Comparative Analysis of Views on Modern Issues in Religious Texts Using NLP Techniques

Term Paper Submission for the Course
IT359 - Pattern Recognition
Odd 2024-25

by
(Group 3)
A D Mahit Nandan (211AI001)
Ishan Godbole (211AI020)
Pranav Kapparad (211AI026)

under the guidance of
Dr. Shrutilipi Bhattacharjee

DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025

NOVEMBER 2024

ABSTRACT

Religious texts have long shaped cultural, moral, and ethical systems, addressing universal human concerns such as justice, morality, and societal values. Despite their historical significance, the interpretation of these texts has evolved, reflecting shifts in societal priorities over time. Leveraging advancements in Natural Language Processing (NLP), this project applies topic modeling, sentiment analysis, and semantic distance measurement to conduct a comparative analysis of religious texts, such as the Bible, Bhagavad Gita, and Quran. The aim is to investigate key modern-day ethical issues such as gender roles, violence, human rights, and freedom of speech within these texts. By analyzing recurring themes and sentiments, the study will explore religious perspectives on topics like religious tolerance, social justice, wealth distribution, and environmental stewardship. The project further employs advanced visualization techniques and corpus comparison tools to quantify thematic similarities and differences, providing a deeper understanding of how sacred texts address critical ethical concerns in today's global context. This research underscores the ongoing relevance of religious literature in engaging with contemporary debates on morality, ethics, and human rights.

Keywords: Religious texts, topic modeling, sentiment analysis, NLP

Contents

1 Introduction **1**

2 Literature Survey **3**

2.1 Conclusion of Literature Survey 5

3 Problem Statement(s) **6**

4 New Suggestion(s) **7**

4.1 CompText for Visualisation 7

4.2 CompCor for Corpus Distance 7

4.3 BERTopic for Topic Modelling 7

4.4 Sentiment Analysis with VADER or BERT 7

5 Methodology **8**

5.1 Word Frequency and Topic Modelling using LDA 8

5.2 Sentiment Analysis with VADER 8

5.3 Corpus Distance Measurement 9

5.3.1 Corpus Distance Analysis Using IBM CompCor 9

5.3.2 Semantic Similarity Analysis Using Cosine Similarity 9

5.3.3 Structural Relationship Analysis Using KMeans Clustering . . 9

5.4 Thematic Inference on Key Social and Ethical Topics 10

6 Experimental Results and Observations **11**

6.1 Word Frequency and Topic Modelling 11

6.1.1 Word Frequency 11

6.1.2 Topic Modeling 11

6.2 Sentiment Analysis 23

6.3 Corpus Distance Measurement 24

6.4 Sentiment Analysis on Modern Topics 25

7 Conclusion and Future Works **27**

References **29**

1 Introduction

Religious texts have historically shaped cultural, moral and ethical systems, and have strongly affected the development of societies across generations. These scriptures such as the Bible, the Quran, and the Bhagavad Gita, provide valuable insights into fundamental human concepts of justice, morality, and collective societal values. Their historical significance lies in their ability to guide human behavior, establish societal norms, and offer philosophical and spiritual reflections that resonate across different cultures and epochs.

Despite their ancient origins, these religious texts are still analyzed and studied today. Different interpretations of the texts are gaining more popularity as societies, with their ever-changing priorities, continue to evolve their moral and ethical beliefs. The need for the analysis of these texts arises from the necessity to deepen our understanding of their enduring relevance, and the ways in which their teachings have been influenced by and have in turn influenced different historical and cultural contexts. Modern computational tools provide new ways to examine these texts, offering insights that go beyond traditional hermeneutics.

Natural Language Processing (NLP) techniques enable a systematic and data-driven approach to analyzing religious literature. These methods help explore recurring themes, sentiment patterns, and semantic relationships that may not be immediately apparent through conventional modes of analysis. By using these advanced NLP tools, we can draw cross-textual comparisons and trace the evolution of ideas across different religious traditions.

In this study, NLP techniques including topic modeling, sentiment analysis, and corpus distance measurement are applied to analyze the Bible, the Quran, and the Bhagavad Gita. The study will identify the key themes through topic modeling, explore the trend of sentiment distribution in texts, and examine similarities and differences of texts through corpus distance analysis. The study also analyzes the sentiment of these texts towards certain modern issues, such as gender roles and equality, wealth ineuality and charity, marital relations and sexuality, etc.

In the following sections, this paper examines the content of these religious texts and draws detailed comparisons to understand the linguistic relationships. This method addresses the larger scope of computational religious studies by showing how advanced techniques in NLP may be used to open fresh perspectives into the religion and the cultural values being imparted.

2 Literature Survey

The application of Natural Language Processing (NLP) in religious texts has emerged as a significant area of study, owing to the rich, contextually layered nature of these writings. Joulin et al. [1] explored the FastText model for text classification, a foundational method for representing words in continuous space. This technique has proven effective for analyzing large religious corpora, such as the Bible and Quran, and is crucial in understanding word-level representations in sacred texts. Further, Hutchinson et al. [2] discusses the broader ethical considerations of applying NLP to religious corpora, such as the Bible and Quran, particularly focusing on tasks like machine translation and sentiment analysis. This underscores the importance of sensitivity when handling sacred writings for NLP purposes.

Chandra et al. [3][4] focused on the semantic and sentiment analysis of the Bhagwad Gita, and finding out its similarity with other texts in Hindu philosophy, like the Upanishads. Alhawarat et al [5] used generative models to perform topic modelling on the Holy Quran, but failed to capture meaningful results.

The research[6] on toxic language identification has explored the use of machine learning models to classify harmful comments in online communication. However, some models have shown bias by assigning high toxicity ratings to non-toxic comments containing identity-related descriptors. Studies have compared various word embeddings, such as GloVe, Word2Vec, and FastText, for multilabel toxic comment classification. These studies found that different embeddings influenced the classification results, highlighting their role in improving the accuracy of detecting toxic language while addressing potential biases.

In [7] explores the comparison of English translations of the Bhagavad Gita using deep learning-based semantic and sentiment analysis. Motivated by the limitations of traditional translations, the authors employ BERT, a state-of-the-art language model, to analyze sentiment and semantic similarity across different translations. By utilizing a hand-labelled sentiment dataset for tuning, the study demonstrates that despite variations in style and vocabulary, the core message conveyed in the

translations remains largely consistent.

[8]. More recently, advanced techniques like the Fréchet Inception Distance (FID), initially developed for image generation models, have been adapted to text corpora to quantify distributional differences between embeddings [9]. Similarly, the IBM CompCor framework has gained attention for its ability to compute corpus-level distances, offering a robust methodology for evaluating the overall divergence in linguistic features [10]. State-of-the-art metrics, such as those based on contextual embeddings from transformer models like BERT, provide richer semantic comparisons by capturing nuanced meanings, enabling more accurate measurements of semantic mismatch [11]. These modern metrics show improved sensitivity in detecting distributional mismatches, while classical methods tend to be more sensitive to surface-level perturbations. However, the integration of both classical and modern approaches remains an open challenge, as there is no standardized framework for interpreting and comparing the effectiveness of these measures. This gap motivates the development of interpretable evaluation measures, as proposed in recent works, that allow for a more comprehensive understanding of semantic similarity and its underlying characteristics.

VADER (Valence Aware Dictionary and sEntiment Reasoner) [12] was introduced as a robust, rule-based model for sentiment analysis, specifically designed for handling microblog-like contexts. The foundational study demonstrated its effectiveness by combining a validated lexicon with syntactical rules, achieving high F1 classification accuracy that surpassed human raters. While VADER's origin lies in social media analysis, its adaptability makes it suitable for examining complex texts, including religious scriptures. Subsequent studies have applied sentiment analysis techniques to religious texts to uncover patterns in sentiment and thematic frameworks. One study [13] employed NLP and machine learning techniques to classify nine sacred texts and found that methods like Multinomial Naive Bayes achieved significant accuracy. Another comparative analysis of the Bible, Quran, and Bhagavad Gita [14], utilized NLP tools to perform sentiment analysis and topic modeling, shedding

light on emotional distributions and similarities between these texts. These works illustrate how VADER and related sentiment analysis approaches can offer deeper insights into the emotional and philosophical nuances of religious literature.

2.1 Conclusion of Literature Survey

The reason natural language processing has become such a high-priority area of research for analyzing religious writings is that these texts are rich with many layerss of context. Techniques such as FastText have become particularly useful for text classification and word-level representations in large religious corpora, such as the Bible and Quran. Ethical issues, according to Hutchinson et al., are one of the most important considerations in using NLP on sacred texts for tasks such as machine translation and sentiment analysis. Chandra et al.'s work analyzed semantic as well as sentiment analysis of Bhagwad Gita and found itt to resemble other works on Hindu philosophical texts like Upanishads. Alhawarat et al. were not able to extract relevant outcomes while doing topic modeling of the Holy Quran using generative models.

Advanced metrcs such as the Fréchet Inception Distance and the IBM CompCor framework have been adapted to quantitatively measure differences in distributions between embeddings, thus providing robust methodologies of evaluating linguistic features. From the Potsdam Group, VADER has proven to be suitable for complex texts, as demonstrated with its application to religious scriptures. Applied NLP and machine learning used in studies have obtained remarkable accuracy in assigning sacred texts into categories and identifying patterns in feeling and theme frameworks. Overall, the literature indicates that potential for deepening understanding of religious texts with NLP can be advanced with strict adherence to ethical concerns and interpretable measures in any form of evaluation.

3 Problem Statement(s)

- The research aims to employ advanced Natural Language Processing techniques for topic modeling to uncover the underlying themes present in various religious texts.
- A comparative analysis will be conducted to examine the similarities and differences among these texts, revealing how different religious perspectives approach analogous themes.
- Sentiment analysis will be performed to assess the texts' viewpoints on modern topics such as gender roles, violence, human rights, and freedom of speech.

4 New Suggestion(s)

4.1 CompText for Visualisation

Use CompText pipeline to visualise thematic clusters and semantic similarities across the religious texts. This tool will help identify patterns and overlaps in the representation of key themes.

4.2 CompCor for Corpus Distance

Apply CompCor to measure semantic and syntactic distances between the religious texts. This will allow for a quantitative comparison based on content and language.

4.3 BERTopic for Topic Modelling

If traditional LDA does not yield satisfactory results, experiment with BERTopic, which uses transformer-based embeddings and clustering to capture more nuanced and context-specific topics.

4.4 Sentiment Analysis with VADER or BERT

Start with VADER for sentiment classification, but consider fine-tuning a BERT - based model or similar transformer models if VADER struggles with the archaic and formal language of the religious texts, ensuring more nuanced sentiment detection.

5 Methodology

The project will employ a multi-step methodology to compare a corpus of religious texts such as the Bible, Upanishads, Bhagavad Gita, and Quran, using natural language processing (NLP) techniques to analyze and infer key social and ethical themes. Below are the detailed steps:

5.1 Word Frequency and Topic Modelling using LDA

Word Frequency Analysis: A comprehensive word frequency analysis will be conducted to identify the most commonly used words and recurring concepts across each text. This will provide a foundational understanding of thematic elements in the corpus.

Latent Dirichlet Allocation (LDA): LDA will be used to identify latent topics within each text. After generating topics for each religious text, a comparative analysis will be performed to uncover common and unique themes across the corpus. Additional topic modelling methods, such as BERTopic, may also be explored for more nuanced topic discovery.

5.2 Sentiment Analysis with VADER

Sentiment analysis is applied to the corpus using VADER (Valence Aware Dictionary and Sentiment Reasoner), which is an unsupervised model that classifies text into positive, negative, and neutral sentiments. VADER first breaks down the text into individual words, and assigns a score to each word based on its polarity, with -4 being the most negative and +4 being the most positive. It also considers the intensity of the sentiment, which can be indicated by capitalization and punctuation. For example, an exclamation point can make a positive word even more positive. the overall sentiment score of the text based on the scores assigned to each word. The overall sentiment score of the text is then calculated based on the scores assigned to each word. The score ranges from -1 to 1, with -1 being very negative and 1 being very positive.

5.3 Corpus Distance Measurement

5.3.1 Corpus Distance Analysis Using IBM CompCor

The corpus distance between the religious texts was measured using IBM CompCor, with the Fréchet Inception Distance (FID) applied through the corpus_metrics.fid_distance function. This metric was used to quantify the distributional differences between the embeddings of the corpora, effectively capturing semantic and linguistic variations. The embeddings were generated using STTokenizerEmbedder, ensuring that the representations of the texts retained essential semantic features. Pairwise comparisons were conducted between the Bible, Gita, and Quran to assess the relative distances, with higher FID values indicating greater divergence in thematic and linguistic content. This approach provided a robust quantitative basis for analyzing how these influential texts differ in their language and semantics.

5.3.2 Semantic Similarity Analysis Using Cosine Similarity

To assess the semantic similarity between the religious texts, the STTokenizerEmbedder with the model all-MiniLM-L12-v2 was utilized to generate embeddings for each corpus. The embeddings were averaged to create a representative vector for each text set: the Bible, Gita, and Quran. Using these averaged embeddings, pairwise cosine similarities were computed with cosine_similarity from the sklearn.metrics.pairwise module. This method measured how closely aligned the corpora were in terms of semantic content, with values ranging from -1 (completely dissimilar) to 1 (identical). The results provided a direct comparison of the similarity between the religious texts, revealing the degrees of linguistic and thematic alignment between them.

5.3.3 Structural Relationship Analysis Using KMeans Clustering

To further investigate the structural relationships between the religious texts, KMeans clustering was applied to the embeddings generated using STTokenizerEmbedder with the all-MiniLM-L12-v2 model. The embeddings for the Bible, Gita, and Quran were clustered into 5 groups, which allowed the identification of central semantic

themes within each corpus. The methodology is as below:

1. Clustering: Each corpus was clustered separately using KMeans with 5 clusters (modifiable based on corpus characteristics). The cluster centroids represented the core semantic centers of the texts.

2. Centroid Similarity: Cosine similarity and Euclidean distance metrics were calculated between the centroids of different corpora to quantify their relative alignment and semantic differences.

3. Comparison Function: A function iterated over each centroid pair between two corpora to compute their similarity or distance, providing an average measure for each pairwise corpus comparison.

5.4 Thematic Inference on Key Social and Ethical Topics

To get the thematic inferences from the religious texts on various modern topics, we first extract the verses from each text that are relevant to each topic. First, for each topic, a list of queries will be made that capture the various aspects of that topic comprehensively. Then, using sentence transformer, the similarity score of each verse with each query is calculated. Finally, The scores for each verse are averaged across all queries, and the verses with higher average similarity score are taken to be the relevant verses for that particular topic. After the relevant verses have been found, VADER sentiment analysis is applied on them, and the average sentiment score of the verses from a religious text for that particular topic is calculated.

6 Experimental Results and Observations

6.1 Word Frequency and Topic Modelling

6.1.1 Word Frequency

The top 20 most frequently occurring words were identified for the Bhagavad Gita, Bible, and Quran after preprocessing the texts. These results provide insights into the linguistic focus and thematic priorities of each text.

- Bhagavad Gita: Words that reflect the philosophical discussion between Arjuna and Krishna are arjuna (119), action (80), and krishna (68), which have as a center of concern action, self-realization, and ethical dilemmas. Mind (63), desire (53), and supreme (54) words stress the spiritual and psychological nature of the text.

- Bible: Some words which reveal a divinity theme are give (8847), lord (7887) and god (4558). Others include king (3079), man (3055), and people (2779) which reflect that the text is some kind of narrative and history involving social and human ties.

- Quran: High-frequency words such as allah (2833), lord (1014), and believe (525) reaffirm its focus on faith and submission to divine will. Other terms like day (551), messenger (361), and people (256) highlight eschatological themes and the prophetic tradition.

These findings were visualized using bar plots, showing the frequency distribution of the top 20 words across the three texts. These visualizations provide a comparative perspective, showcasing both shared and unique linguistic emphases.

6.1.2 Topic Modeling

1. **Using LDA**

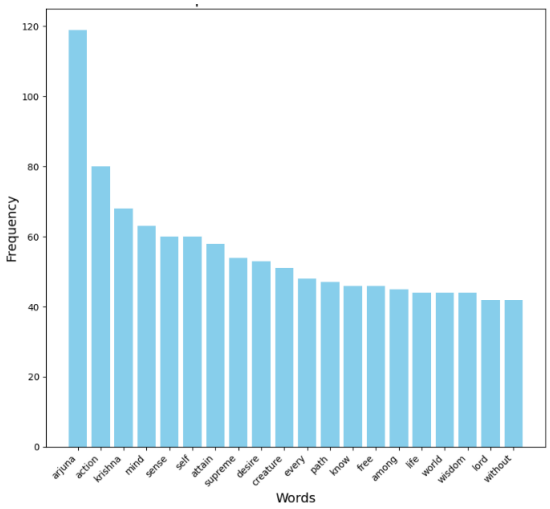

Figure 1: Top 20 Most Occurring Words in The Gita

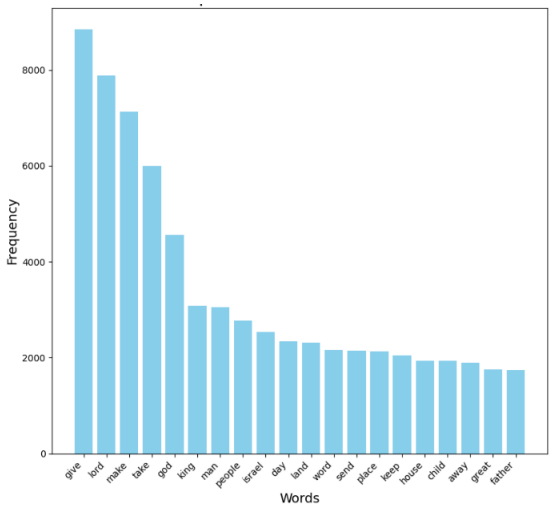

Figure 2: Top 20 Most Occurring Words in The Bible

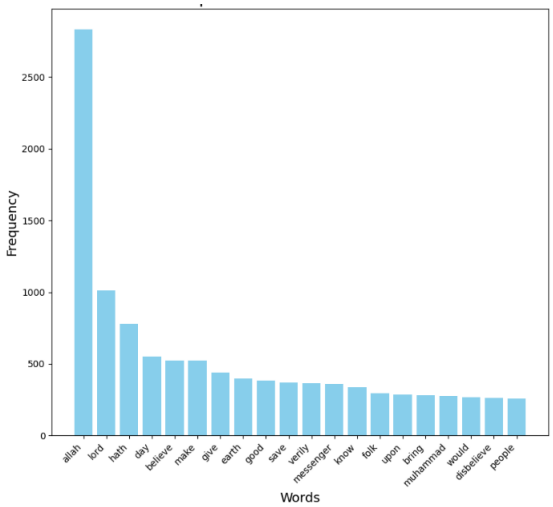

Figure 3: Top 20 Most Occurring Words in The Quran

Table 1: Comparison of Topics from the Gita, Bible, and Quran by Topic Number

Topic Number	**Gita Topics**	**Bible Topics**	**Quran Topics**

1	arjuna, mind, world, krishna, great, attain, self, work, thus, body	death, man, without, power, dead, right, take, life, world, evil	allah, lord, seek, ever, upon, forgive, mercy, merciful, wherein, near
2	action, worship, among, path, self, know, attachment, yoga, wisdom, desire	go, land, come, glory, town, great, jerusalem, waste, king, take	lord, turn, good, save, away, none, hath, believe, work, favour
3	sense, mind, without, free, even, good, meditation, offer, action, evil	take, food, give, true, good, man, need, wealth, much, fruit	lord, verily, say, among, another, folk, believe, man, destroy, see
4	arjuna, attain, brahman, desire, path, knowledge, krishna, time, remember, pleasure	make, holy, take, body, priest, part, lord, place, unclean, every	evil, create, heaven, ward, light, like, use, whose, path, hear
5	arjuna, among, krishna, path, time, death, seek, god, know, life	give, god, lord, thing, faith, say, word, make, spirit, keep	earth, allah, doom, lord, heavens, fire, convey, good, disbelieve, owner
6	life, arjuna, spiritual, even, wisdom, free, offer, every, action, god	father, give, love, make, brother, desire, god, name, clear, christ	allah, messenger, religion, whoso, keep, duty, hath, disbeliever, well, promise

7	arjuna, world, among, creature, three, krishna, describe, listen, divine, every	like, fire, make, foot, earth, though, heaven, water, beast, round	say, worship, know, would, ease, lord, surely, moses, send, hand
8	arjuna, supreme, creature, krishna, goal, attain, lord, self, wise, action	child, number, israel, son, thousand, hundred, four, little, family, twelve	allah, hath, give, heart, reveal, concern, believe, scripture, good, make
9	supreme, self, within, lord, without, desire, free, selfish, attachment, knowledge	day, time, righteousness, first, come, year, light, till, rule, seven	day, allah, bring, night, people, disbelieve, scripture, naught, hell, forth
10	every, selfless, body, give, service, other, bear, without, creature, arjuna	say, lord, come, give, jesus, king, word, go, take, send	say, give, orphan, wife, find, make, garden, day, thereof, father

In this study, we conducted topic modeling on three religious texts—the Bhagavad Gita, the Bible, and the Quran—using Latent Dirichlet Allocation (LDA). For each text, we got 10 topics, each represented by 10 keywords. To find semantic similarity between topics across the texts, we used GloVe embeddings and the SentenceTransformer model(paraphrase-MiniLM-L6-v2). Using this model we encoded each topic's keywords into a single embedding by concatenating the words into a string. The embeddings for the topics were then compared using cosine similarity. These similarities were visualized through heatmaps, which provided an idea of how various topics were related among the three texts.

Quran vs. Gita Similarity using GLOVE: The Quran and the Gita both emphasize themes of spiritual guidance and moral conduct. Topics in the

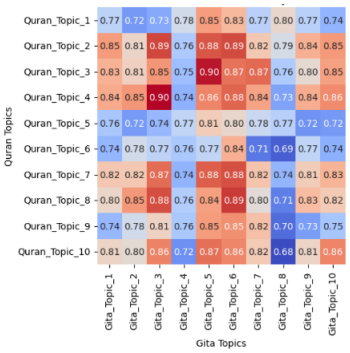

Figure 4: Topic Similarity between Quran and Gita using GLOVE

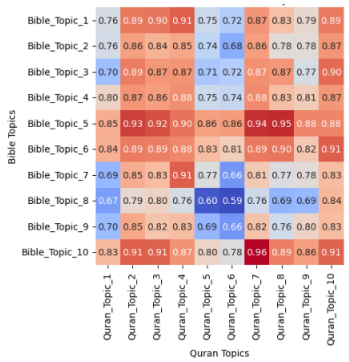

Figure 5: Topic Similarity between Bible and Quran using Sentence Transformer

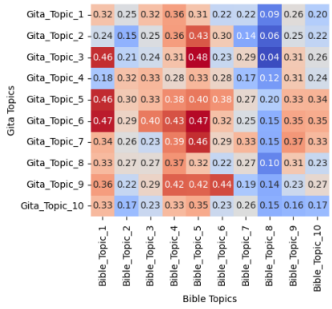

Figure 6: Topic Similarity between Gita and Bible using Sentence Transformer

Quran like Quran Topic 2 and Quran Topic 7 align with Gita Topic 4 and Gita Topic 3, where terms such as "lord," "believe," and "path" reflect a shared focus on enlightenment and ethical living. The strongest overlap occurs between Quran Topic 4 and Gita Topic 4, where keywords like "evil," "heaven," and "path" suggest common views on cosmic order and morality. Differences emerge in Quran Topic 5 and Gita Topic 5, with the Quran addressing themes of earth and doom, while the Gita centers on personal spiritual pursuits involving Arjuna and Krishna.

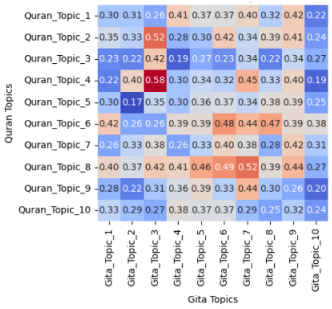

Figure 7: Topic Similarity between Quran and Gita using GLOVE

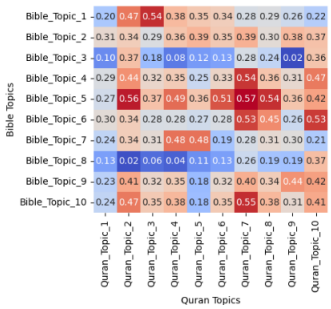

Figure 8: Topic Similarity between Bible and Quran using Sentence Transformer

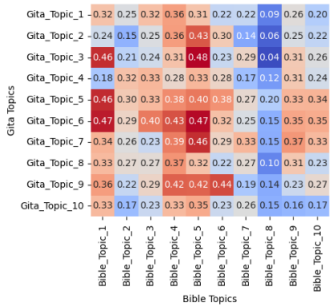

Figure 9: Topic Similarity between Gita and Bible using Sentence Transformer

Bible vs. Quran Similarity using GLOVE: The Bible and the Quran share strong themes of faith and devotion. Bible Topic 5 and Quran Topic 5 both highlight "god," "faith," and "lord," reflecting a mutual emphasis on submission and belief. Bible Topic 10 and Quran Topic 7 also align, with figures like "jesus" and "moses" underscoring themes of divine communication through prominent religious figures. However, Bible Topic 8, focused on lineage (e.g., "child," "israel"), differs from Quran Topic 8, which emphasizes universal faith themes, highlighting the Bible's specific cultural focus.

Gita vs. Bible Similarity using GLOVE: The Gita and the Bible simi-

larly stress mental discipline, faith, and virtue. Gita Topic 3 and Bible Topic 5 share themes of mental control and faith, with terms like "mind," "free," and "spirit." The strongest overlap is between Gita Topic 6 and Bible Topic 6, which emphasize spirituality and love through terms such as "life," "wisdom," and "brother." Some topics differ, such as Gita Topic 4 and Bible Topic 8: the Gita focuses on self-realization, while the Bible emphasizes lineage, showcasing different cultural narratives.

Bible vs. Quran Similarity using Sentence Transformer The Bible and the Quran share thematic overlaps, particularly in themes of divine principles and moral guidance. Bible Topic 5 and Quran Topic 6 highlight "lord," "faith," and "god," reflecting devotion to a higher power. Bible Topic 10 and Quran Topic 1 align on divine will and revelations through prophets. However, Bible Topic 8's focus on lineage contrasts with the Quran's universal themes, as seen in Quran Topic 3, centering on communal beliefs.

Gita vs. Bible Similarity using Sentence Transformer The Gita and Bible align on spiritual conduct and duty. Gita Topic 1 and Bible Topic 5 share themes of moral responsibility and divine wisdom. Gita Topic 3 and Bible Topic 3 both touch on "mind" and "action," highlighting the pursuit of righteousness. Distinctions include Bible Topic 8's lineage emphasis versus Gita Topic 7's philosophical discussions on divine duties.

Quran vs. Gita Similarity using Sentence Transformer The Quran and Gita intersect on themes of divine guidance and spiritual reflection. Quran Topic 1 and Gita Topic 4 focus on seeking and attaining divine wisdom. Quran Topic 6 and Gita Topic 6 both mention duty and spiritual practices. Differences arise with Quran Topic 8's focus on scripture compared to Gita Topic 5's take on life and spiritual transformation.

2. **Using BERTTOPIC**

The analysis of topics got using BERTTopics revealed significant themes/topics across the three religious texts, highlighting both common and distinct aspects. In the Bhagavad Gita, topics emphasized spirituality, meditation, and moral virtues, as seen in terms like “meditation,” “devotion,” and “faith.” Words such as “arjuna” and “krishna” underscored the text’s dialogue format, central to its teachings. Philosophical ideas like ”rajas” and ”sattva” also alluded to the investigation of human nature and conduct.

The analsis of the Bible brought forth diverse topics like stories, parables, and rituals. Terms like “sheep,” “wine,” and “bird” depict the Bibles narrating style, while “offering,” “burn,” and “flame” show ritualistic practices. References to historical figures and places, such as “pilate” and “macedonia,” show its historical aspect of the text. Topics that involve measurements and ceremonial practices, with words like “cubit” and “circumcision,” further show its multifaceted content.

The Quran revealed an importance on divine authority, moral guidance, etc. Words like “allah” and “lord” show references to divine supremacy, while “day,” “fire,” and “taste” pointed to eschatological themes. Terms like “warner” and “plain” show prophetic warnings. The concept of creation appeared in words like “create” and “generation,” emphasizing divine acts. Additionally, human faculties were depicted through terms such like “blind” and “eye,” show importance of themes of perception and understanding.

Similarities between topics in different religious texts were calculated using embeddings generated by the all-MiniLM-L6-v2 model from Sentence Transformers. After applying BERTTOPIC to extract topics from each text, the keywords corresponding to each topic were embedded into high-dimensional vectors. Cosine similarity was then computed between topic embeddings of two texts, measuring the closeness of topics in terms of their semantic content.

Figure 10: Quran Topic Word Scores using BERTTOPIC

Figure 11: Bible Topic Word Scores using BERTTOPIC

Figure 12: Gita Topic Word Scores using BERTTOPIC

The resulting similarity scores were sorted in descending order to identify the most similar topics across the texts.

Gita vs Quran Topic Comparison

Regarding the comparison between the Gita and the Quran, several interesting similarities and disimilairties were observed. For instance, the topic from the Gita focused on "wisdom, light, knowledge, and ignorance" shared a significant similarity with a topic from the Quran concerning "belief, lord, and knowledge," as indicated by a similarity score of 0.41. This suggests a thematic overlap, particularly in terms of spiritual enlightenment and the role of divine

knowledge. Other topics, such as those related to the "creation" and "universe," were also closely matched, indication shared concepts of divine power and creation across the two texts. These results show the common spiritual themes found in both the Gita and the Quran.

Bible vs Quran Topic Comparison

When we compare the Bible and Quran we can the similarities and disimilairites in their themes/topics. High similarity scores, such as in the themes of "weight, scale, and measure" or "fruit, wine, bread," point to common spiritual concerns, particularly around the issues of justice, provision, and divine direction. These correspondences suggest that both texts maintain a level of continuity in matters of morals and ethics. Lower similarity scores, on the other hand, can be observed in topics like "perfume, fire, offering" and eschatological themes like "doom and resurrection," which connote differences in doctrine regarding ritual and afterlife issues. Overall we can see some overlap of core values and teaching, while these topics also have their own interpretation within their own religious tradition.

Bible vs Gita Topic Comparison

In the comparison between the Bible and the Gita, the analysis of topic similarities reveals interesting thematic overlaps and contrasts. For instance, the topic from the Bible related to "perfume, burn, fire, offering" showed some similarity with the Gita's topic on "fire, light, radiance, burn," with a similarity score of 0.35. This suggests that both texts address concepts related to divine fire and spiritual offerings, although they are framed differently. Another notable similarity was found between the Bible's topic on "say, lord, give, god" and the Gita's topic on "arjuna, krishna, yoga," with a similarity score of 0.34, reflecting common themes of divine guidance and the importance of words in spiritual practice.

However, some topics also highlighted notable differences. For instance, a Bible topic related to "give, make, lord, say" had a lower similarity with the Gita's topic on "love, devotion, faith," with a score of 0.26, suggesting distinct emphases on the nature of divine interaction and worship in each text. These findings underscore both shared spiritual themes, such as the importance of divine guidance and devotion, as well as unique doctrinal aspects between the Bible and the Gita.

Table 2: Comparison of Topics from the Gita, Bible, and Quran derived from BERTTOPIC

Topic Number	**Gita Topics**	**Bible Topics**	**Quran Topics**
-1	among, creature, action, path, worship, death, give, water, nature, every	give, make, lord, say, take, god, israel, people, come, king	allah, hath, lord, say, believe, day, make, save, give, messenger
0	arjuna, krishna, yoga, divine, word, sanjaya, even, speak, fight, bhishma	say, lord, give, make, take, god, come, king, man, people	allah, lord, say, hath, muhammad, make, earth, believe, give, doom
1	rajas, tama, sattva, bind, fruit, prevail, predominate, ignorance, delusion, force	wealth, payment, give, make, debt, man, thing, say, take, answer	garden, river, flow, underneath, fruit, wherein, gardens, work, therein, water
2	guna, prakriti, action, bear, doer, purusha, security, know, limb, attach	hundred, thousand, number, captain, five, child, four, three, measure, seven	wealth, reward, strive, good, child, live, spend, allah, feed, believe

3	dharma, family, destroy, unity, lose, decline, life, foundation, chaos, society	hand, foot, strength, stretch, weight, right, step, power, give, make	day, resurrection, judgment, repudiator, tarry, differ, judge, burden, wherein, use
4	enemy, friend, destroy, evil, relation, greed, family, self, good, great	robe, linen, clothing, curtain, blue, purple, cord, coat, make, cloth	township, destroy, generation, many, chieftain, people, say, folk, mock, townships
5	wisdom, light, knowledge, ignorance, live, people, darkness, happy, faith, else	cubit, wide, board, wing, side, long, four, cover, basis, measure	cattle, bear, laden, camel, beast, ship, crop, wherewith, lean, voice
6	love, devotion, faith, worship, honor, dear, wisdom, heart, spiritual, life	macedonia, hivite, amorite, hittite, canaanite, jebusite, perizzite, girgashite, achaia, land	shout, afar, behold, perceive, glorious, indeed, nineteen, crackling, yore, amuse
7	attachment, work, selfish, desire, free, pleasure, reward, action, personal, selfless	truly, man	hill, earth, spread, mountain, firm, place, therein, shake, gather, become
8	meditation, mind, practice, sense, self, discipline, spiritual, still, awareness, path	part, agreement, body, division, make, together, join, might, christ, good	warner, plain, send, bringer, tiding, marvel, bearer, muhammad, goeth, good

9	brahman, sacrifice, offer, attain, ritual, proper, scripture, give, veda, perform	breast, breastplate, milk, birth, baby, woman, faith, give, take, child	jinn, wizard, magic, humankind, jinni, serpent, say, madman, fling, neither
10	lord, creature, birth, supreme, creation, universe, realize, form, god, eternal	record, history, act, book, rest, king, israel, judah, power, jeroboam	hand, right, record, salutation, book, leave, left, later, recording, white
11	fire, mouth, light, radiance, burn, moon, vishnu, shine, stand, infinite	light, table, support, vessel, put, front, room, place, opposite, side	favour, deny, lord, nevermore, supporter, firmly, madman, forasmuch, king, establish
12		record, history, act, book, rest, king, israel, judah, power, jeroboam	measure, draw, weigh, scale, near, weight, heavy, other, whose, mutually
13		light, table, support, vessel, put, front, room, place, opposite, side	couch, recline, eye, lovely, gaze, face, fair, wide, one, cushion

6.2 Sentiment Analysis

Using the VADER sentiment analysis, we obtain the sentiment scores for each verse of each book. Figure 13 shows the distribution of sentiment scores for the 3 religious texts. Here, we should note that the sentiment scores do not represent the moral positive or negative sentiment, but rather the use of positive or negative words, regardless of the context they are used in. From the figure, we see that the Bible and

the Quran have a higher percentage of neutral verses, while the Gita has a higher percentage of positively classed verses. This can be interpreted the Gita uses a higher percentage of words that have positive sentiments, regardless of the context they are used in.

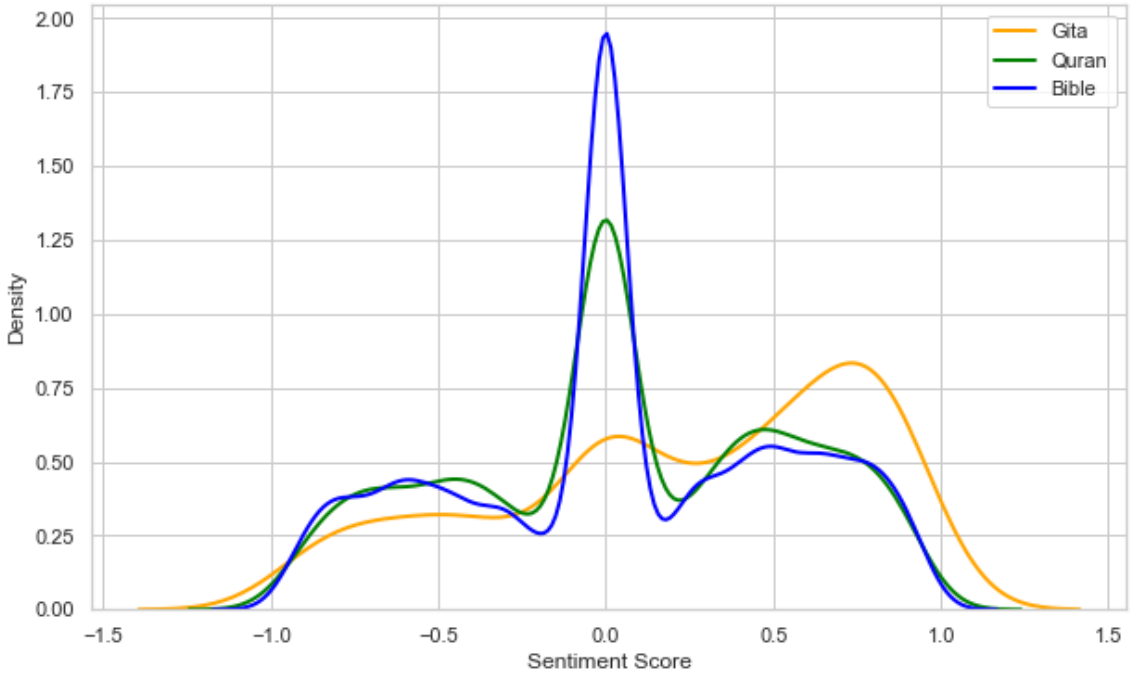

Figure 13: Distribution of Sentiment Scores of Religious Texts

6.3 Corpus Distance Measurement

The analysis of corpus distances and similarities between the religious texts revealed several key insights. The FID distance (Table 3) indicated that the Bible vs Quran pair had the closest distributional alignment (0.1664), while the Bible vs Gita and Gita vs Quran comparisons showed higher values, suggesting more divergence. In terms of semantic similarity, cosine similarity of averaged embeddings (Table 3) showed the Bible vs Quran pair had the highest alignment (0.8980), with the Bible vs Gita and Gita vs Quran pairs demonstrating lower similarities. KMeans clustering

further reinforced these findings, with centroid-based cosine similarity (Table 4) being highest for Bible vs Quran (0.7035) and the lowest Euclidean distance (0.4170), indicating closer semantic structures, while the Bible vs Gita comparison showed greater dissimilarity.

6.4 Sentiment Analysis on Modern Topics

The following topics were taken for the sentiment analysis of the religious texts on modern issues:

- Gender roles and gender equality
- Violence, war, peace, and forgiveness
- Slavery, human rights, and dignity
- Religious tolerance, interfaith relations, and apostasy
- Freedom of speech and expression
- Marital relations and sexuality
- Wealth, greed and charity
- Environment, preservation, and sustainability
- Criminal justice, law, and punishment
- Science, knowledge, and rationality

The number of relevant verses for each topic in each text are shown in figure 14.

The sentiment for each topic for each text is then taken as the average of the sentiments of the verses for that topic of that particular text. The topic-wise sentiments are shown in figure 15.

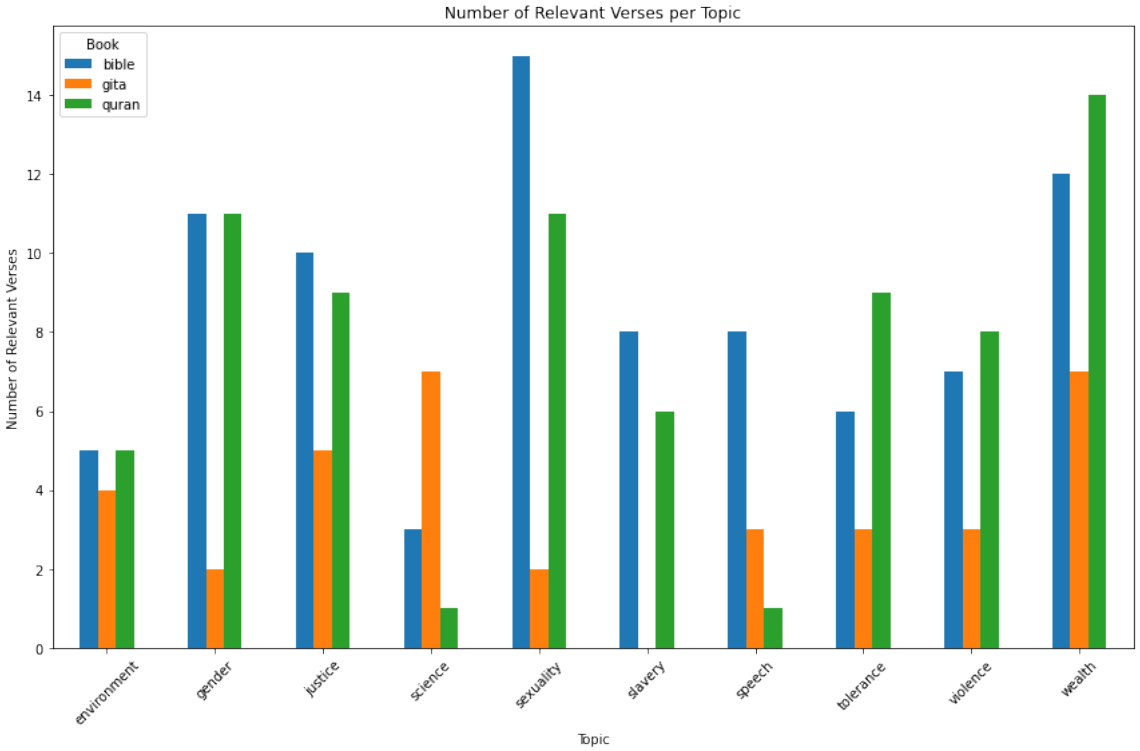

Figure 14: Number of Relevant Verses for each Modern Topics in Religious Texts

As we see from the topic-wise sentiments, most topics have a positive sentiment for all the texts. Topics with the most negative sentiment are violence and justice. Looking at the context of the books, which narrate stories of war, gives an understanding as to why the sentiments are mostly negative for these two topics.

Table 3: FID Distance and Cosine Similarity

Comparison	**FID Distance**	**Cosine Similarity**
Bible vs Gita	0.3868	0.7278
Bible vs Quran	0.1664	0.8980
Gita vs Quran	0.3801	0.7757

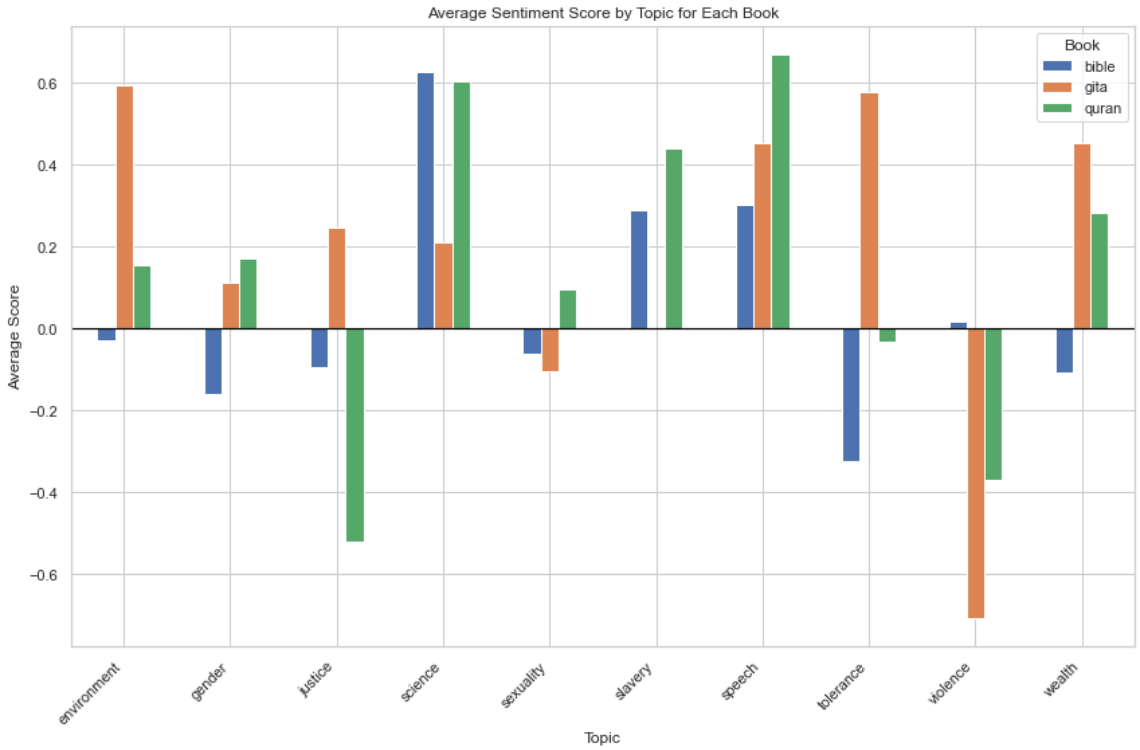

Figure 15: Sentiment Towards Modern Topics in Religious Texts

7 Conclusion and Future Works

Natural Language Processing (NLP) techniques have proven to be highly effective in analyzing and comparing the themes of the Bible, Quran, and Bhagavad Gita. By using methods such as Latent Dirichlet Allocation (LDA) for topic modeling, VADER for sentiment analysis, and corpus comparison, we uncover shared themes across these sacred texts. The analysis highlights unique philosophical dialogues and moral imperatives in each text while also revealing common ethical frameworks and human values that connect them. It also addresses the sentiments of these texts towards various modern issues. This research offers valuable insights into how religious texts continue to influence and shape cultural, ethical, and moral systems in modern day society.

Table 4: Average Cosine Similarity and Euclidean Distance

Comparison	**Avg. Cosine Sim.**	**Avg. Euclidean Dist.**
Bible vs Gita	0.5423	0.5284
Bible vs Quran	0.7035	0.4170
Gita vs Quran	0.6226	0.4911

Further research of religious texts could expand the scope by including texts from other diverse religious traditions. Advanced NLP techniques like transformer-based models for sentiment analysis and topic modeling could offer more detailed insights into the complex language and meanings within these scriptures. Furthermore, studying how interpretations of these texts have changed over time, particularly in response to societal changes, could provide a broader view of their continued relevance.

References

[1] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2017.

[2] B. Hutchinson, "Modeling the sacred: Considerations when using religious texts in natural language processing," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 1029–1043.

[3] 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, p. e0273476, 2022.

[4] R. Chandra and V. Kulkarni, "Semantic and sentiment analysis of selected bhagavad gita translations using bert-based language framework," *IEEE Access*, vol. 10, pp. 21 291–21 315, 2022.

[5] M. Alhawarat, "Extracting topics from the holy quran using generative models," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, pp. 288–294, 2015.

[6] A. Abbasi, A. R. Javed, F. Iqbal, N. Kryvinska, and Z. Jalil, "Deep learning for religious and continent-based toxic content detection and classification," *Scientific Reports*, vol. 12, no. 1, p. 17478, 2022.

[7] R. Chandra and V. Kulkarni, "Semantic and sentiment analysis of selected bhagavad gita translations using bert-based language framework," *IEEE Access*, vol. 10, pp. 21 291–21 315, 2022.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018. [Online]. Available: https://arxiv.org/abs/1706.08500

[10] G. Kour, S. Ackerman, E. Farchi, O. Raz, B. Carmeli, and A. Anaby-Tavor, "Measuring the measuring tools: An automatic evaluation of semantic metrics for text corpora," in *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2022)*. Association for Computational Linguistics, 2022.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[12] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[13] Y. M. Peuriekeu, V. D. Noyum, C. Feudjio, A. Goktug, and E. Fokoue, "A text mining discovery of similarities and dissimilarities among sacred scriptures," *arXiv preprint arXiv:2102.04421*, 2021.

[14] P. Goel and R. Arsiwala, "A comparative study of religious scriptures using natural language processing," *Journal of Student Research*, vol. 13, no. 2, 2024.