Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02355-w OPEN

Check for updates

Dissecting *The Analects*: an NLP-based exploration of semantic similarities and differences across English translations

Liwei Yang[1] & Guijun Zhou[1]✉

The Analects, a classic Chinese masterpiece compiled during China's Warring States Period, encapsulates the teachings and actions of Confucius and his disciples. The profound ideas it presents retain considerable relevance and continue to exert substantial influence in modern society. The availability of over 110 English translations reflects the significant demand among English-speaking readers. Grasping the unique characteristics of each translation is pivotal for guiding future translators and assisting readers in making informed selections. This research builds a corpus from translated texts of *The Analects* and quantifies semantic similarity at the sentence level, employing natural language processing algorithms such as Word2Vec, GloVe, and BERT. The findings highlight semantic variations among the five translations, subsequently categorizing them into "Abnormal," "High-similarity," and "Low-similarity" sentence pairs. This facilitates a quantitative discourse on the similarities and disparities present among the translations. Through detailed analysis, this study determined that factors such as core conceptual words, and personal names in the translated text significantly impact semantic representation. This research aims to enrich readers' holistic understanding of *The Analects* by providing valuable insights. Additionally, this research offers pragmatic recommendations and strategies to future translators embarking on this seminal work.

[1] School of Foreign Languages, Northeast Normal University, Changchun City, Jilin Province, China. ✉email: zhougj589@nenu.edu.cn

Introduction

The Analects, a Chinese classic encapsulating the teachings and philosophies of Confucius, were compiled by his disciples and their subsequent generations during the Warring States Period (476–221 BC) of Chinese history. Comprising 20 books, each with several chapters, *The Analects* are concise in structure. Though many of the passages are short, they often carry profound meaning (Lin, 2010: i). The doctrines presented in *The Analects* have exerted a profound influence on Chinese history (Watson, 2007; Chin, 2014). The sayings fundamentally regard ethics, guiding Confucius' disciples and their successors by strengthening resolve, shaping conduct, and preparing for practical life crises. Over the centuries, the teachings have proved highly effective (Brooks & Brooks, 1998: viii).

The original text of *The Analects* is written in Classical Chinese. Non-Chinese readers, depend on translations of the text for understanding Confucius. For them, the reliability of a translation is a basic expectation (Ni, 2017: 18). Throughout the history of English translations of *The Analects*, in order to better interpret and elucidate the profound connotations of *The Analects*, translators are critical of prior translations and attempt improvements (Chesterman, 2000). Concurrently, retranslation serves to meet the demands of the target culture (Desmidt, 2009). From the initial English rendition of *The Analects* in 1691 up to 2022, more than 110 English translations have been produced. These translations have not only facilitated the introduction of this classic work to Western readers but have also broadened the range of interpretive possibilities available to readers. However, a prevailing challenge with all the translations of *The Analects* is the distortion evident during translation (Roberts, 2020: 11), although *The Analects* has been one of the most influential books in China for the last 2500 years, many of the English translations of this book have been incomprehensible (Hedstrom, 2020: i). For instance, James Legge's rendering of "天 Tian" as "Heaven"; this choice evokes connotations linked to the Judeo-Christian tradition, which are not present in Chinese culture (Ames & Rosemont, 1998: 46).

For native English speakers, interpreting the intricate philosophies and nuanced content within *The Analects* is often considered a daunting task owing to the profound differences in linguistic structures and the cultural distinctions between English and Classical Chinese. This linguistic and cultural disparity poses added challenges for translators and compounds the reading difficulties for Western readers, they often have to rely solely on their own interpretations to grasp the complex semantics of *The Analects* (Slingerland, 2003: viii), which is certainly not conducive to their better understanding of its original meaning. Therefore, the differences and similarities among various English translations of *The Analects* have also garnered considerable attention and reflection from scholars. In reviewing this field, numerous scholars have explored the variances and rationales underlying different translations of *The Analects*, attributing the discrepancies to factors including translators' life experiences, academic backgrounds, expertise in Sinology, bilingual proficiency, intended translation purposes, and chosen translation strategies (Yang, 2014; Hou & Sun, 2019; Liu (2023)). While these studies offer insights into various English translations of *The Analects*, they frequently result in interpretations grounded in subjectivity. Several corpus-based studies have shed light on the macro-linguistic characteristics in different translations of *The Analects* using quantitative analysis. Though these analyses provide some objectivity, their emphasis predominantly lies in theoretical exploration, frequently omitting the pragmatic considerations needed to deepen readers' understanding of *The Analects*.

As translation studies have evolved, innovative analytical tools and methodologies have emerged, offering deeper insights into textual features. Among these methods, NLP stands out for its potent ability to process and analyze human language. The current research underscores the multifaceted functionalities of NLP, encompassing text generation (Koplin, 2023; Seifossadat & Sameti, 2023; He et al., 2022), textual data mining (Shahbazi & Byun, 2022; Gutierrez et al., 2021; Green, 2015), phonetics (Iliev & Ilieva, 2023; Nissan, 2017), sentiment analysis (Ma et al., 2023; Li et al., 2022; Oh & Yi, 2022) and Semantic similarity computation (Chang et al., 2023; Jiang et al., 2017; Iosif & Potamianos, 2015). Within digital humanities, merging NLP with traditional studies on *The Analects* translations can offer more empirical and unbiased insights into inherent textual features. This integration establishes a new paradigm in translation research and broadens the scope of translation studies.

This study employs natural language processing (NLP) algorithms to analyze semantic similarities among five English translations of *The Analects*. To achieve this, a corpus is constructed from these translations, and three algorithms—Word2-Vec, GloVe, and BERT—are applied to assess the semantic congruence of corresponding sentences among the different translations. Analysis reveals that core concepts, and personal names substantially shape the semantic portrayal in the translations. In conclusion, this study presents critical findings and provides insightful recommendations to enhance readers' comprehension and to improve the translation accuracy of *The Analects* for all translators.

Materials and methods

Sample selection. A translation should convey the necessary information, while taking into account readers' responses. Consequently, the acceptability of a translation emerges as a crucial factor, especially when the goal is to promote reader acceptance of Chinese Classical texts. In this regard, some researchers have utilized a Python crawler to collect data on the volume of reviews, downloads, and readership for various English translations of *The Analects* from platforms such as Amazon, Goodreads, Archive, Google Scholar, and PDF-Drive. This data is then used to gauge the attention each translation attracts from readers (Yang & Zhou, 2022). Building upon previous research, this study selects five high acceptability English translations of *The Analects* by D. C. Lau, James Legge, William Jennings, Edward Slingerland, and Burton Watson as research samples.

Corpus building. This study obtains high-resolution PDF versions of the five English translations of *The Analects* through purchase and download. The first step entailed establishing preprocessing parameters, which included eliminating special symbols, converting capitalized words to lowercase, and sequentially reading the PDF file whilst preserving the English text. Subsequently, this study aligned the cleaned texts of the translations by Lau, Legge, Jennings, Slingerland, and Watson at the sentence level to construct a parallel corpus. The original text of *The Analects* was segmented using a method that divided it into 503 sections based on natural section divisions. This study further subdivided these segments using punctuation marks, such as periods (.), question marks (?), and semicolons (;). However, it is crucial to note that these subdivisions were not exclusively reliant on punctuation marks. Instead, this study followed the principle of dividing the text into lines to make sure that each segment fully expresses the original meaning. Finally, each translated English text was aligned with its corresponding original text.

During our study, this study observed that certain sentences from the original text of *The Analects* were absent in some English translations. To maintain consistency in the similarity calculations within the parallel corpus, this study used "None" to represent untranslated sections, ensuring that these omissions did not impact our computational analysis. The analysis encompassed a total of 136,171 English words and 890 lines across all five translations.

Table 1 illustrates this process using lines 0–2 as examples. The complete corpus can be accessed on figshare: https://doi.org/10.6084/m9.figshare.23931291, specifically in Attachment 4.

Modeling of semantic similarity calculation. Evaluating translated texts and analyzing their characteristics can be achieved through measuring their semantic similarities, using Word2Vec, GloVe, and BERT algorithms. This study conduct triangulation method among three algorithms to ensure the robustness and reliability of the results.

The Word2Vec algorithm, introduced by Mikolov et al. (2013) under the aegis of Google, is a predominant model in Natural Language Processing (NLP) for unsupervised learning of semantic knowledge from a large text corpus. Word2Vec encompasses a suite of models architectured to generate word embeddings, where words with similar meanings would have similar vector representations. It represents words in a high-dimensional continuous vector space. Within this dimensional space, words with semantic similarities are spatially proximate. Example Explanation: In the sentence "The cat sits on the mat", word2vec will represent "cat" and "sits" by vectors, say [0.2, −0.4, 0.7, …] and [−0.1, 0.6, −0.3, …] respectively. Each number in the vector could represent an aspect of the word's relationship to its neighboring words in the training corpus. For example, if "cat" often appears near the word "pet," then one dimension in the vector could somewhat represent the concept of "pet-ness."

The Global Vectors for Word Representation (GloVe) model is designed to create word embeddings by considering global statistics of a corpus (Pennington et al., 2014). It constructs a large matrix of word co-occurrence probabilities and factorizes this matrix to produce embeddings. Example Explanation: For words like "cat" and "sits" from the same sentence, GloVe might represent them as [0.3, −0.1, 0.8, …] and [0.2, 0.5, −0.2, …] respectively. In contrast to word2vec, the embedding for "cat" in GloVe not only considers its immediate contextual environment but also for how frequently "cat" co-occurs with every other word in the corpus, allowing it to capture more global semantic information about the word.

BERT (Bidirectional Encoder Representations from Transformers) employs transformer architecture to integrate contextual information bidirectionally (left-to-right and right-to-left), allowing each word to have a dynamic embedding influenced by its surrounding words. It enables the model to grasp the meaning of each word in a more nuanced way. Example Explanation: When we have the word "cat" in the sentence "The cat sits on the mat," BERT generates an embedding like [0.9, −0.2, 0.3, …]. If "cat" appears in a different context, its embedding will be different, say [−0.1, 0.7, 0.2, …]. This dynamic representation allows BERT to capture the nuanced meanings and usages of words based on different contexts they appear in. The BERT algorithm is universally acknowledged as one of the most exhaustive algorithms extant in the domain.

All these models aim to provide numerical representations of words that capture their meanings. While word2vec and GloVe generate static embeddings, meaning the representation of a word is fixed regardless of context, BERT creates dynamic embeddings, indicating that it takes into account the contextual environment of a word within a sentence to formulate its representation.

In this study, Python 3.6 to implement the NLP semantic similarity algorithmic models. These models computed similarity

Table 1 A corpus of five English translations of *The Analects*.

NO.	James Legge	D. C. Lau	William Jennings	Edward Slingerland	Burton Watson
0	The Master said, Is it not pleasant to learn with a constant perseverance and application?	The Master said, 'Is it not a pleasure, having learned something, to try it out at due intervals?	To learn," said the Master, "and then to practise opportunely what one has learnt—does not this bring with it a sense of satisfaction?	The Master said, "To learn and then have occasion to practice what you have learned—is this not satisfying?	The Master said, Studying, and from time to time going over what you've learned— that's enjoyable, isn't it?
1	Is it not delightful to have friends coming from distant quarters?	Is it not a joy to have friends come from afar?	To have associates in study coming to one from distant parts—does not this also mean pleasure in store?	To have friends arrive from afar—is this not a joy?	To have a friend come from a long way off—that's a pleasure, isn't it?
2	Is he not a man of complete virtue, who feels no discomposure though men may take no note of him?	Is it not gentlemanly not to take offence when others fail to appreciate your abilities?'	And are not those who, while not comprehending all that is said, still remain not unpleased to hear, men of the superior order?	To be patient even when others do not understand—is this not the mark of the gentleman?"	Others don't understand him, but he doesn't resent it—that's the true gentleman, isn't it?

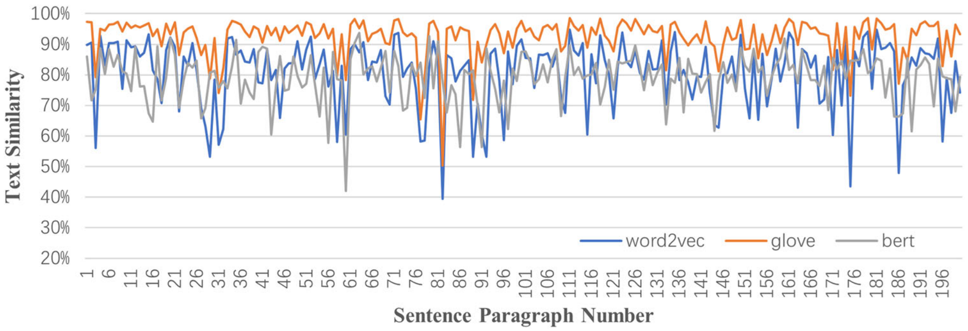

Fig. 1 Textual Similarity Analysis in the Jennings vs. Slingerland Dataset. Varied colored lines illustrate the computational trends of Word2Vec, GloVe, and BERT Algorithms.[1].

based on the corpus established above. This study presented the English translations in Excel format and divided them into separate sheets corresponding to each translator. The program was structured as follows: ① Import Python modules such as re, pandas, streamlit, and numpy to handle text processing and data management; ② Import the calculation models of Word2Vec, GloVe, and BERT to perform the semantic similarity calculation; ③ This study set the upload and save formats for files, ensuring the program correctly processed the input and output data. ④ This study encapsulated the calculation procedure within a user-friendly interface for ease of operation. Following the outlined logic, the developed semantic similarity comparison model can compute and display the semantic similarity of the text. The Python code is uploaded to figshare: https://doi.org/10.6084/m9.figshare.23931291, Attachment 2. The operation procedure of the calculation program is shown in Attachment 5

Results

Overall trend of semantic similarity of sentences pairs. This study assigned a unique number to each sentence in the corpus, with numbers ranging from 0 to 889. This study subsequently utilized NLP models - Word2Vec, GloVe, and BERT—to compute the semantic similarity between the corresponding sentences spanning the five translations of *The Analects*. Each translation of *The Analects* contains 890 sentences. The comparison between sentence pairs across the five translations yields 8,900 results per algorithm. Thus, across the three employed algorithms—Word2Vec, GloVe, and BERT—a total of 26,700 results are generated. collectively generated by the three algorithms. Figure 1 demonstrates the semantic similarity between Jennings and Slingerland's translations of *The Analects* first 200 sentences, as calculated by the Word2Vec, GloVe, and BERT algorithms. For all translation sentence pair detailed calculation results, please refer to figshare: https://doi.org/10.6084/m9.figshare.23931291, Attachment 3.

The x-axis represents the sentence numbers from the corpus, with sentences 0-199 taken as an example due to space limitations. For each sentence number on the x-axis, a corresponding semantic similarity value is generated by each algorithm. These values are then connected from 0 to 199 to form a line graph. The similarity values generated by the three algorithms can be compared. The y-axis represents the semantic similarity results, ranging from 0 to 100%. A higher value on the y-axis indicates a higher degree of semantic similarity between sentence pairs.

Through the analysis of our semantic similarity calculation data, this study finds that there are some differences in the absolute values of the results obtained by the three algorithms. Several factors, such as the differing dimensions of semantic word vectors used by each algorithm, could contribute to these dissimilarities. However, this study does not delve into these underlying causes. Figure 1 primarily illustrates the performance of three distinct NLP algorithms in quantifying semantic similarity. As depicted in Fig. 1, although there are variations in the absolute values among the algorithms, they consistently reflect a similar trend in semantic similarity across sentence pairs. This suggests that while the selection of a specific NLP algorithm in practical applications may hinge on particular scenarios and requirements, in terms of overall semantic similarity judgments, their reliability remains consistent. For example, a sentence that exhibits low similarity according to the Word2Vec algorithm tends to also score lower on the similarity results in the GloVe and BERT algorithms, although it may not necessarily be the lowest. In contrast, sentences garnering high similarity via the Word2Vec algorithm typically correspond with elevated scores when evaluated by the GloVe and BERT algorithms. To provide an objective comparison of the semantic similarity across the English translations of *The Analects* by Legge, Lau, Jennings, Slingerland, and Watson, this study calculated the average semantic similarity scores using all three algorithms and then sorted these results.

Distribution of semantic similarity of sentence pairs. Taking the average of the semantic similarity scores generated by the three embedding models to measure the ranking and distribution among sentence pairs, this study shows the distribution of the similarity results of the corresponding sentence pairs among the five English translations of *The Analects*, as shown in Table 2.

Since each translation contains 890 sentences, pairing the five translations produces 10 sets of comparison results, totaling 8900 average results.

The data presented in Table 2 elucidates that the semantic congruence between sentence pairs primarily resides within the 80–90% range, totaling 5,507 such instances. Moreover, the pairs of sentences with a semantic similarity exceeding 80% (within the 80–100% range) are counted as 6,927 pairs, approximately constituting 78% of the total amount of sentence pairs. This forms the major component of all results in the semantic similarity calculations. Most of the semantic similarity between the sentences of the five translators is more than 80%, this demonstrates that the main body of the five translations captures the semantics of the original Analects quite well.

Conversely, the outcomes of semantic similarity calculations falling below 80% constitute 1,973 sentence pairs, approximating 22% of the aggregate number of sentence pairs. Although this subset of sentence pairs represents a relatively minor proportion, it holds pivotal significance in impacting semantic representation amongst the varied translations, unveiling considerable semantic variances therein. To delve deeper into these disparities and their

Table 2 Distribution of similarity calculation results of the five English translations of *The Analects*.

Sentence pair	Threshold Value				
	≤60%	60% ~ 70%	70% ~ 80%	80% ~ 90%	90% ~ 100%
	Sentence pair Quantity				
Legge VS Lau	6	26	150	584	124
Legge VS Jennings	9	39	157	514	171
Legge VS Slingerland	4	26	162	565	133
Legge VS Watson	8	19	159	519	185
Lau VS Jennings	9	39	194	576	72
Lau VS Slingerland	3	10	88	525	264
Lau VS Watson	6	22	133	592	137
Jennings VS Slingerland	11	42	212	545	80
Jennings VS Watson	14	39	227	561	49
Slingerland VS Watson	5	24	130	526	205

Table 3 Some samples of abnormal results.

NO.	Translator1	Translator2	Word2Vec	GloVe	BERT	AVG
410	Those known for virtuous conduct: Yan Hui...(ES)	None (WJ)	0.090	-	0.293	0.192
289	The Master said, Extravagance leads to...(ES)	None (BW)	0.058	-	0.362	0.210
486	Not for the sake of wealth, But simply for ...(ES)	None (BW)	0.149	-	0.298	0.224
203	The Master said, Yung is right in ...(Lau)	None (ES)	0.237	-	0.235	0.236
289	The Master said, Extravagance means...(Lau)	None (BW)	0.167	-	0.348	0.258
397	When summoned by his lord, he would ...(ES)	None (BW)	0.291	-	0.409	0.350
345	The Master said, I have yet to meet the man...(Lau)	None (WJ)	0.276	-	0.434	0.355

Note: JL-James Legge; **Lau**-D. C. Lau; **WJ**-William Jennings; **ES**-Edward Slingerland; **BW**-Burton Watson.

foundational causes, a more comprehensive and meticulous analysis is slated for the subsequent sections.

Analysis of calculation results

Abnormal results. The semantic similarity analysis yielded 33 outcomes categorized as abnormal, signifying significant deviation from expected values. Some examples of abnormal results are shown in Table 3. For more detailed results, please refer to figshare: https://doi.org/10.6084/m9.figshare.23931291, Attachment 3, "Abnormal Results."

In Table 3, "NO." refers to the specific sentence identifiers assigned to individual English translations of *The Analects* from the corpus referenced above. "Translator 1" and "Translator 2" correspond to the respective translators, and their translations undergo a comparative analysis to ascertain semantic concordance. Translations from the five translators undergo pairwise comparisons. The columns labeled "Word2Vec," "GloVe," and "BERT" present outcomes derived from their respective semantic similarity algorithms. Subsequently, the "AVG" column presents the mean semantic similarity value, computed from the aforementioned algorithms, serving as the basis for ranking translations by their semantic congruence. By calculating the average value of the three algorithms, errors produced in the comparison can be effectively reduced. At the same time, it provides an intuitive comparison of the degrees of semantic similarity.

For instance, a value in the table, such as "0.090" under the Word2Vec column, implies that the renditions of the 410th sentence in the corpus of *The Analects* by two translators exhibit a mere 9% semantic concordance (with 100% being the maximum), indicating that the translated texts are almost entirely dissimilar in meaning.

As delineated in Section 2.1, all aberrant outcomes listed in the above table are attributable to pairs of sentences marked with "None," indicating untranslated sentences. When the Word2Vec and BERT algorithms are applied, sentences containing "None" typically yield low values. The GloVe embedding model was incapable of generating a similarity score for these sentences. This study designates these sentence pairs containing "None" as Abnormal Results, aiding in the identification of translators' omissions. These outliers scores are not employed in the subsequent semantic similarity analyses.

High-similarity sentence pairs. Upon comparing the five English translations within the aforementioned corpus, this study found that 6,927 of all sentence pairs demonstrated a semantic similarity exceeding 80%. These sentence pairs represent 77.83% of the corpus. Although the translations employ varying linguistic expressions, their semantic conveyance remains largely similar. For detailed calculation results, see figshare: https://doi.org/10.6084/m9.figshare.23931291, Attachment 3, titled "High Similarity Results."

Among the five translations, only a select number of sentences from Slingerland and Watson consistently retain identical sentence structure and word choices, as in Table 4. The three embedding models used to evaluate semantic similarity resulted in a 100% match for sentences NO. 461, 590, and 616. In other high-similarity sentence pairs, the choice of words is almost identical, with only minor discrepancies. However, as the semantic similarity between sentence pairs decreases, discrepancies in word selection and phraseology become more pronounced.

Low-similarity sentence pairs. Out of the entire corpus, 1,940 sentence pairs exhibit a semantic similarity of ≤ 80%, comprising 21.8% of the total sentence pairs. These low-similarity sentence pairs play a significant role in determining the overall similarity between the different translations. They further provide valuable insights into the characteristics of different translations and aid in identifying potential errors. By delving deeper into the reasons

Table 4 Some samples of High-similarity sentence pair.

NO.	Translator1	Translator2	Word2Vec	GloVe	BERT	AVG
616	Qu Boyu sent a messenger to Confucius. (ES)	Qu Boyu sent a messenger to Confucius. (BW)	1.000	1.000	1.000	1.000
461	Sima Niu asked about the gentleman. (ES)	Sima Niu asked about the gentleman. (BW)	1.000	1.000	1.000	1.000
590	Zilu asked about the complete person. (ES)	Zilu asked about the complete person. (BW)	1.000	1.000	1.000	1.000
757	Tzu-chang asked Confucius about benevolence. (Lau)	Zizhang asked Confucius about humaneness. (BW)	1.000	1.000	0.959	0.986
471	Zigong asked about governing. (ES)	Zigong asked about government. (BW)	1.000	1.000	0.957	0.986
417	When Yan Yuan died, the Master said, Alas! Heaven is destroying me! Heaven is destroying me! (JL)	When Yan Yuan died, the Master said, Ah, Heaven is destroying me! Heaven is destroying me! (BW)	1.000	1.000	0.937	0.979

Note: JL-James Legge; **Lau**-D. C. Lau; **WJ**-William Jennings; **ES**-Edward Slingerland; **BW**-Burton Watson.

Table 5 Some samples of Low-similarity sentence pair.

NO.	Translator1	Translator2	Word2Vec	GloVe	BERT	AVG
79	He replied, It is beyond my power. (WJ)	Ran Qiu replied, I was not. (ES)	0.231	0.494	0.545	0.423
218	How admirable Hui is! (Lau)	Aye, a right worthy soul was he! (WJ)	0.318	0.210	0.837	0.455
84	Then, said the other, rules of ceremony require to have a background! (WJ)	Zixia said, so ritual comes afterward? (BW)	0.333	0.504	0.545	0.461
43	Zi Xia asked what filial piety was. (JL)	Tzu-hsia asked about being filial. (Lau)	0.292	0.475	0.629	0.465
41	Zi You asked what filial piety was. (JL)	To a like question put by Tsz-yu, (WJ)	0.410	0.403	0.698	0.504

Note: JL-James Legge; **Lau**-D. C. Lau; **WJ**-William Jennings; **ES**-Edward Slingerland; **BW**-Burton Watson.

behind this substantial difference in semantic similarity, this study can enable readers to gain a better understanding of the text of *The Analects*. Furthermore, this analysis can guide translators in selecting words more judiciously for crucial core conceptual words during the translation process.

The sentences in *The Analects*, as rendered by the five different translators, display considerable variations in terms of words and sentence structure. This diversity in translation, which can range from variations in word choice to distinct sentence structures, can confuse readers trying to understand the original text, as they may encounter different interpretations in different translations. The goal of analyzing these differences is to assist readers in more accurately comprehending the original meaning of *The Analects*. This analysis further aims to offer guidance to translators striving for greater accuracy in their renditions. Table 5 above provides a snapshot of the calculations, whereas a detailed tabulation of results is available in figshare: https://doi.org/10.6084/m9.figshare.23931291, Attachment 3, titled "Low Similarity Results."

Discussion

The discussion of the similarities. As previously discussed, this study has set an 80% threshold to ascertain the semantic similarity between sentence pairs. This cut-off point is derived from the distribution of similarity scores presented in Table 2. Then further segmented the range between 80% and 100% into four distinct intervals: 80–85%, 85–90%, 90–95%, and 95–100%. This study counted the number of translated sentences falling within these intervals and calculated their corresponding percentages. The results of this analysis are presented in Table 6.

Within the similarity score intervals of 80–85% and 85–90%, the distributions of sentences across all five translators is more balanced, each accounting for about 20%. However, translations by Jennings present fewer instances in the highly similar intervals of 95–100% (1%) and 90–95% (14%). Contrastingly, Slingerland's translation features a higher percentage of sentences with similarity scores within the 95–100% interval (30%) and the 90–95% interval (24%) compared to the other translators. Watson's translation also records a substantially higher percentage (34%) within the 95–100% range compared to other translators.

A detailed examination of Jennings's translation unveils a remarkably distinct sentence structure. The original text of *The Analects* is characterized as a dialogue or a question-and-answer session, with the majority of sentences initiating with phrases such as "the master said," "Confucius said," and "Confucius' disciple said." Jennings argues that in perhaps three-fourths of the number of paragraphs, a sentence is introduced by the formula "The Master said," which in English, after a while, becomes wearisome (Jennings, 1895: 35). Jennings's methods of "inversion" and "combining sentences under one head" lead to his translations exhibiting a lower similarity to those of other translators. In 1895, a time when few translations were available for his reference, Jennings daringly made structural adjustments to his translation. His boldness in restructuring the translation demonstrates his deep understanding of Chinese culture. This approach, adopted with the readers' convenience in mind, represented a unique and innovative strategy in that epoch.

The Slingerland and Watson translations of *The Analects* were published in 2003 and 2007 respectively. Given its extensive interpretative history, *The Analects* furnishes a wealth of reference material for contemporary translators. This is evident in Slingerland's translation, where he acknowledges: "The task of translating the primary text of *The Analects* was considerably eased by the labor of previous translators, upon whose work I have built and whose well-turned phrases I have, in many cases, been entirely helpless to improve upon" (Slingerland, 2003: x). Slingerland's translation of *The Analects* utilizes an array of para-texts to elucidate each sentence, a method aligned with the concept of a "thick translation." Slingerland's translation is specifically designed for classroom instruction, and the para-texts help students better understand *The Analects*. Thus, it is reasonable to categorize Slingerland's translation as a classic example of a "thick translation" interpretation of *The Analects*. Watson (2007), on the other hand, acknowledges that his translation draws upon 11 different translations. Contrarily, Watson does not heavily rely on para-texts. Nevertheless, he provides multiple translations for certain complex sentences in

Table 6 Distribution of different translators' sentences in the high similarity interval.

Distribution interval	Legge	Lau	Jennings	Slingerland	Watson
80% ~ 85% occurrence number	873	858	927	839	899
(Percentage for the range)	(20%)	(20%)	(21%)	(19%)	(20%)
85% ~ 90% occurrence number	1309	1419	1269	1322	1299
(Percentage for the range) number	(20%)	(21%)	(19%)	(20%)	(20%)
90% ~ 95% occurrence number	602	569	371	649	539
(Percentage for the range)	(22%)	(21%)	(14%)	(24%)	(20%)
95% ~ 100% occurrence number	11	28	1	33	37
(Percentage for the range)	(10%)	(25%)	(1%)	(30%)	(34%)

the main text. This unique approach effectively maintains the balance between translation accuracy and readability for readers throughout the entire text.

The discussion of the differences. The data displayed in Table 5 and Attachment 3 underscore significant discrepancies in semantic similarity (values ≤ 80%) among specific sentence pairs across the five translations, with a particular emphasis on variances in word choice. As mentioned earlier, the factors contributing to these differences can be multi-faceted and are worth exploring further.

Analysis of high-frequency words. The translation of *The Analects* contains several common words, often referred to as "stop words" in the field of Natural Language Processing (NLP). These words, such as "the," "to," "of," "is," "and," and "be," are typically filtered out during data pre-processing due to their high frequency and low semantic weight. Similarly, words like "said," "master," "never," and "words" appear consistently across all five translations. However, despite their recurrent appearance, these words are considered to have minimal practical significance within the scope of our analysis. This is primarily due to their ubiquity and the negligible unique semantic contribution they make. For these reasons, this study excludes these two types of words-stop words and high-frequency yet semantically non-contributing words from our word frequency statistics.

Table 7 provides a representation that delineates the ranked order of the high-frequency words extracted from the text. This visualization aids in identifying the most critical and recurrent themes or concepts within the translations.

This study has categorized the high-frequency words into two main groups. The first category consists of core conceptual words in the text, which embody cultural meanings that are influenced by a society's customs, behaviors, and thought processes, and may vary across different cultures. These recurrent words in *The Analects* include key cultural concepts such as "君子 Jun Zi, 小人 Xiao Ren, 仁 Ren, 道 Dao, 礼 Li," and others (Li et al., 2022). A comparison of sentence pairs with a semantic similarity of ≤ 80% reveals that these core conceptual words significantly influence the semantic variations among the translations of *The Analects*. The second category includes various personal names mentioned in *The Analects*. Our analysis suggests that the distinct translation methods of the five translators for these names significantly contribute to the observed semantic differences, likely stemming from different interpretation or localization strategies.

Table 8a, b display the high-frequency words and phrases observed in sentence pairs with semantic similarity scores below 80%, after comparing the results from the five translations. This set of words, such as "gentleman" and "virtue," can convey specific meanings independently. Furthermore, when combined with other words, these terms can form semantically rich phrases like "good man," "mean man," and "superior man." These high-frequency words and phrases further underscoring the importance of the essential core concepts found in *The Analects*.

Table 8c displays the occurrence of words denoting personal names in *The Analects*, including terms such as "zi, Tsz, Tzu, Lu, Yu," and "Kung." These terms can appear individually or in combination with other words and often represent important characters within the text. The translation of these personal names exerts considerable influence over the variations in meaning among different translations, as the interpretation of these names may vary among translators.

The translators of *The Analects*, particularly the five referenced in this study, allocate considerable space in their appendices to explaining core concepts such as "君子 Jun Zi" and "仁 Ren." Furthermore, they devote considerable space to listing the characters involved in *The Analects* and briefly introducing them. Considering the aforementioned statistics and the work of these scholars, it is evident that the translation of core conceptual terms and personal names plays a significant role in shaping the semantic expression of *The Analects* in English.

The effect of core conceptual words on semantic representations. For comparative analysis, this study has compiled various interpretations of certain core conceptual terms across five translations of *The Analects*.

The table presented above reveals marked differences in the translation of these terms among the five translators. These disparities can be attributed to a variety of factors, including the translators' intended audience, the cultural context at the time of translation, and the unique strategies each translator employed to convey the essence of the original text. The term "君子 Jun Zi," often translated as "gentleman" or "superior man," serves as a typical example to further illustrate this point regarding the translation of core conceptual terms.

In his translation, Legge almost uniformly translates "君子 Jun Zi" as "superior man," whereas Jennings offers a more diverse range of translations. Although Jennings also frequently uses "superior man" as Legge does, he argues that it is sometimes challenging to consistently translate "君子 Jun Zi" with the same English term (Jennings, 1895: 31). As a result, he supplements his translation with additional terms such as "gentleman," "great man," "noble-minded man," and "masterly man." D. C. Lau, Slingerland, and Watson, use the word "gentleman" to translate "君子 Jun Zi." As the gentleman is the ideal moral character, it is not to be expected that a man can become a gentleman without a great deal of hard work or cultivation, as the Chinese call it (Lau, 1979). According to William Soothill, the term "君子 Jun Zi" is roughly equivalent to the concept of a "gentleman" in its most noble and virtuous sense (Soothill, 1910). In Lin Yutang's translation of *The Analects*, "gentleman" is most often used to correspond to "君子 Jun Zi," while a few translations use "superior man" (Lin, 1941). Ezra Pound's translation uses "gentleman" a few times (Pound, 1969). Arthur Waley also used "gentleman" in his translation (Waley, 1997). From these

Table 7 Ranking subject words in the corpus with similarity ≤80%.

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
You	857	Indeed	123	Filial	73	Zhang	58
Man	640	Zigong	120	Well	71	Son	58
One	439	Person	116	Ran	69	Official	57
Will	305	Words	114	Zixia	69	Common	57
Zi	255	Zilu	111	Practice	68	State	55
Gentleman	241	Government	99	Right	67	Rules	55
Without	241	Duke	95	Chi	65	Wise	51
Tsz	233	Learning	95	Mind	65	Time	51
Tzu	209	Small	87	Fan	64	Shu	51
Virtue	207	Gong	86	Humaneness	64	Simply	50
Lu	204	Three	86	Great	62	Propriety	50
People	192	Office	82	Music	62	Xia	50
Good	164	Chang	82	Ritual	62	Ji	49
Yu	163	Friends	82	Zhong	61	Word	49
Men	154	Ch	77	Heaven	60	True.	48
Way	153	Mean	75	Chung	59	Goodness	48
Superior	145	Zizhang	75	Yen	59	Love	48
Kung	141	Petty	73	Disciple	59	benevolent	39

Table 8 a The combination words frequency of "君子 Jun Zi" and "小人 Xiao Ren" in Low similarity texts. b. The word frequency of other Core Conceptual Words in Low similarity texts. c. The word frequency of personal names in Low similarity texts.

		man/men		person/people	
Word	**Frequency**	**Combination of words 1**	**Frequency**	**Combination of words 2**	**Frequency**
superior	145	superior+man/men	130	/	/
small	87	small+man/men	36	/	/
petty	73	petty+man/men	41	petty+person	24
mean	75	mean+man	24	mean+people	1
good	164	good+man	6	good+people/person	6
common	57	common+man/men	15	common+people/person	34
great	62	great+man	9	/	/
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
gentleman	241	virtue	207	Way	153
Humaneness	64	Ritual	62	Right	67
Heaven	60	Wise	51	Propriety	50
Goodness	48	benevolent	39	Rules	55
Zi	255	Tsz	233	Tzu	209
Lu	204	Yu	163	Kung	141
Zilu	111	Zigong	120	Gong	86
Chang	82	Zizhang	75	Zixia	69
Ran	69	Chi	65	Fan	64
Zhong	61	Yen	59	Chung	59
Zhang	58	Shu	51	Xia	50

observations, one could infer that translators D. C. Lau, Slingerland, and Watson likely drew inspiration from the translations by William Soothill, Arthur Waley, Lin Yutang, and Ezra Pound. "Gentleman" comes from the French word "gentil homme," meaning "noble, noble people." "君子 Jun Zi" means "son of a gentleman," originally referring to nobility. Over time, both terms evolved within their respective cultural contexts to denote individuals of moral character, demonstrating an intriguing parallel between the moral traditions of these distinct cultures.

The observations regarding translation differences extend to other core conceptual words in *The Analects*, a subset of which is displayed in Table 9 due to space constraints. Translators often face challenges in rendering core concepts into alternative words or phrases while striving to maintain fidelity to the original text. Yet, even with the translators' understanding of these core concepts, significant variations emerge in their specific word choices. These variations, along with the high frequency of core concepts in the translations, directly contribute to differences in semantic representation across different translations.

The effect of personal names on semantic representations. Ancient Chinese personal names typically comprised a "formal name," which served as their social identifier, and a "style name," which often complemented the "formal name", often expressing a character trait or personal philosophy. "Formal names", assigned in early childhood, were typically used while addressing elders or superiors, while "style names", adopted during adulthood, were commonly used in interactions among peers or friends. In addition to formal and "style names", ancient Chinese individuals also adopted "pseudonyms," "nicknames," and "aliases," all of which are reflected in the text of *The Analects*. In third-person narratives, the "formal name" was commonly used, while in first-person speech or when Confucius addressed his disciples, the

Table 9 A comparison of the main core conceptual words in the five translations of *The Analects*.

Chinese and pronunciation	James Legge	D. C. Lau	William Jennings	Edward Slingerland	Burton Watson
君子 Jun Zi	superior man accomplished scholar	Gentleman	superior man great man masterly man Men of loftier mind noble-minded man gentleman ideal man good man	Gentleman	Gentleman
小人 Xiao Ren	mean man small man lower people	small man	common man small-minded man inferior man common person	petty person petty man common person	petty man
仁 Ren	Virtue Virtuous Benevolent	Benevolent Benevolence know the man	Philanthropy good-will fellow-men Philanthropic Philanthropy right feeling proper feelings right likings good feeling good nature	Goodness Good Character	Humaneness Humane

"style name" was often employed (Slingerland, 2003: xi). The complex system of personal names in *The Analects* can easily lead to confusion in translation, particularly because the text may use different names for the same individual within different contexts. For instance, in *The Analects*, the historical figure "Zhong You" (542-480 BC) is also referred to by the style names "Zi Lu" and "Ji Lu," and is additionally known by the nickname "Yu." Similarly, "Ran Qiu" (522-? BC) is also known by the style name "Zi You," and has the nicknames "Ran You" and "Ran Zi," or is referred to as "Qiu." For English-speaking readers, unfamiliar with the complexities of the ancient Chinese naming system, these interchangeable names can cause confusion, potentially leading to significant discrepancies in the interpretation and understanding of various translations Table 10.

The translators, Legge, D. C. Lau, and Watson, strictly adhere to the original text of *The Analects* when translating personal names. This approach effectively preserves the nomenclature used in the original text. However, readers without extensive background knowledge might find it challenging to connect the characters with their varying names throughout the text. Importantly, this does not indicate a translation flaw; instead, it underscores the three translators' deep respect for the original text. In an effort to enhance reader comprehension, Jennings and Slingerland simplified the translations of names in *The Analects*. Jennings (1895: 35) acknowledges the confusion surrounding the multiple names and opts for a single, consistent name in his translation. At times, he even omits the name entirely when a personal pronoun suffices. Similarly, Slingerland, aiming to serve students with limited background in Sinology, ensures a consistent representation of names throughout the text.

While some translators faithfully mirror the original text, capturing the unique aspects of ancient Chinese naming conventions, this approach may necessitate additional context or footnotes for readers unfamiliar with these conventions. Conversely, certain translators opt for consistency in translating personal names, a method that boosts readability but may sacrifice the cultural nuances embedded in *The Analects*. The simplification of personal names in translation inevitably affects the translation of many dialogues in the original text. This practice can result in the loss of linguistic subtleties and tones that signify distinct identities within particular contexts. Such nuances run the risk of being overlooked when attempting to communicate the semantics and context of the original text.

Enhancing Comprehension of The Analects: Perspectives of Readers and Translators. As delineated in the introduction section, a significant body of scholarly work has focused on analyzing the English translations of *The Analects*. However, the majority of these studies often omit the pragmatic considerations needed to deepen readers' understanding of *The Analects*. Given the current findings, achieving a comprehensive understanding of *The Analects*' translations requires considering both readers' and translators' perspectives.

For readers, the core concepts in *The Analects* transcend the meaning of single words or phrases; they encapsulate profound cultural connotations that demand thorough and precise explanations. For instance, whether "君子 Jun Zi" is translated as "superior man," "gentleman," or otherwise. It is nearly impossible to study Confucius's thought without becoming familiar with a few core concepts (LaFleur, 2016), comprehending the meaning is a prerequisite for readers. The same principle applies to the personal names in *The Analects*. Ancient Chinese names often carry specific meanings and backgrounds. Various forms of names, such as "formal name," "style name," "nicknames," and "aliases," have deep roots in traditional Chinese culture. Whether translations adopt a simplified or literal approach, readers stand to benefit from understanding the structure and significance of ancient Chinese names prior to engaging with the text. Most proficient translators typically include detailed explanations of these core concepts and personal names either in the introductory or supplementary sections of their translations. Nevertheless, these explanations alone may not suffice for readers. If feasible, readers should consult multiple translations for cross-reference, especially when interpreting key conceptual terms and names. However, given the abundance of online resources, sourcing accurate and relevant information is convenient. Readers can refer to online resources like Wikipedia or academic databases such as the Web of Science. While this process may be time-consuming, it is an essential step towards

Table 10 A comparison of the main personal names in the five translations of *The Analects*.

Classical Chinese Name	James Legge	D. C. Lau	William Jennings	Edward Slingerland	Burton Watson
仲由 (Zhong-You)	Zhong You	Chung Yu	Tsz-lu	Zilu	Zhongyou “Zilu”
子路 (Zi-Lu)	Zilu	Tzu-lu	Yu		Zilu
由 (You)	You	Yu			You “Zilu”
季路 (Ji-Lu)	Ji Lu	Chi-lu			Zhongyou
子贡 (Zi-Gong)	Zi Gong	Tzu-kung	Tsz-kung	Zigong	Zigong
赐 (Ci)	Ci	Ssu			Si “Zigong”
冉有 (Ran-You)	Ran You	Ch’iu	Yen Yu	Ran Qiu	Ran You
冉子 (Ran-Zi)	Ran	Jan Ch’iu	Yen		Ran Qiu
冉求 (Ran-Qiu)	Ran Qiu	Jan Tzu			Qiu “Ran You”
求 (Qiu)	Qiu	Jan Yu			Qiu “Ran Qiu”
		Yu			Qiu
颜回 (Yan-Hui)	Yan Hui	Yen Hui	Yen Hwui	Yan Hui	Yan Hui
颜渊 (Yan-Yuan)	Yen Hui	Yen Yuan	Yen Yuen	Hui	Yan Yuan
回 (Hui)	Yan Yuan	Hui	Hwui		Hui
	Hui				
子张 (Zi-Zhang)	Zi Zhang	Tzu-chang	Tsz-chang	Zizhang	Zizhang
师 (Shi)	Shi	Shih			Shi
张 (Zhang)	Zhang	Chang			
子夏 (Zi-Xia)	Zi Xia	Tzu-hsia	Tsz-hia	Zixia	Zixia
商 (Shang)	Shang	Shang	You		Shang “Zixia”
					Shang
曾子(Zeng-Zi)	Philosopher Zeng	Tseng Tzu	Scholar Tsang	Master Zeng	Master Zeng
参 (Shen)	Shen	Ts’an!	Tsang Sin		Shen “Master Zeng”
	disciple Zeng	Ch’ai	Tsang Si		Shen “Zeng Shen”
			Tsang		
子游 (Zi-You)	Zi You	Tzu-yu	Tsz-yu	Ziyou	Ziyou
偃 (Yan)	Yan	Yen			
言游 (Yan-You)	Yan You	Yen Yu			

improving comprehension of *The Analects*. From readers cognitive enhancement perspective, this approach can significantly improve readers’ understanding and reading fluency, thus enhancing reading efficiency.

For translators, in the process of translating *The Analects*, it is crucial to accurately convey core conceptual terms and personal names, utilizing relevant vocabulary and providing pertinent supplementary information in the para-text. The author advocates for a compensatory approach in translating core conceptual words and personal names. This strategy enables the translator to maintain consistency with the original text while providing additional information about the meanings and backgrounds. This approach ensures simplicity and naturalness in expression, mirrors the original text as closely as possible, and maximizes comprehension and contextual impact with minimal cognitive effort.

The five translators examined in this study have effectively achieved a balance between being faithful to the original text and being easy for readers to accept by utilizing apt vocabulary and providing essential para-textual information. As English translations of *The Analects* continue to evolve, future translators can further enhance this work by summarizing and supplementing paratextual information, thereby building on the foundations established by their predecessors. By integrating insights from previous translators and leveraging paratextual information, future translators can provide more precise and comprehensive explanations of core concepts and personal names, thus enriching readers’ understanding of these terms.

Conclusion

This study employs sentence alignment to construct a parallel corpus based on five English translations of *The Analects*. Subsequently, this study applied Word2Vec, GloVe, and BERT to quantify the semantic similarities among these translations. The similarities and dissimilarities among these five translations were evaluated based on the resulting similarity scores. This study discusses the high similarity observed between translations. Our analysis reveals that while Slingerland and Watson’s translations exhibit more highly similar sentences compared to others, their robust para-texts and multiple translations of complex sentences can augment readers’ understanding while preserving the essence of the original text. The Jennings’ translation considered the readability of the text and restructured the original text, which was a very reader-friendly innovation at the time. What Jennings modified were the high-frequency, low-weight parts. Despite this structural change slightly impacting the semantic similarity with other translations, it did not significantly affect the semantic representation of the main body of *The Analects* when considering the overall data analysis.

The analysis of sentence pairs exhibiting low similarity underscores the significant influence of core conceptual words and personal names on the text’s semantic representation. The complexity inherent in core conceptual words and personal names can present challenges for readers. To bolster readers’ comprehension of *The Analects*, this study recommends an in-depth examination of both core conceptual terms and the system of personal names in ancient China. By doing so, readers can greatly improve their cognitive abilities during the reading process. Furthermore, this study advises translators to provide comprehensive paratextual interpretations of core conceptual terms and personal names to more accurately mirror the context of the original text. Such an approach can further improve the text’s comprehensibility.

This study ingeniously integrates natural language processing technology into translation research. Using quantitative research methods, it presents the semantic differences between different translations of the same work in a more refined, systematic, and in-depth manner, hoping to attract more researchers’ attention to and in-depth exploration of semantic differences between translations. The semantic similarity calculation model utilized in this study can also be applied to other types of translated texts. Translators can employ this model to compare their translations degree of similarity with previous translations, an approach that does not necessarily mandate a higher similarity to predecessors. This allows them to better realize the purpose and function of translation while assessing translation quality.

There are also some limitations in this study. For instance, the analysis only took into account three key factors: the macro-structure of translations, core vocabulary, and personal names. Considering that multiple factors influence translation differences, it would be beneficial for future studies to explore additional research perspectives. Such an approach could broaden the scope of inquiry, encompass a more diverse range of texts, and offer systematic support for readers and translators striving to enhance their comprehension and translation of texts.

Data availability

The datasets generated during and/or analysed during the current study are available in the figshare repository, https://doi.org/10.6084/m9.figshare.23931291.

Received: 8 August 2023; Accepted: 6 November 2023;
Published online: 05 January 2024

Note

[1] "text similarity" in Fig. 1 represents the computed semantic similarity between any two aligned sentences from the translations, averaged over three algorithms.

References

Ames RT, Rosemont HJ (1998) The Analects of Confucius: A Philosophical Translation. The Ballantine Publishing Group, New York

Brooks EB, Brooks AT (1998) The original analects, sayings of Confucius and his successors. Columbia University Press, New York

Chang CY et al. (2023) JCF: Joint coarse- and fine-grained similarity comparison for plagiarism detection based on NLP. Journal of Supercomputing. https://doi.org/10.1007/s11227-023-05472-0

Chesterman, A (2000). A Causal Model for Translation Studies. Intercultural Faultlines, 15–27. https://doi.org/10.4324/9781315759951-2

Chin AP (2014) Confucius, The Analects(Lunyu). Penguin Group, New York

Desmidt I (2009) Retranslation revisited. Meta 54(4):669–683. https://doi.org/10.7202/038898ar

Green C (2015) An analysis of the relationship between cohesion and clause combination in English discourse employing NLP and data mining approaches. Digital Scholarship in the Humanities 30(3):326–343. https://doi.org/10.1093/llc/fqu012

Gutierrez E, Karwowski W, Fiok K, Davahli MR, Liciaga T, Ahram T (2021) Analysis of Human Behavior by Mining Textual Data: Current Research Topics and Analytical Techniques. Symmetry 13(7):1276. https://doi.org/10.3390/sym13071276

He X, Nassar I, Kiros J, Haffari G, Norouzi M (2022) Generate, Annotate, and Learn: NLP with Synthetic Text. Transactions of the Association for Computational Linguistics 10:326–343. https://doi.org/10.1162/tacl_a_00492

Hedstrom MW (2020) Foreword. In R. K. Li, Confucius Analects (論語): A New Translation with Annotations and Commentaries. iUniverse

Hou YQ, Sun Y (2019) A Corpus-Based Comparative Analysis of Cohesive Devices in Two English Translations of The Analects of Confucius. International Journal of Languages. Literature and Linguistics 5(4):247–252. http://www.ijlll.org/vol5/236-AU0015.pdf

Iliev Y, Ilieva G (2023) A Framework for Smart Home System with Voice Control Using NLP Methods. Electronics 12(1):116. https://doi.org/10.3390/electronics12010116

Iosif E, Potamianos A (2015) Similarity computation using semantic networks created from web-harvested data. Natural Language Engineering 21(1):49–79. https://doi.org/10.1017/S1351324913000144

Jennings W (1895) The Confucian analects: A translation, with annotations and an introduction. George Routledge and Son, London & New York

Jiang YC, Bai W, Zhang XP, Hu JJ (2017) Wikipedia-based information content and semantic similarity computation. Information Processing & Management 53(1):248–265. https://doi.org/10.1016/j.ipm.2016.09.001

Koplin JJ (2023) Dual-use implications of AI text generation. Ethics and Information Technology. https://doi.org/10.1007/s10676-023-09703-z

LaFleur RA (2016) Books That Matter: The Analects of Confucius. The Great Courses, Chantilly

Lau DC (1979) The Analects. Penguin Group, London & New York

Li LY, Johnson J, Aarhus W, Shah D (2022) Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit. Computers & Education 176. https://doi.org/10.1016/j.compedu.2021.104354

Lin YT (1941) The Wisdom of Confucius. Hua Guang Book Company, Shanghai

Lin WS (2010) Getting to Know Confucius-A New Translation of The Analects. Foreign Language Press, Beijing

Liu Z (2023) A Corpus-Based Study on the Spanish Translation of 道 (dao) in The Analects. CLINA 8(2):135–161. https://doi.org/10.14201/clina202282135161

Ma L, Pahlevan Sharif S, Ray A, Khong KW (2023) Investigating the relationships between MOOC consumers' perceived quality, emotional experiences, and intention to recommend: an NLP-based approach. Online Information Review 47(3):582–603. https://doi.org/10.1108/OIR-09-2021-0482

Mikolov T, Chen K, Corrado G & Dean J (2013) Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR 2013), Scottsdale, AZ, May 2–4. https://doi.org/10.48550/arXiv.1301.3781

Ni PM (2017) Understanding the Analects of Confucius, A New Translation of Lunyu with Annotations. State University of New York Press, New York

Nissan E (2017) In the Garden and in the Ark: The belles letters, a etiological tales, and narrative explanatory trajectories-The concept of an architecture combining phono-semantic matching, and NLP story-generation. Digital Scholarship in the Humanities 32(4):859–886. https://doi.org/10.1093/llc/fqw040

Oh YK, Yi J (2022) A symmetric effect of feature level sentiment on product rating: an application of bigram natural language processing (NLP) analysis. Internet Research 32(3):1023–1040. https://doi.org/10.1108/INTR-11-2020-0649

Pennington J, Socher R & Manning C (2014) GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. Accessed Oct: 25–29. https://aclanthology.org/D14-1162.pdf

Pound E (1969) Confucius: The Unwobbling Pivot, the Great Digest, The Analects. Illustrated ed. New Directions, New York

Roberts M (2020) The Analects, Conclusions and Conversations of Confucius. University Of California Press, Oakland

Seifossadat E, Sameti H (2023) Improving semantic coverage of data-to-text generation model using dynamic memory networks. Natural Language Engineering: 1-26. https://doi.org/10.1017/S1351324923000207

Shahbazi Z, Byun YC (2022) NLP-Based Digital Forensic Analysis for Online Social Network Based on System Security. International Journal of Environmental Research and Public Health 19(12):7027. https://doi.org/10.3390/ijerph19127027

Slingerland E (2003) Analects: With selections from traditional commentaries. Hackett Publishing Company, Indianapolis

Soothill WE (1910) The Analects of Confucius. The F. H. Revell Company, Yokohama

Watson B (2007) The Analects of Confucius. Columbia University Press, New York

Waley A (1997) The Analects. Shanghai. Foreign Language Education Press, Shanghai

Yang LH (2014) A Comparative Study of the English Versions of The Analects by Legge and Ku Hungming. Theory and Practice in Language Studies 4(1):65–69. https://www.academypublication.com/issues/past/tpls/vol04/01/10.pdf

Yang LW, Zhou GJ (2022) A semantic similarity analysis of multiple English translations of the analects: Based on a natural language processing algorithm. Frontiers in Psychology. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.992890/full

Author contributions

LY: organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. GZ: gave the paper guidance and some research ideas. All authors contributed to manuscript revision, read, and approved the submitted version.

Competing interests

The authors declare no competing interests.

Ethical Approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-02355-w.

Correspondence and requests for materials should be addressed to Guijun Zhou.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.