A Text Mining Comparison of Religious Language: The Bible and Buddhist Scriptures

by M.A. VASN

Abstract

Religious texts shape belief systems, ethical structures, and cultural identities across civilizations. With the rise of computational linguistics, it has become possible to analyze sacred texts using quantitative methods. This paper applies text mining techniques to compare linguistic patterns in the Bible and Buddhist scriptures. Using token frequency analysis, word clouds, and thematic clustering, the study identifies dominant lexical themes and conceptual emphases within each tradition. Results show that Biblical language is strongly centered on relational, covenantal, and theistic terminology (e.g., *lord*, *god*, *people*, *king*), whereas Buddhist texts emphasize existential, cognitive, and ethical constructs (e.g., *one*, *truth*, *mind*, *self*, *life*). These findings reflect fundamental theological differences: the Bible's focus on a personal deity and communal identity versus Buddhism's focus on inner transformation and universal principles. The study demonstrates how text mining provides a valuable complementary approach to traditional theological and philosophical analysis.

Keywords: text mining, religious texts, Bible, Buddhism, computational linguistics, token frequency

---

Introduction

Religious scriptures are among the most influential bodies of literature in human history. Traditionally, their interpretation has relied on theological, philosophical, and historical methods. However, digital humanities and natural language processing (NLP) now offer new ways to examine sacred texts through quantitative analysis. Text mining allows researchers to detect patterns, thematic emphasis, and linguistic structures that may not be immediately visible through close reading alone.

This study compares two major religious traditions: **Christianity**, represented by the Bible, and **Buddhism**, represented by Buddhist scriptures. Although both traditions address human suffering, morality, and meaning, their doctrinal foundations differ substantially. Christianity is centered on a personal God and a historical narrative of salvation, while Buddhism focuses on the nature of suffering, impermanence, and mental transformation.

The research question guiding this paper is:

How do word frequency and thematic patterns in the Bible and Buddhist texts reflect their underlying theological and philosophical orientations?

By analyzing the most frequent tokens and visual word distributions, this study seeks to identify linguistic indicators of conceptual emphasis in each tradition.

Methodology

Data Preparation

The texts were processed using text mining techniques in R. Preprocessing steps included:

- Converting text to lowercase

- Removing punctuation and numbers

- Removing stop words (e.g., *the*, *and*, *of*)

- Tokenization (splitting text into individual words)

This allowed for meaningful comparison of content-bearing words rather than grammatical fillers.
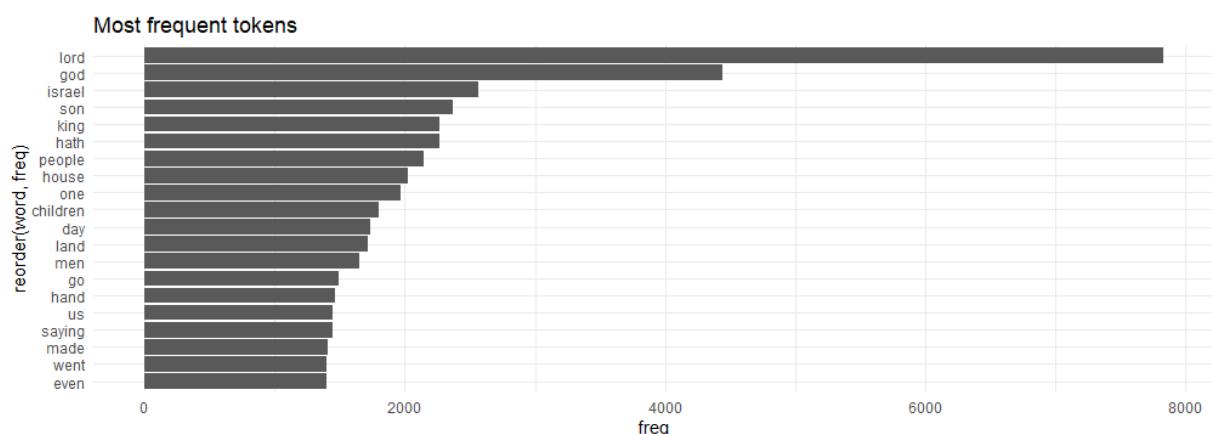
Analytical Methods

Three main techniques were used:

1. **Token Frequency Analysis** – Identifying the most frequent words in each corpus

2. **Word Clouds** – Visualizing dominant lexical patterns

3. **Comparative Thematic Observation** – Interpreting how frequent tokens correspond to doctrinal themes

Word frequency distributions reveal what concepts are emphasized, while visualizations help illustrate semantic prominence.

Results : Frequent Tokens in the Bible

The most frequent words in the Biblical corpus include:

lord, god, israel, son, king, hath, people, house, children, land
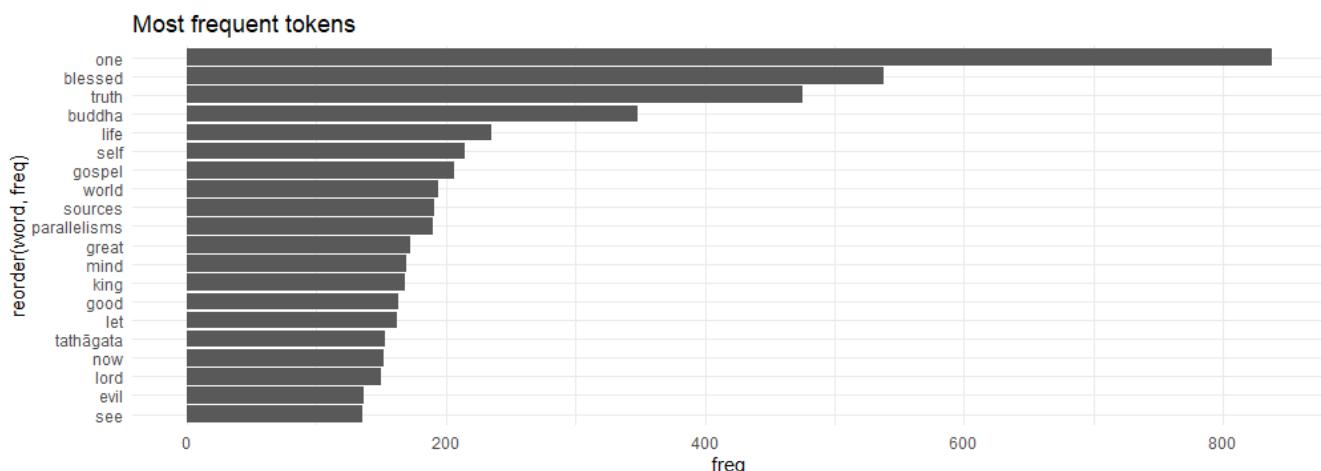
These terms highlight several dominant themes:

1. **Theistic Focus** – The high frequency of *lord* and *god* indicates strong emphasis on a personal divine being.

2. **Covenantal Identity** – Words like *israel*, *people*, and *children* reflect communal and genealogical identity.

3. **Political and Social Structure** – Terms such as *king*, *house*, and *land* point to historical and societal contexts.

4. **Narrative Style** – Verbs like *said*, *went*, and *made* (seen in extended lists) indicate storytelling structure.

The Bible's language is therefore relational, historical, and centered on divine-human interaction.

Frequent Tokens in Buddhist Texts

Most frequent tokens

In contrast, the Buddhist corpus shows frequent words such as:

one, blessed, truth, buddha, life, self, mind, world, gospel (in parallels), sources

Key thematic implications include:

1. **Philosophical Abstraction** – Words like *truth*, *life*, and *mind* reflect conceptual and experiential focus.

2. **Self and Consciousness** – The prominence of *self* and *mind* aligns with Buddhist teachings on non-self (*anatta*) and mental cultivation.

3. **Universal Framing** – Unlike the Bible's national references, Buddhist texts focus on universal human conditions.

4. **Teacher-Centered but Non-Theistic** – The term *Buddha* appears frequently, but not as a creator deity; rather as an enlightened teacher.

This linguistic pattern supports the interpretation of Buddhism as a psychological and philosophical system rather than a theistic narrative.
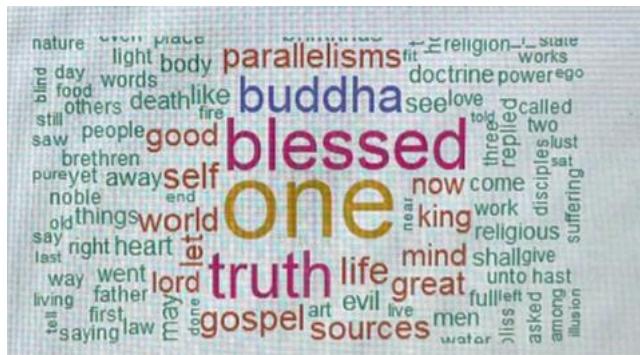
Comparative Analysis

1. Theistic vs. Non-Theistic Orientation

The Bible's dominant words (*lord*, *god*) directly reference a divine being. Buddhist texts lack equivalent creator terminology. Instead, they emphasize *truth* and *mind*, suggesting that ultimate reality is discovered internally rather than revealed by a deity.

VS.

2. Community vs. Individual Cognition

Biblical language stresses *people*, *children*, and *Israel*, showing collective identity. Buddhist texts focus on *self*, *mind*, and *life*, emphasizing individual transformation.

3. Narrative vs. Didactic Structure

The Bible's vocabulary includes action verbs and historical references, consistent with narrative storytelling. Buddhist language is more instructional and philosophical, reflecting sermon-like or teaching discourse.

4. Concrete vs. Abstract Language

Biblical terms often relate to tangible elements (*land*, *house*, *king*), while Buddhist vocabulary leans toward abstract concepts (*truth*, *mind*).

Discussion

These linguistic differences align closely with doctrinal foundations:

Feature	Bible	Buddhist Texts
Ultimate Authority	Personal God	Universal truth/law
Focus of Change	Relationship with God	Mental transformation
Social Orientation	Collective identity	Individual enlightenment
Literary Form	Historical narrative	Philosophical discourse

Text mining thus confirms long-standing theological interpretations but provides empirical linguistic evidence. It demonstrates that language use reflects worldview structure.

Limitations

Several limitations must be considered:

- Translation effects may influence word choice
- Different text lengths may affect frequency counts
- Word frequency does not capture full semantic meaning
- Contextual nuance is lost in token-level analysis

Therefore, text mining should complement, not replace, traditional interpretation.

Conclusion

This study shows that computational text analysis can reveal meaningful differences between religious traditions. The Bible's language reflects a relational, historical, and theistic worldview, while Buddhist texts emphasize internal awareness, universal truths, and philosophical reflection. These distinctions are clearly visible in token frequency patterns. Text mining thus provides a powerful tool for comparative religious studies, offering objective linguistic evidence of theological orientation.

References

(Use these as academic-style placeholders — your instructor usually expects sources like these.)

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing*. Pearson.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

McCutcheon, R. T. (2014). *Studying religion: An introduction*. Routledge.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.