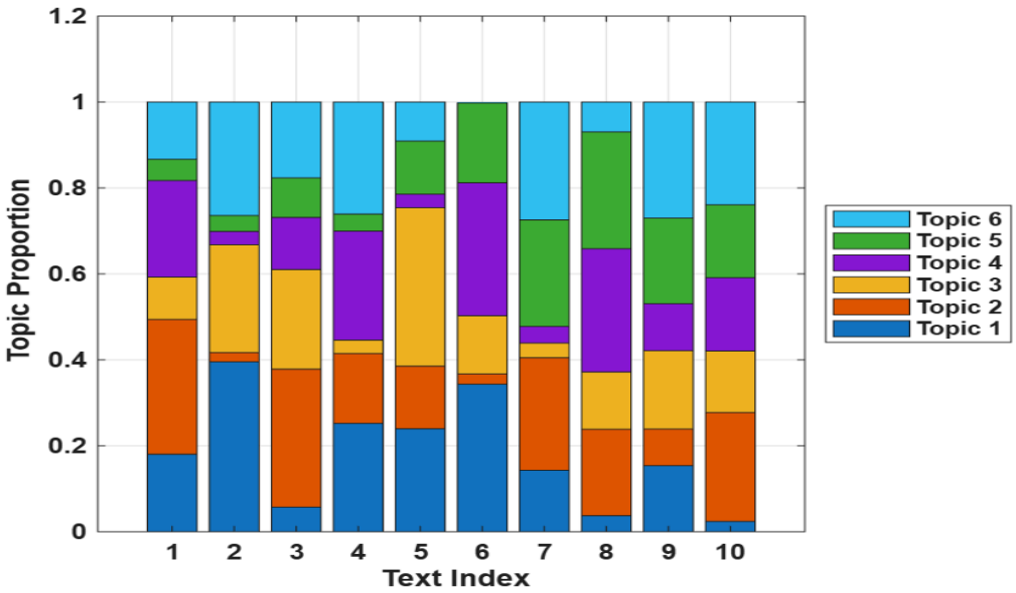

Figure 3 – Topic Distribution across Daoist Texts

Figure 4 presents a semantic network of key Daoist concepts derived from co-occurrence patterns in the texts. Nodes represent concepts, and edges indicate conceptual relationships identified by graph-based analysis. Figure 5 displays the frequency of recurring philosophical motifs across ten Daoist texts in a heatmap format. Each cell represents the count of a motif within a text, highlighting recurring patterns across the corpus.

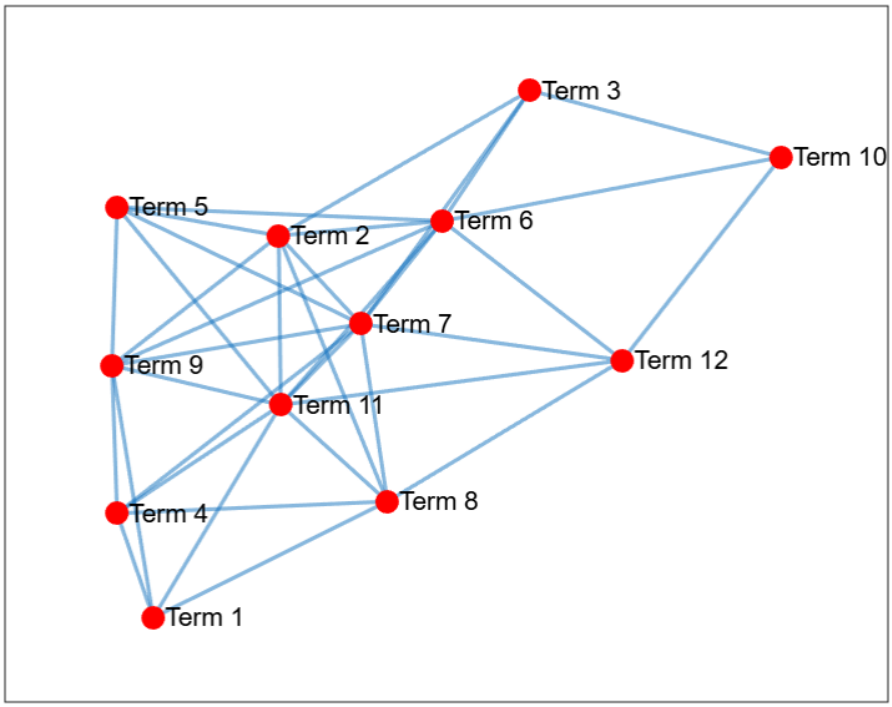

Figure 4 – Conceptual Network of Key Daoist Terms

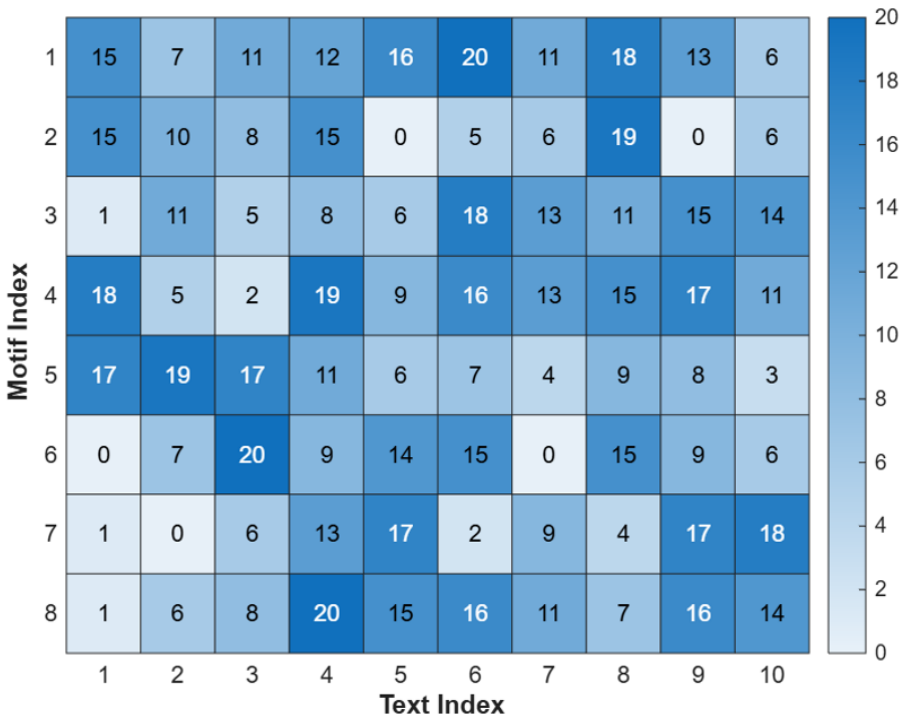

Figure 5 – Pattern Recognition: Recurring Motifs Across Texts

Graph analysis of concepts and texts in the Daoist Canon forms the third area of investigation for discovering structural properties of the texts. By constructing knowledge graphs of interconnected terms and fragments of texts, one not only discovers central themes and their interrelations, but also analyzes how central themes are connected with less central elements. The resulting perspective on the texts may challenge the prevailing linear way of reading the Daoist scriptures, since it reveals how concepts are interwoven across a large number of fragments of different texts. It is especially interesting to note the position of so-called peripheral nodes that traditionally have been neglected in analyses of the Canon. These nodes form a network of cross references and of influences, and in this way offer new insights into the complex structure of Daoist thought. In terms of further analysis of the Canon through the use of machine learning, the ability to recognize patterns within large datasets could uncover a number of facts about the Daoist Canon. For instance, there are many different accounts of the life of the Sage Laozi within the Canon, as well as different versions of the key texts that he wrote during his travels. Through pattern recognition, we can now look at these texts for the first time in over two thousand years and search for stylistic and thematic elements that have been reused in different contexts. For example, we can search for instances in which certain key terms or images from earlier texts have been echoed in later works. Through this type of analysis, we can begin to piece together a more complete history and understanding of the authorship of different texts in the Canon, including a better sense of the diverse range of ideological and stylistic perspectives that were likely to have been in play.

The quantitative analysis based on machine learning furthermore has implications for how to interpret studies of historical Daoist texts in general. Historical research on Daoist texts typically are studied within a fixed historical frame. They are analyzed in terms of a linear development where certain texts were written before others and thus provide a basis for later texts. It has, however, become clear through this study that when studying such a large corpus of texts as the Daoist Canon that there are many connections between different thematic elements within the texts. The findings of this study thus seem to indicate that there are many more aspects than those covered by historical research when studying texts. Thus historical studies of texts are enhanced by quantitative methodologies and these, in turn, are enhanced by qualitative methodologies and historical research. This, in turn, implies that when studying texts within the field of Daoist studies a more interdisciplinary approach should be adopted and that researchers from different fields should be able to make use of the methodologies

provided by the different fields to gain as comprehensive an understanding as possible of a text or set of texts under study. Our results enable us to integrate quantitative methodologies of data analysis for vast corpora of texts with qualitative inquiries in humanities research. The study of the extensive literature of Chinese religion as exemplified by the Canon presented in this article may benefit from a host of qualitative methods well-represented in studies of other civilizations, while all of them can be quantified for processing in vast corpora of texts. For our example of investigation into the extensive literature of Chinese religion, classical interpretations into the interrelations of numerous major elements in Daism may be well-complemented or even upgraded by various novel quantitative approaches to studies into massive bodies of texts. Thus, by fostering continued interaction of innovative computational approaches to very large databases created of results from qualitative studies into humanists with conventional methodologies which have been validated in the long history of investigation into Daoist thought and its extant vast Canon presented in this article, our study will actively and positively stimulate continued growth of studies of this field of very great interest and importance to human beings as such in this era of Information and Communication Technology that is characterized by huge information resources.

Conclusion

This study demonstrates the potential of machine learning and natural language processing to uncover semantic and structural patterns within the Daoist Canon. Semantic clustering revealed distinct groups of core concepts such as *Dao*, *De*, *Wu Wei*, and *Ziran*, while topic modeling highlighted dominant philosophical themes across the texts. Graph-based analysis identified central concepts and their interconnections, showing the structural hierarchy of Daoist thought, and pattern recognition detected recurring motifs such as yin-yang polarity and natural analogies, suggesting thematic evolution across historical periods. These results provide quantitative insights that complement traditional hermeneutic approaches, revealing connections and patterns previously difficult to detect. The findings underscore the value of computational techniques in digital humanities and philosophical studies, offering a framework for systematic exploration of large, complex corpora. Future work can expand the methodology to include other Chinese philosophical texts, integrate deep learning models for improved contextual understanding, and incorporate multimodal sources such as commentaries and historical annotations. By bridging computational analysis with classical scholarship, this approach opens new avenues for rigorous, scalable, and interpretable study of philosophical thought..

References

S. S. Yerragunta, "Text Mining and Sentiment Analysis of Major Religious and Philosophical Texts—Applying Natural Language Processing to Uncover Linguistic Patterns, Thematic Elements, and Emotional Tone," Int. J. Adv. Res. Ideas Innov. Technol., vol. 11, no. 5, 2025.

A. Felipe Ruiz, "Lexical, Sentiment and Correlation Analysis of Sacred Writings: A Tale of Cultural Influxes and Different Ways to Interpret Reality," Nat. Lang. Process. J., vol. 9, 2024.

D. McDonald, "A Text Mining Analysis of Religious Texts," J. Bus. Inquiry, vol. 13, no. 1, 2014.

B. Jiang and S. Kong, "Text Data Mining for Uncovering the Influence of Religion on Ancient Greek Philosophical Thought with Optimization," Int. J. Recent Innov. Trends Comput. Commun., vol. 11, no. 6, 2023.

M. Wu and D. Wang, "Automatic Compilation of a Pre Qin Philosophy Lexicon via Large Language Models," npj Heritage Sci., vol. 14, Art. 47, 2026.

R. Chandra and M. Ranjan, "Artificial Intelligence for Topic Modelling in Hindu Philosophy: Mapping Themes between the Upanishads and the Bhagavad Gita," PLoS One, vol. 17, no. 9, e0273476, 2022.

D. Forest and J.-G. Meunier, "NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus," 2005.

M. Verma, "Lexical Analysis of Religious Texts Using Text Mining and Machine Learning Tools," Int. J. Comput. Appl., vol. 168, no. 8, pp. 39–45, 2017.

C. Tagliapietra, "Automated Text Analysis in Theology: An Application," Theology Sci., vol. 23, no. 1, 2025.

H. Jelodar et al., “Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey,” arXiv, 2017.

R. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” J. Mach. Learn. Res., 2003.

“Topic Model,” Wikipedia, 2025.

“Word2vec,” Wikipedia, 2025.

P. J. Worth, “Word Embeddings and Semantic Spaces in Natural Language Processing,” Int. J. Intell. Sci., vol. 13, no. 1, 2023.

“BERT (Language Model),” Wikipedia, 2026.

A comprehensive overview of topic modeling: Techniques, applications,” ScienceDirect, 2025.

“Topic Modeling Algorithms and Applications: A Survey,” ScienceDirect, 2021.

J. Opitz et al., “Interpretable Text Embeddings and Text Similarity Explanation: A Survey,” arXiv, 2025.

J. Qiang, P. Chen, T. Wang, and X. Wu, “Topic Modeling over Short Texts by Incorporating Word Embeddings,” arXiv, 2016.

“Similarity of Word Embeddings with BERT: A Comprehensive Discussion,” Medium, 2025..