ABSTRACT

The Bhagavad Gita, a central sacred text in Hinduism, has transcended cultural boundaries with its profound teachings. However, the application of cutting-edge Machine Learning (ML) techniques to simplify its learning process remains largely unexplored. Prior research has predominantly focused on lexical and linguistic aspects, neglecting semantic understanding. The emergence of Transformers in Natural Language Processing (NLP) offers new possibilities for analyzing spiritual texts like the Bhagavad Gita. This study introduces a novel approach by integrating the BERTopic algorithm to perform comprehensive topic modeling. Ten distinct topics were identified, with the model achieving a coherence score of 0.61 and diversity score of 0.30, demonstrating its effectiveness. Sentence Transformer embeddings in conjunction with Principal Component Analysis (PCA) were identified as the optimal document representation, and K-Means clustering was employed. Notably, the identified topics align closely with the Bhagavad Gita's teachings, showcasing the efficacy of this approach within the chosen domain. This research pioneers advanced NLP investigations into the Bhagavad Gita, bridging a significant gap in the field. Previous studies had relied on rudimentary linguistic approaches, necessitating a fresh perspective. This study not only marks a pioneering achievement but also establishes a crucial foundational framework for future investigations. It identifies the BERTopic methodology as a promising tool for NLP analysis in this context and offers a methodology with potential for broader adoption in the research community. In essence, this research breaks new ground by applying advanced NLP techniques to the Bhagavad Gita, opening doors to deeper insights and fostering future inquiries in this domain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI……Artificial Intelligence

BERT…….....Bidirectional Encoder Representation from Transformer

BTm……...Biterm Topic Model

LSI……...Latent Semantic Indexing

CTM……...Correlated Topic Model

HDBSCAN……Hierarchical Density Based Spatial Clustering of Applications with Noise

GPT……...Generative Pretrained Transformer

LDA……Latent Dirichlet Allocation

LSTM……Long Short Term Memory

ML……..Machine Learning

MLM……...Masked Language Model

MMR……...Maximum Marginal Relevance

MT …….... Machine Translation

NER……...Named Entity Recognition

NMF……...Non Matrix Factorisation

NLP……Natural Language Processing

NSP……Next Sentence Prediction

PCA……...Principal Component Analysis

PMI……Pointwise Mutual Information

OCR……...Optical Character Recognition

RNN……Recurrent Neural Networks

SVD……...Singluar Value Decomposition

TFIDF……... term Frequency Inverse Document Frequnecty

UMAP……...Uniform Manifold Approximation and Projection

USE….....Universal Sentence Encoder

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Numerous scholars have undertaken in-depth investigations into the realm of religion from a philosophical vantage point, with the overarching objective of gaining comprehensive insights into fundamental themes, practices, and beliefs. This intellectual journey has ultimately coalesced into the established discipline of "philosophy of religion," a trajectory notably delineated by the seminal contributions of Murray and Rea(Murray and Rea, 2008) and further expounded upon by Meister(Meister, 2009). Within the ambit of Hindu philosophy, a multifaceted tapestry of diverse schools of thought and foundational principles has emerged, including but not confined to the elucidation of concepts such as yoga, karma, and Brahman. These philosophical tenets constitute integral components of the broader tapestry of Hindu metaphysical and spiritual inquiry. The corpus of Hindu philosophical teachings, deeply interwoven with millennia of theological contemplation, derives its sustenance from a myriad of sacred texts. Notably, the four Vedas, namely the Rigveda, Yajurveda, Samaveda, and Atharvaveda, serve as seminal sources of wisdom and spiritual elucidation. Additionally, the Upanishads, Puranas, and epic narratives such as the Mahabharata and Ramayana, among others, contribute to the rich reservoir of knowledge underpinning Hindu philosophical discourse. Furthermore, it is noteworthy that beyond the geographical boundaries of the East, a discernible proliferation of Western scholars has manifested a burgeoning interest in exploring the profound insights enshrined within Hindu sacred texts. This cross-cultural intellectual exchange has engendered a prolific body of translations and interpretations, signifying a convergence of Eastern and Western philosophical thought, as exemplified by the work of Renard(Renard, 1995) .

The Mahabharata, a monumental literary work composed in the Sanskrit language, stands as the preeminent Hindu epic, as affirmed by Rajagopalachari(Rajagopalachari, 1970). Within its sprawling narrative, it recounts the epic saga of a colossal conflict between two factions of royal lineage, both stemming from a common ancestral heritage. The Bhagavad Gita, often referred to as the "Song of the Lord," represents an integral component of the Mahabharata, heralded as a sacred scripture within the Hindu faith. It is celebrated for encapsulating the fundamental tenets of Hindu philosophy(Gandhi, 2010). Amidst the backdrop of impending warfare, Prince

Arjuna finds himself in a profound moral quandary, grappling with the ethical dilemma of engaging in battle against his own kin. This pivotal moment in the narrative sets the stage for a profound dialogue between Prince Arjuna and Lord Krishna, who serves as Arjuna's charioteer. This philosophical discourse, ensuing in a question-answer format, ultimately crystallizes into what is known as the Bhagavad Gita. The text, composed in contemporary Sanskrit and delivered in verse form, has undergone countless translations over the years, each aiming to disseminate its profound teachings. The Bhagavad Gita comprises 18 chapters, each dedicated to expounding upon distinct facets of Hindu philosophy. The reverberations of the Bhagavad Gita extend far beyond the confines of religious discourse, permeating diverse domains of human existence, including psychology, management, personality development, mental health, politics, economics, and leadership. This seminal text has been subjected to linguistic and semantic analysis, illuminating pathways to a deeper comprehension and propagation of its philosophical wisdom.Within the realm of Artificial Intelligence (AI), Natural Language Processing (NLP) emerges as a critical subfield, replete with methodologies and techniques geared toward the analysis of textual data, encompassing syntactic, structural, and semantic dimensions. NLP research encompasses a gamut of tasks, such as sentiment analysis, parsing, part-of-speech tagging, topic modeling, semantics, named-entity recognition (NER), optical character recognition (OCR), question-answer systems, speech recognition, summarization, and machine translation, as comprehensively elucidated by Chowdhury(Chowdhury, 2003). The evolution of NLP over the years reflects a paradigm shift from traditional statistical approaches to contemporary Bayesian statistics and, most notably, deep learning methodologies. Deep learning architectures, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models, have demonstrated enhanced efficacy in capturing the intricate relationships between words and their meanings, thereby facilitating the development of more robust language models, as delineated by Hochreiter and Schmidhuber(Hochreiter and Urgen Schmidhuber, 1997). Subsequently, the advent of encoder-decoder architectures, fortified with attention mechanisms, ushered in a transformative era in NLP research, epitomized by the groundbreaking work of Vaswani et al(Vaswani et al., 2017) with the introduction of Transformers. A preeminent exemplar within this framework is the Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model boasting an extensive parameter count exceeding 300 million. BERT's intrinsic capability to furnish contextual embeddings, as elucidated by Devlin et al.(Devlin et al., 2018), has firmly established it as the cornerstone of numerous NLP tasks, accentuating its profound influence on the contemporary landscape of AI-driven linguistic analysis.

Topic modelling is an NLP task that helps uncover latent topics/themes, that are meaningful and interpretable, within a collection of texts/documents. It is an unsupervised technique and does not need any labelled data. Based on the underlying approach to modelling, topic models are categorized into – Algebraic, Fuzzy, Probabilistic & Neural (Abdelrazek et al., 2023).

Topic modeling represents a pivotal endeavor within the domain of Natural Language Processing (NLP), striving to unveil latent themes and topics that reside concealed within textual corpora. The discernible objective is to derive a comprehensible and meaningful interpretation of the inherent content. Importantly, topic modeling is classified as an unsupervised learning technique, obviating the necessity for annotated or labeled data, thus positioning it as a versatile tool for knowledge extraction from textual archives.The categorization of topic models is predicated on the foundational methodology employed in the modeling process. These categories encompass the following key paradigms: Algebraic, Fuzzy, Probabilistic, and Neural(Abdelrazek et al., 2023). Each of these approaches possesses unique characteristics and brings forth diverse capabilities in discerning latent topics and themes within textual datasets. This thesis embarks on an exploration of these distinct topic modeling frameworks, aiming to elucidate their merits and demerits, with a view toward contributing to the ever-evolving landscape of text analysis and information retrieval.

1.2 Problem Statement

Apply unsupervised techniques like topic modeling and clustering to identify key themes from Bhagavad Gita.

1.3 Aim & Objectives

The primary objective of this research endeavour is to employ unsupervised Machine Learning (ML) and Natural Language Processing (NLP) techniques, specifically focused on topic modelling and clustering, in order to gain comprehensive insights into the principal thematic elements of the Bhagavad Gita. This scholarly pursuit is rooted in the aspiration to elucidate and comprehend the underlying motifs and subject matter of this revered Hindu scripture. The overarching ambition of this research transcends its immediate scope. By unraveling the thematic complexities of the Bhagavad Gita through advanced computational methodologies, it aspires to set a precedent for the systematic exploration of analogous sacred texts within Hinduism and, more broadly, within the spectrum of world religions. This approach is envisioned to yield a streamlined and automated framework for comparative studies across

diverse religious scriptures, thus facilitating a deeper understanding of their nuanced content and interconnections.

The research objectives articulated herein are intricately derived from the central aim of this study and are delineated as follows:

- To discern and establish the most suitable document representation scheme for the Bhagavad Gita, one that aligns optimally with the requirements of topic modelling.
- To undertake a comprehensive assessment and comparative analysis of various topic modelling techniques, with a view to discerning the most efficacious method for modelling the thematic structure of the Bhagavad Gita text.
- To meticulously apply the selected topic modelling methodology, thereby elucidating, interpreting, and identifying the salient key themes embedded within the textual corpus of the Bhagavad Gita.

1.4 Research Question

The following research questions are suggested for each of the research objective as highlighted as follows.

- What is the most appropriate document representation scheme for the Bhagavad Gita that aligns optimally with the requirements of topic modelling?
- How can various topic modelling techniques be comprehensively assessed and compared to determine the most efficacious method for modelling the thematic structure of the Bhagavad Gita text?
- Through the selected topic modelling methodology, what salient key themes can be elucidated, interpreted, and identified within the textual corpus of the Bhagavad Gita?

1.5 Scope of the Study

1.5.1 In scope

The following aspects are in scope for this research:

- Within the framework of this research, a singular translation will be judiciously chosen to serve as a representative exemplar, thus constituting a concrete manifestation of the proposed proof of concept.
- This research is delimited by its exclusive engagement with the English language and the Bhagavad Gita in its written textual form, excluding any consideration of oral recitations or alternative linguistic representations.

- The assessment of similarity and distinctiveness across various thematic dimensions shall be rigorously conducted exclusively through the application of quantitative metrics, eschewing qualitative evaluations.
- A systematic exploration of diverse topic modeling methodologies shall be undertaken, followed by a rigorous comparative analysis to ascertain their relative efficacies in the context of the research objectives.

1.5.2 Out of scope

The following aspects are in scope for this research:

- Apply same approach to other translations to check for consistency of results
- Apply same approach to other Hinduism texts
- Apply same approach to sacred texts of other religion to other texts
- Evaluation of results by domain experts in Hindu philosophy
- Verification by domain experts
- Run exhaustive comparison across all topic modelling approaches to evaluate the best

1.6 Significance of the Study

This research endeavours to offer a concise exposition of the fundamental themes contained within the Bhagavad Gita, eschewing an exhaustive examination of its intricate prose and extensive narrative. This deliberate simplification augments the dissemination of its salient teachings beyond the confines of the scholarly realm, thereby reaching a broader global audience. The envisaged consequence of the heightened propagation of these Bhagavad Gita tenets entails an anticipated enhancement in the overall quality of human existence.This topical analysis framework, as applied herein, holds the potential for expansion to encompass other sacred scriptures intrinsic to the Hindu tradition, thereby further enriching the tapestry of philosophical insights available for contemplation and practical application in daily life. Notably, the underlying premise posited by the author of this research project posits that its outcomes may potentially serve as a blueprint for analogous investigations into the sacred texts of diverse religious traditions, transcending the boundaries of Hinduism.Central to this hypothesis is the notion of a substantial congruence among the core thematic elements propagated by diverse religious doctrines across the globe. This profound revelation, if substantiated, holds the promise of mitigating interfaith conflicts and is deserving of heightened emphasis as the contemporary world ardently aspires towards global harmony and peace.

1.7 Structure of the Study

This thesis is systematically structured into six chapters. Chapter 1 introduces the study, providing background information, research aims and questions, and highlighting the significance of the research. Chapter 2 serves as the foundational framework, conducting a comprehensive literature review that includes insights from the Bhagavad Gita, an exploration of Natural Language Processing (NLP), topic modelling, and the intersection of religious scriptures and machine learning techniques. Chapter 3 delves into the methodology, detailing data collection methods and the meticulous selection of methodologies for topic modelling. Chapter 4 presents preliminary investigative outcomes and the experimental framework. Chapter 5 offers a comprehensive exposition of the final experimental outcomes, discussing findings and limitations. Finally, Chapter 6 concludes the thesis by summarizing research findings, providing recommendations for future research, and highlighting the study's contributions to existing knowledge. This structured approach ensures a systematic and coherent presentation of the research journey.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Chapter 2 of this thesis undertakes a comprehensive and meticulous examination of the pertinent literature, serving as the foundational framework for the subsequent research endeavors. Within this chapter, Section 2.2 is dedicated to an exhaustive exploration of the insights derived from the Bhagavad Gita and their pragmatic implications. Subsequently, Section 2.3 delves into an exhaustive investigation of Natural Language Processing (NLP) and the diverse formulations through which it can be effectively deployed. Likewise, Section 2.4 embarks on an intricate examination of topic modeling, offering a profound understanding of its nuances and applications. Further enriching the scholarly discourse, Sections 2.5 and 2.6 scrutinize the intersection of religious scriptures and machine learning techniques, scrutinizing their convergence and the potential applications of the latter on the textual content of the former. Finally, Section 2.7 contributes to the scholarly dialogue by engaging in an in-depth discourse concerning the existing lacunae and challenges within the domain, offering a nuanced perspective on the areas that demand further investigation and elucidation. This comprehensive overview establishes the conceptual underpinnings and intellectual context that underscore the subsequent chapters of this thesis.

The literature review endeavors to discern antecedent research endeavors pertaining to the application of machine learning techniques to the textual corpus of the Bhagavad Gita. The quest for relevant literature was guided by a set of carefully selected keywords, encompassing not only "Bhagavad Gita" but also key terms such as "machine learning," "artificial intelligence," "natural language processing," and "data science." To ensure a rigorous selection of scholarly sources, the search was principally conducted on renowned academic platforms, with primary emphasis on Google Scholar and the Liverpool John Moore’s University (LJMU) digital library. These platforms offer extensive access to a diverse array of peer-reviewed journals and authoritative databases, thereby facilitating the identification of the most pertinent and credible sources within the research domain.

2.2 Study of knowledge from Bhagavad Gita and its applications

Over the years, the Bhagavad Gita has garnered renewed attention, not only as a spiritual scripture but also as a source of inspiration and wisdom for various disciplines, including organisational management, psychology, healthcare, leadership, governance, and economics. Scholars, researchers, and practitioners from different fields have delved into the depths of the Gita, seeking to extract its relevance and insights in the context of contemporary challenges and aspirations. This section of the literature review embarks on a journey through the multifaceted landscape of studies that have illuminated the practical applications of the Bhagavad Gita's teachings. Each study in this review adds a unique facet to the ever-evolving understanding of how the Gita's principles can shape and enrich our lives, from personal well-being to the broader realms of societal and organisational dynamics.

Mukherjee's (Mukherjee, 2017) recommendation for organisations to draw management lessons from the Bhagavad Gita encompasses a wide spectrum of essential themes, making it a valuable source for guiding principles in various aspects of organisational dynamics. The Bhagavad Gita's teachings on duty/action, anger management, work culture, commitment, mental health, and goal setting serve as a rich reservoir of insights that can empower organisations to navigate complex challenges. Nagappa's qualitative and conceptual study (Nagappa, 2023) underscores the profound impact of life management principles derived from the Bhagavad Gita. This study not only emphasizes the positive outcomes when these principles are applied in both human and corporate contexts, but it also delves into the intricate nuances of themes such as stress, suffering, balance, focus, knowledge, dedication, and the significance of smart work. By adopting these principles, organisations can aspire to foster a healthier work environment and enhance overall well-being. Analogous to analyses of other religious texts such as the Quran and Bible, Muniapan's exploration (Muniapan and Satpathy, 2013) of the Bhagavad Gita's relevance in business management, particularly in the realm of social responsibility, highlights a critical aspect of organizational ethics. This study's focus on duty and action provides a robust framework for integrating social responsibility at three essential levels: the individual, the corporate entity, and the global community. Organisations that embrace these principles can become catalysts for positive societal change while aligning with their core values.

Jeste's (Jeste and Vahia, 2008) mixed qualitative/quantitative study on the Bhagavad Gita's inspiration in psychology reflects the timeless wisdom embedded in this ancient text. The

exploration of concepts such as duty over action, renunciation of pleasure, and materialistic desires underscores the deep relevance of these principles in the fields of psychiatry and psychotherapy. These insights provide mental health professionals with alternative perspectives and strategies for addressing complex psychological challenges. Pandurangi's (Pandurangi et al., 2014)work sheds light on the therapeutic potential of the Bhagavad Gita's verses, offering a unique perspective on integrating cognitive behaviour and metacognitive principles into psychotherapy. The identification of specific verses with relevance to psychotherapeutic practices provides a structured framework for counsellors and therapists seeking to leverage the timeless wisdom of the Bhagavad Gita in their interventions. Keshavan's (Keshavan, 2020) principles for resilience, drawn from the Bhagavad Gita's three paths – the path of knowledge, the path of action, and the path of meditation – offer valuable guidance in navigating challenging situations, such as the unprecedented global crisis brought about by the COVID-19 pandemic. By embracing these paths, individuals and organizations can develop resilience and adaptability, enabling them to thrive even in the face of adversity. In a timely extension of Keshavan's work, Das(Das and Behura, 2021) highlights the relevance of the Bhagavad Gita for first responders and frontline workers, who bear the brunt of the pandemic's impact. The Gita's teachings on duty, action, and maintaining equanimity in the face of overwhelming challenges provide a source of solace and inspiration for these dedicated individuals, offering a framework for coping and finding meaning in their critical roles. Menon et al.'s (Menon et al., 2021) exploration of moral injury experienced by frontline healthcare workers during the COVID-19 pandemic sheds light on the emotional turmoil faced by those at the forefront of healthcare. The Bhagavad Gita's guidance on managing uncertainty, helplessness, and anguish offers a path to emotional resilience and motivation, helping healthcare professionals navigate the emotional toll of their vital work. Bhatia et al. (Bhatia et al., 2013) draw parallels between Cognitive Behaviour Therapy (CBT) and the principles of the Bhagavad Gita, advocating for a more holistic and spiritually grounded approach in psychotherapy. By incorporating these principles, therapists can create a more comprehensive and integrated therapeutic experience that addresses not only cognitive aspects but also the deeper dimensions of the human experience. In the realm of chronic illness management, Kalra et al.'s (Kalra et al., 2018) recommendation to adopt coping mechanisms from analogies within the Bhagavad Gita is particularly relevant. The lessons from the verses of Arjuna and Krishna offer a rich source of metaphors and insights that can empower individuals facing long-term health challenges to cope more effectively and maintain a positive outlook. Furthermore, Kalra et al.'s (Kalra et al., 2017)exploration of the ideal attributes and behaviours of a good physician, aligning with the

teachings of Krishna in the Bhagavad Gita, highlights the profound connection between medical ethics and spiritual wisdom. Physicians embodying qualities such as knowledge, care, leadership, humanity, control, and compassion, as emphasized in the Bhagavad Gita, can contribute significantly to the well-being of their patients and the overall healthcare ecosystem. Dhillon's comprehensive study (Dhillon, 2023)on the relevance and applications of the Bhagavad Gita in psychology underscores the historical and philosophical depth of Eastern traditions in contrast to the predominantly Western origins of modern psychological theories. The Gita's timeless insights offer a unique perspective that predates modern psychological developments, holding the potential to enrich contemporary psychological practice and contribute to the holistic well-being of individuals.

Bhagavad Gita was also studied from a political and leadership perspective. Pandey's(Pandey, 2017) economic lens brings a fresh perspective to the Bhagavad Gita's applicability, demonstrating how its principles extend beyond spiritual and personal domains. The insights from the Gita can be translated into actionable strategies for governance, production, productivity, development, self-reliance, and the effective division of labour. By incorporating these principles, economic systems can be designed with a more balanced and sustainable approach. Simpson & Cunha (Simpson and Pina e Cunha, 2021) contribute a valuable leadership model derived from the Bhagavad Gita, encompassing four fundamental principles: self-leadership, servant leadership, holistic systems, and a higher purpose. This holistic leadership framework not only aligns with the Gita's teachings but also serves as a comprehensive guide for creating value within the context of business management. By fostering these principles, organizations can create a culture that transcends mere profit and focuses on meaningful and sustainable value creation for all stakeholders. Satpathy & Muniapan(Satpathy et al., 2013) extend their exploration to the realm of good governance, revealing the striking alignment between institutional governance principles and those derived from the Bhagavad Gita. The Gita's emphasis on ethical conduct, duty, and responsibility offers valuable guidance for governance at various levels, fostering transparency, accountability, and ethical decision-making. Sudhakar's(Purna Sudhakar, 2014) application of the Bhagavad Gita to project management introduces a novel perspective on the attributes and behaviours of successful projects, the values that underpin them, and the role of team members and leaders. By embracing the qualities advocated in the Bhagavad Gita, project managers can create a conducive environment for successful project execution, emphasizing values, goals, teamwork, and effective leadership.

The diverse array of studies presented in this review collectively paints a comprehensive picture of the Bhagavad Gita's enduring relevance in various aspects of human life. By traversing this journey through the Gita's impact on organizational management, psychology, healthcare, economics, leadership, and governance, we aim to uncover the deep-rooted wisdom that transcends time and continues to guide and inspire individuals and societies on their quests for well-being, resilience, and meaningful existence.

In conclusion, our journey through the diverse studies illuminates the Bhagavad Gita's enduring relevance in shaping diverse aspects of human existence. Its teachings serve as a guidepost, providing insights into the profound truths of life, the nature of duty, and the pursuit of knowledge. As we embrace the wisdom of the Gita, we embark on a path towards a more enlightened, compassionate, and resilient world—one where the principles of duty, action, and compassion can pave the way for a brighter and more harmonious future.

2.3 Natural Language Processing (NLP)

Human communication finds its foundation in the utilization of language as a medium for the efficient conveyance of information. Language is an intricate system governed by codified rules, is characterized by symbolic representations, and assumes a pivotal role in facilitating interpersonal interactions and information dissemination. Natural Language Processing (NLP) is a critical field within Artificial Intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language in a way that's both meaningful and useful. NLP involves developing algorithms and models that can process and analyse text and speech data, allowing machines to perform tasks that traditionally required human understanding. NLP, is situated at the cross of linguistic analysis and computational methodologies, and embarks upon the deconstruction of linguistic intricacies and their subsequent representation in algorithmic and model-driven forms. Some of the key problem formulations within NLP include:

- Machine Translation: Translating text or speech from one language to another. This is exemplified by tools like Google Translate
- Language Modelling: Building models that can predict the likelihood of a sequence of words, which serves as the foundation for many NLP tasks.
- Summarization: Creating concise summaries of longer texts, useful for news articles or research papers.

- Language Understanding: Extracting meaning from text by understanding relationships between words, phrases, and concepts.
- Text Generation: Creating coherent and contextually relevant text, which can be used for chatbots, content creation, and more
- Sentiment Analysis: Determining the emotional tone of a piece of text, often used in social media monitoring and customer feedback analysis
- Named Entity Recognition (NER): Identifying and classifying entities mentioned in text, such as names of people, places, organizations, and more
- Text Classification: Categorizing text into predefined classes or categories, like spam detection or topic categorization
- Speech Recognition: Converting spoken language into written text, used in voice assistants like Siri or Alexa
- Question Answering: Developing systems that can comprehend and answer questions posed in natural language, as seen in systems like IBM Watson's Jeopardy-playing AI.

NLP has found diverse application across a wide spectrum of domains. Noteworthy among these are the domains of healthcare, medical science, financial endeavours, and market research. Within such arenas, NLP has demonstrated its versatility through the pursuit of tasks spanning medical record scrutiny, sentiment analysis within financial discourse, pharmacological discovery facilitated by textual data extraction, and consumer sentiment comprehension. Multiple reviews on NLP have been done. Khurana (Khurana et al., 2023) looked into current, past and future trends including challenges.

Most of foundational work in NLP originated in Machine Translation (MT) and happened since the 1960s and 1980s. Early 2000s paved way for more sophisticated approaches. Bengio (Bengio et al., 2003) established a seminal framework for language modelling, employing a neural network-centric approach to approximating joint probabilities of co-occurrence of words. This undertaking yielded a discernibly enhanced efficacy compared to prevailing n-gram-based language models of the era. However, challenges were observed from a perspective of contextual length, investment in training, and computational resource intensiveness. Work by Collobert (Collobert and Weston, 2008) astutely recognized prevailing challenges within the domain and advocated for the employment of semi-supervised learning in conjunction with multi-task learning paradigms. This approach entails the construction of layered neural networks, wherein shared parameters are instantiated across distinct tasks encompassing tasks

such as Part-of-Speech tagging, named entity recognition, semantic role labelling, chunking, and language modelling. As anticipated, the adoption of this method yielded enhanced performance outcomes concerning targeted tasks compared with models of a more generalized nature.

A pivotal juncture within NLP emerged through the work of Mikolov (Mikolov et al., 2013a) wherein groundbreaking research was made regarding the feasibility of encapsulating words within a vectorized format. This advancement culminated in the inception of the widely acclaimed word2vec framework, thus heralding a transformative era in linguistic representation. This departure from prior methodologies, such as the conventional bag-of-words paradigm employed for textual representation, signifies a paradigm shift of considerable significance. Subsequent to this yet another word embedding methodology emerged, as espoused by Pennington(Pennington et al., 2014) through introduction of Glove embeddings.
The next major paradigm shift started in 2013-14 with the application of larger and different types of architectures of neural networks in modelling NLP tasks as sequence-to-sequence formulations. Convolution Neural Networks (CNNs), were traditionally designed and used for image processing and computer vision related tasks. However, several applications of CNNs to NLP tasks was experimented with. Wang(Wei Wang and Jianxun Gang, 2018) elaborated on the application of CNN models to NLP tasks. Similarly, other NLP tasks were also experimented with the CNNs like Classification (Santoro et al., n.d.; Socher et al., n.d.; Tan et al., 2022), Translation (Luong et al., 2014), Question Answering (Wiese et al., 2017), Summarisation (Yu et al., 2018). The next major paradigm shift was in the modelling NLP tasks as sequence-to-sequence problems was through Recurrent Neural networks (RNNs) and their variants - Long Short-term memory (LSTM) and Gated Recurrent Units (GRUs). RNN based architectures demonstrated better results than any of the previous state-of-the-art models as they were able to capture dependency across the sequence. However, a significant drawback was that they were unable to capture dependencies over a longer sequence, which is often the case with text data and NLP tasks. With the advent of Attention mechanism(Bahdanau et al., 2014) and Transformer architecture(Vaswani et al., 2017), the world of NLP has been revolutionised and the current state-of-the-art is deeply rooted based on these mechanism and architectures. Attention mechanism is able to capture relationships between words and retain the memory even for longer sequences.

2.4 Topic modelling

Topic modelling is a natural language processing (NLP) technique used in text analysis to discover hidden thematic structures in a collection of documents. The primary goal of topic modelling is to automatically identify and extract topics or themes from a large corpus of text data without any prior knowledge of what those topics might be. It's a way to organize and summarize text data, making it easier to understand and explore large text collections.

In the context of comprehensive literature exploration pertaining to the domain of topic modelling, a meticulous examination was undertaken through the perusal of multiple survey and review publications(Jelodar et al., 2019; Vayansky and Kumar, 2020; Yamunathangam et al., 2021; Osuntoki et al., 2022; Abdelrazek et al., 2023). Notably, Abdelrazek's review(Abdelrazek et al., 2023), being the most recent and all-encompassing, encompasses diverse facets of topic modelling, including algorithmic typology, evaluation methodologies, practical applications, computational tools, available datasets and benchmarks, as well as the discernment of evolving research trends. Vayansky's review(Vayansky and Kumar, 2020) presciently suggests a departure from the prevalent Latent Dirichlet Analysis (LDA) model, advocating instead for the adoption of modelling paradigms characterized by enhanced flexibility and the capacity to capture intricate inter-topic relationships, particularly embracing the dynamic aspect of temporal evolution amongst topics. Furthermore, Yamunathangam's comprehensive analysis(Yamunathangam et al., 2021) offers invaluable insights, explicitly tailored to the scale of textual data under consideration. This is particularly pertinent in the contemporary context where the proliferation of textual data sources from social media platforms necessitates nuanced and scalable modelling approaches.

Topic modelling, owing to its unsupervised nature, finds myriad applications spanning diverse domains and holds particular significance within the realm of Natural Language Processing (NLP) for subsequent tasks. The extraction of interpretable semantic themes serves as an initial step, often preceding any classification task(Sun et al., n.d.) Additionally, it facilitates feature extraction. Furthermore, topic modelling is instrumental in the comparative analysis of two corpora, enabling a comprehensive examination of their topic distribution for enhanced semantic comprehension. Such comparisons contribute to a nuanced understanding of similarity and coherence. Haghighi(Haghighi and Vanderwende, n.d.) pioneered the application of topic modelling to document summarization, albeit the absence of rigorous quantitative comparison methodologies. The utility of topic modelling extends across a wide spectrum of domains,

exemplified by Jeong's (Jeong et al., 2019) utilization in the realm of social media for product planning, where topics and associated sentiments are subject to rigorous analysis. Sun(Sun and Yin, 2017) employed topic modelling techniques to dissect research papers within the field of transportation, providing invaluable insights into research trends. Ambrosino (Ambrosino et al., 2018) conducted a meticulous study of topic trends within the field of economics over time intervals. Multiple instances of topic modelling applications have emerged in various domains, including software and technology(2011 26th IEEE/ACM International Conference on Automated Software Engineering., 2011; Institute of Electrical and Electronics Engineers, 2012; Dit et al., 2013; Hemmati et al., 2017), the airline industry(Srinivas and Ramachandiran, 2020) , and bioinformatics(Liu et al., 2016) . These diverse applications underscore the versatility and potential of topic modelling techniques in knowledge discovery and analysis.

In the realm of topic modelling, a notable facet of consideration resides in the manifold variations encompassing underlying assumptions, probability distributions, interpretational paradigms, and evaluation methodologies. Notably, the trajectory of research endeavours within topic modelling gained significant impetus during the 1990s with the introduction of the Latent Semantic Indexing (LSI)(Deerwester et al., 1990) methodology. It is imperative to acknowledge that LSI and Non-negative Matrix Factorization (NMF)(Lee and Seung, 1999) represent foundational algebraic models within this domain, both principally rooted in matrix factorization techniques and the utilization of a bag-of-words representation to encapsulate textual data structures. Amongst the pantheon of topic modelling methodologies, the most widely embraced one is the Latent Dirichlet Allocation (LDA)(Blei et al., 2003a), characterized by its probabilistic framework. Nevertheless, LDA has grappled with several intrinsic challenges, including the constraints imposed by the bag-of-words representation, issues of robustness, stability, order, and the necessity for diligent hyperparameter tuning. These challenges have precipitated subsequent advancements, notably exemplified by the Correlated Topic Model (CTM)(Blei and Lafferty, n.d.), founded upon the fundamental assumption of logistic distribution. CTM has sought to mitigate the shortcomings associated with its predecessor. In the domain of topic modelling, a distinctive subcategory pertains to the modelling of short textual data. In this context, variants of LDA(Hofmann, 1999) have encountered significant difficulties owing to the inherent sparsity of such data. Consequently, this predicament has ushered in the emergence of the Biterm Topic Model (BTM), which has been devised to address the unique challenges posed by short textual corpora. Furthermore, it is pertinent to elucidate that certain modelling approaches have been conceptualized, predicated

upon the incorporation of fuzzy algorithms. These innovative approaches(Yang, 1993) are characterized by the amalgamation of dimensionality reduction techniques and clustering methodologies, all orchestrated to discern and encapsulate semantic similarity patterns within the textual data.

The contemporary landscape of topic modelling within the field of Natural Language Processing (NLP) is predominantly underpinned by the paradigm of deep learning. Neural topic models seamlessly integrate deep learning methodologies into multiple facets of the topic modelling process, effectively addressing the limitations inherent in preceding iterations of topic modelling techniques such as Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Indexing (LSI). A recurring theme among neural topic models involves the harnessing of neural embeddings, which serve as representations that map words and documents into an n-dimensional space.

Noteworthy contributions to the advancement of topic modelling techniques include the work of Dieng et al.(Dieng et al., n.d.), who discerned shortcomings associated with LDA in the context of expansive vocabularies. Their proposal involves an embedded topic model, extending LDA through the incorporation of embeddings. This variant effectively caters to corpora characterized by substantial vocabularies, albeit persisting performance challenges.
Furthermore, the investigations of Das et al. (Das et al., n.d.) and Nguyen et al. (Nguyen et al., n.d.) have augmented LDA-based models by introducing a novel inference strategy employing Gibbs sampling. This embedding-based approach demonstrates adaptability to out-of-vocabulary terms, further enhancing its utility. The evolution of topic modelling techniques extends to LDA2VEC(Moody, 2016), a model that amalgamates Word2Vec(Mikolov et al., 2013a) embeddings with LDA, resulting in a synergistic learning approach. The present state-of-the-art of topic modelling methodologies builds upon the seminal work of Grootendorst(Grootendorst, 2022) and Sia (Sia et al., 2020) . Their approach constitutes a multifaceted process founded on pretrained embeddings, clustering, and topic representation models, each capable of autonomous operation. BERTopic(Grootendorst, 2022), an exemplar in this domain, relies on BERT-based(Devlin et al., 2018) embeddings and employs HDBSCAN(McInnes and Healy, 2017) for clustering, signifying a contemporary culmination of advancements in the field. Neural topic models offer numerous advantages in terms of performance and scalability compared to conventional topic models.

2.5 Text analysis of Bhagavad Gita and other religious texts

During the linguistic scrutiny of the Spanish rendition of the Bhagavad Gita, Stein (2012) perceptively discerned an abundant presence of metaphors and multi-word expressions within the text. This observation, while offering valuable insights, also emphasized the intricate nature inherent in sacred scriptures. The application of the graphical local grammars method, employing semantic categorization, yielded commendable yet rudimentary outcomes. Nevertheless, a pivotal inquiry remains unresolved, pertinently addressing the relationship between the multi-word expressions in the original Sanskrit rendition of the Bhagavad Gita and their transference within the utilized Spanish translation. The requisite groundwork encompassing pertinent data preprocessing and the pursuit of a more intricate analytical methodology remain unexplored avenues that merit further investigation. The statistical examination was executed by Rajput and colleagues(Rajput et al., 2019) on English, French, and Hindi translations, juxtaposed against the source Sanskrit text. Evaluation of diverse statistical distribution metrics unveiled the linguistic wealth and distinctiveness of Sanskrit in terms of lexical and vocabulary structures, with negligible disparities observed amidst the translated renditions.

Rajandran's study(Rajandran, 2017) discerned the substantial presence of metaphoric expressions rooted in the human corporeal structure and ancient life within the Bhagavad Gita. This comprehensive investigation, characterized by its focus on linguistic and semantic aspects, yields profound elucidation of the fundamental themes and communicative content inherent in the Bhagavad Gita. A pressing desire emerges for a computational framework integrating machine learning and algorithmic techniques to extend and operationalize this methodological approach. The influence of the translator's subjective inclinations is apparent in the eventual skew of the outcomes. In their recent work, Chandra(Chandra and Kulkarni, 2022) undertook a commendable endeavour to employ enhanced and advanced methodologies for conducting a comprehensive semantic and sentiment analysis of diverse translations of the Bhagavad Gita. The judicious selection of datasets encompassing translations by historical figures (Mahatma Gandhi, Shri Purohit Swami, and Eknath Easwaran) spanning different epochs, media, and domains is well-founded.

In recent times, considerable advances have been witnessed in the realm of deep learning, particularly evident in the enhanced capacity of Transformer-based models to encapsulate

semantic nuances and extract meaningful representations. This progress substantiates the assertion that Transformer-based models stand at the forefront of state-of-the-art approaches. Notably, Chandra and Ranjan(Chandra and Ranjan, 2022) conducted a comparative analysis of thematic elements within the Bhagavad Gita and the Upanishads. Employing measures of topic coherence and cosine similarity, they demonstrated a substantial degree of convergence, with an approximate 70% thematic similarity between the Upanishads and the Bhagavad Gita, a finding of considerable significance. In the process of elucidating these thematic parallels, Chandra and Ranjan adopted an innovative approach, leveraging dimensionality reduction techniques coupled with clustering methodologies. This unique strategy facilitated the visualization of these topics in a two-dimensional space, a novel contribution that extends the boundaries of the domain. Such an approach not only sheds new light on the comparative analysis of these ancient texts but also introduces a pioneering methodology with the potential to reshape research paradigms in this field. Karekar and colleagues(Karekar et al., 2023) embarked upon an endeavor to construct a dialogic AI agent designed for engagement with the Bhagavad Gita. This scholarly pursuit entails a captivating proposition, involving the integration of ChatGPT-like attributes from the OpenAI framework into a bespoke system tailored for the Bhagavad Gita. In this intellectual endeavor, Karekar seeks to enhance upon previous iterations of similar functionality, as exemplified by the GitaGPT system developed by a Google engineer. Noteworthy is the incorporation of cutting-edge models fostered by OpenAI. The present scholarly contribution expounds upon a rudimentary and unembellished methodology employed in the instantiation of the aforementioned system. Regrettably, the narrative herein provided does not extend its purview toward matters of usability nor does it indulge in comprehensive evaluative exercises. This lacuna presents an avenue for subsequent explorations, encompassing delineation of discerned use cases, meticulous preprocessing of pertinent data, and the establishment of a robust framework for evaluative pursuits.

2.6 Text analysis of other religious texts

Verma(Verma, 2017) conducted an initial exploration employing lexical analysis on a corpus of ten religious' texts, encompassing prominent works such as the Bible, Dhammapada, Torah, Bhagavad Gita, Tao Te Ching, Guru Granth Sahib, Agama, Quran, Rigveda, and Sarbachan. This analytical endeavor primarily focused on gauging lexical richness as the chosen metric. It is pertinent to acknowledge that the findings yielded by this study were predominantly observational in nature, reflecting a cursory examination of textual features, with a limited

engagement with the underlying profound messages or nuanced meanings encapsulated within these sacred writings.

2.7 Discussion

The literature review indicates that, through its translations, the Bhagavad Gita is keenly studied and followed around the world for its teachings on philosophy of life. However, there has been very little that has been done in applying modern NLP and machine learning techniques to understand its themes for easy interpretation and consumption of teachings. There is a translational loss of meaning with different authors/writers across languages projecting tones based on their own interpretations that add a slight flavour to the version of Bhagavad Gita.
Of the few studies that were relevant, most studies have been based on lexical, syntactic and statistical distributions with focus on words, frequencies, etc. with no focus on the semantics and contextual meaning of words & texts. There is a lack of sophistication and robustness in the machine learning techniques and algorithms that have been used, especially given the recent push towards state-of-the-art deep learning-based models today. This research is deeply inspired by the works of Chandra (Chandra and Kulkarni, 2022; Chandra and Ranjan, 2022)

In the context of this discussion, it is imperative to delve deeper into the implications of the existing literature on the Bhagavad Gita and the untapped potential offered by modern Natural Language Processing (NLP) and machine learning methodologies. Firstly, the popularity of the Bhagavad Gita transcends geographical boundaries and cultural contexts, attracting a global readership. This widespread appeal underscores the relevance and significance of its teachings in contemporary society. The reliance on translations to access these teachings introduces a crucial dimension, as various translators inevitably infuse their interpretations, cultural biases, and linguistic nuances into the text. Consequently, understanding the true essence of the Gita's message becomes a challenge, and there is a pressing need for computational methods to assist in disentangling these nuances.

The existing research, as highlighted in the literature review, has primarily focused on superficial analyses of the Gita's text. Lexical, syntactic, and statistical approaches, such as word frequency analysis, have provided valuable insights into linguistic patterns and structure. However, they fall short in capturing the deeper philosophical and contextual meanings embedded within the text. This limitation is significant, given that the Bhagavad Gita is not merely a collection of words but a repository of profound philosophical concepts and life

lessons. Furthermore, the lack of sophistication and robustness in the machine learning techniques employed in prior studies is a noteworthy concern. In an era marked by rapid advancements in deep learning and artificial intelligence, the application of outdated methodologies can be seen as a missed opportunity. Cutting-edge deep learning models have demonstrated remarkable capabilities in capturing nuanced semantics and contextual information, making them well-suited for the complex task of interpreting philosophical texts like the Bhagavad Gita.

The research cited in this literature review underscores the need for a more comprehensive and sophisticated approach to understanding the Bhagavad Gita. The incorporation of modern NLP and machine learning techniques holds the promise of shedding new light on the text's intricate philosophical themes. By harnessing the power of deep learning-based models, researchers can aim for a more accurate and nuanced interpretation of the Gita's teachings, thus bridging the gap between the original Sanskrit text and its diverse translations. In conclusion, the literature review reveals a compelling opportunity to leverage state-of-the-art machine learning techniques to unravel the deeper philosophical and contextual meanings within the Bhagavad Gita. By addressing the translational challenges and shortcomings of previous research, this study aspires to contribute to a more profound and accessible understanding of this timeless philosophical masterpiece.

2.8 Summary

The literature review emphasizes the need for a more advanced approach to comprehending the Bhagavad Gita, suggesting that modern NLP and machine learning techniques can illuminate its intricate philosophical themes. These technologies offer the potential for a more precise interpretation of the text, bridging the gap between the original Sanskrit and translations. In summary, this review highlights the opportunity to use cutting-edge machine learning to uncover deeper meanings in the Bhagavad Gita, addressing translation challenges and enhancing our understanding of this timeless philosophical work. The literature review reveals that fewer than five papers apply advanced machine learning to Bhagavad Gita text, and fewer than twenty papers employ statistical analysis or algorithmic approaches. This underscores a notable gap in the existing literature regarding the NLP and advanced ML-based exploration of the Bhagavad Gita.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Chapter 3 of this thesis is dedicated to the presentation of the Methodology, a critical component in the pursuit of our research objectives. Section 3.2 initiates an in-depth exploration into the pertinent datasets and their collection methods, which serve as the foundation upon which our investigation rests. In the subsequent sections, we embark on a comprehensive examination aimed at the meticulous identification and selection of suitable methodologies for the execution of topic modelling exercises. This process entails a rigorous evaluation of various techniques and approaches, ensuring that our research methodology is methodically crafted to yield insightful and reliable results.

3.2 Dataset & collection

Upon recognizing the profound importance of the Sanskrit language as elucidated in the literature review, it becomes evident that the most ideal dataset for this research endeavor would comprise the original Sanskrit verses. Regrettably, the author's proficiency in Sanskrit is limited, and employing Sanskrit for the analysis would inherently restrict the accessibility of the ultimate findings. It is noteworthy that the Natural Language Processing (NLP) modules requisite for preprocessing and topic modeling have not yet achieved robust development for the Sanskrit language. Therefore, in the context of this research, the judicious choice is to utilize an English translation. Although numerous scholars worldwide have undertaken the translation of the Bhagavad Gita and have provided scholarly commentaries, our preference gravitates toward translations authored by individuals endowed with profound comprehension of both Sanskrit and Hindu philosophies. In light of this consideration, translations offered by Gandhi (Gandhi, 2010), Easwaran(Easwaran, 1985) , Purohit(Swami, 2003), Prabhupada(Prabhupada and Swami, 1972), and Mukundananda(Swami Mukundananda, n.d.) were systematically assessed to minimize potential biases. As a proof-of-concept, the dataset selected for this analytical pursuit is derived from the translation rendered by Swami Mukundananda(Swami Mukundananda, n.d.), and conveniently accessible in digital format via Kaggle Datasets(Yash Narnaware, 2023). All subsequent analysis on topic modelling applied on this translation can be applied on other translations too. However, it is important to note that for other potential sources, encompassing PDF documents, their inclusion necessitates a scanning process

followed by Optical Character Recognition (OCR) conversion, which regrettably falls beyond the scope of this research due to constraints pertaining to OCR accuracy, feasibility and timeline considerations.

The Bhagavad Gita comprises a total of 18 chapters, housing an approximate count of 700 verses within its corpus. In the context of the dataset employed for this research, it is imperative to delineate its structural composition. Specifically, each individual row within this dataset corresponds to a single verse extracted from the Bhagavad Gita. The columns within this dataset are meticulously designed to encapsulate a comprehensive array of linguistic and semantic facets pertaining to each verse. These columns encompass the key attributes presented in Table 3.1

Table 3.1 List of columns in dataset

Column name	Description
verse_number	Chapter number and verse number within the chapter
verse_in_sanskrit	Original Sanskrit rendition of the verse, retaining the integrity of the ancient language
sanskrit_verse_transliteration	English transliteration of the original Sanskrit rendition of the verse, retaining the integrity of the ancient language
translation_in_english	English translation of the Sanskrit verse, facilitating accessibility and comprehension for non-Sanskrit speakers
meaning_in_english	This column delves deeper into the semantic essence of the verse by providing a concise and elucidating interpretation in the English language
translation_in_hindi	To cater to a broader audience, this column is dedicated to furnishing the Hindi translation of the verse, extending the reach of the dataset to Hindi-speaking scholars and enthusiasts.
meaning_in_hindi	To cater to a broader audience, this column is dedicated to furnishing the Hindi translation of the verse, extending the reach of the dataset to Hindi-speaking scholars and enthusiasts.

This structured dataset serves as the foundational resource underpinning the analytical methodologies and investigations conducted within the purview of this research, enabling a comprehensive exploration of the Bhagavad Gita's textual content and semantic nuances. However, the primary focus will be the column – translation_in_english, which is the key column deriving a straight translation from Sanskrit. The column meaning_in_english will be used for comparison and validation purposes only to ensure analysis from translation aligns with the meaning as explained by the author of the translation.

3.3 Proposed approach for topic modelling

Topic modelling techniques help us identifying underlying semantic themes. Abdelrazek's (Abdelrazek et al., 2023) survey of topic modelling categorises the techniques into four, depending on the underlying algorithm and sophistication – algebraic, fuzzy, probabilistic and neural. For this research as proof-of-concept, we will pick an approach each from algebraic, probabilistic and neural-based and compare their performance. Non-negative matrix factorization (NMF) (Lee D.D and Seung H.S, 1999) is an algebraic model that is extensively used. It attempts to convert a high dimension vector as a product of two matrices which are non-negative. Amongst the probabilistic models, Latent Dirichlet allocation (LDA) (Blei et al., 2003b) is the most widely used (Abdelrazek et al., 2023)

The principal objective of topic modelling is the identification of latent semantic themes characterized by interpretability and robustness. In formulating the fundamental methodology for topic modelling on the selected dataset, a comprehensive evaluation of multiple modelling techniques has been undertaken. Conventional algebraic and probabilistic approaches to topic modelling, such as Latent Dirichlet Allocation (LDA)(Blei et al., 2003a), Non-Negative Matrix Factorization (NMF)(Lee and Seung, 1999), and Latent Semantic Indexing (LSI)(Hofmann, 1999), predominantly rely on the bag-of-words representation for documents, postulating that topics adhere to statistical distributions. Nevertheless, the bag-of-words representation exhibits a fundamental constraint, treating each document as a mere amalgamation of constituent words, neglecting the intrinsic semantic relationships among words. This inherent limitation necessitates a pivotal decision to exclude topic modelling algorithms contingent upon the bag-of-words framework. Additionally, variants of algebraic and probabilistic models confront challenges associated with stability and robustness, as documented in the literature(Abdelrazek et al., 2023). A comprehensive review of the literature delineates a contemporary paradigm that relies on document representation through an embedding approach. Neural topic models

leverage embeddings derived from pretrained deep learning models. An innovative approach proposed by Sia(Sia et al., 2020) capitalizes on embeddings, characterizing topics as principal themes or clusters of these embeddings. Historically, centroid-based clustering algorithms have been employed, a strategy also adopted by Sia. Furthermore, Sia's research underscores that neural topic modelling techniques that integrate embeddings and clustering exhibit noteworthy enhancements in terms of computational efficiency and resource utilization.

In the context of Sia's methodology, a notable constraint pertains to the assumption that topics are confined within spherical regions centered around cluster centroids. Nevertheless, this assumption may not accurately reflect real-world topic representations. To address this limitation, the present study advocates the adoption of the BERTopic(Grootendorst, 2022) framework. BERTopic offers a promising avenue for mitigating the aforementioned shortcomings by constructing topics founded on class-based variations of tf-idf. The schematic representation of the proposed procedural steps is elucidated in Figure 3.1. To initiate the process, it is imperative to commence with relevant preprocessing procedures. Subsequently, the workflow encompasses four pivotal stages:

- Document representation - Embeddings: Initially, documents are transformed into embeddings utilizing pretrained models.
- Document representation - Dimensionality reduction: Dimensionality reduction techniques are applied to encapsulate document embeddings within a lower-dimensional space.
- Clustering of Document Representations: The reduced-dimensional document representations are subjected to clustering algorithms to form cohesive clusters of documents based on their underlying semantic relationships.
- Topic Word Extraction and Topic Representation: Within each document cluster, a mechanism for extracting topic-related keywords and crafting comprehensive topic representations is employed. This step facilitates the coherent delineation and understanding of topics extracted from the corpus.

In effect, this methodological approach enhances the robustness and applicability of topic modelling techniques, addressing the limitations associated with Sia's methodology and aligning with the objectives of this research endeavour.

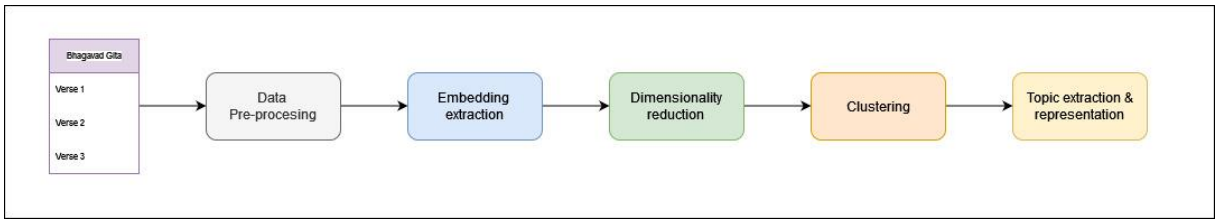

Figure 3.1 Process steps of proposed approach

3.4 Data pre-processing

In the context of this research endeavour, it is imperative to acknowledge that the initial dataset is presented in an unprocessed, raw textual format. Consequently, prior to commencing with the topic modelling exercise, an essential pre-processing phase is mandated to reconfigure the data into an optimal format for analysis. This methodological approach draws inspiration from the seminal work conducted by Chandra(Chandra and Kulkarni, 2022; Chandra and Ranjan, 2022). The delineated pre-processing steps encompass:

- Substitution of Archaic English References: The first step involves the identification and replacement of archaic English references, such as 'thou,' with modern English equivalents to enhance comprehensibility.
- Standardization of Names: Known references to Lord Krishna and Arjuna across various nomenclatures are systematically identified and unified under a single designated name for consistency.
- Lowercasing: To facilitate uniformity and mitigate case-related disparities, all textual data is converted to lowercase.
- Punctuation Removal: Extraneous punctuations are systematically removed from the textual corpus to ensure cleaner and more coherent content.
- Exclusion of Commentary and Redundant Text: Any non-essential commentary text or redundant content that extends beyond the scope of the translated verses is pruned to streamline the dataset.
- Stopword Elimination: Common stop words, which typically contribute minimal semantic value, are eliminated to enhance the dataset's relevance for subsequent analysis.

However, it is crucial to note that, as elucidated in Section 3.2, our proposed methodology is predicated on the utilization of document embeddings derived from pre-trained deep learning models. These embedding models are conventionally pre-trained on extensive corpora, leveraging contextual information to generate embeddings. Consequently, the traditional pre-

processing steps, such as punctuation removal and lowercasing, are deemed counterintuitive within this paradigm, as they may disrupt the integrity of the contextual information encoded within the embeddings. Nevertheless, certain pre-processing steps remain pertinent in preserving the semantic representations within the dataset. Specifically, the removal of archaic English references and the unification of various references to Lord Krishna and Arjuna under a single name are retained as they do not compromise the contextual integrity of the data and enhance its interpretability and relevance in the context of the proposed deep learning-based methodology.

3.5 Document representation

3.5.1 Embeddings

In the context of deep learning-based neural topic modeling methodologies, this research employs Transformer models, as introduced by Vaswani(Vaswani et al., 2017), to derive contextual embeddings serving as the foundation for our data representations. Of particular significance is the utilization of BERT, an extensively pre-trained language model developed by Devlin(Devlin et al., 2018), which has been trained on a substantial corpus and offers the versatility for fine-tuning in various downstream natural language processing (NLP) tasks, including classification and question-answering systems, among others. BERT's training regimen comprises two semi-supervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In the MLM task, a token within a sequence is masked, and the model is trained to predict the masked word, thereby promoting contextual understanding. In the NSP task, two consecutive sentences are presented, and the model is trained to discern the sequential relationships between them. This dual-task training approach enables BERT to capture rich contextual embeddings of words, establishing it as the state-of-the-art choice when contrasted with alternative representations such as GloVe(Pennington et al., 2014) and word2vec(Mikolov et al., 2013a). Depending on the granularity of textual blocks utilized, we have the flexibility to employ either BERTbase or BERTlarge for extracting contextual word embeddings. Expanding the application of BERT to encompass entire documents or sentences, we introduce the Sentence-BERT model, as proposed by Reimers(Reimers and Gurevych, 2019) . This model leverages Siamese networks and triplet network loss functions to generate embeddings that encode the semantic content of sentences. In a similar vein, Distlibert(Sanh et al., 2019) offers a streamlined variant of BERT, which achieves a reduced model size while preserving its performance characteristics. ROBERTa(Liu et al., 2019) represents an evolution of BERT, capitalizing on a comprehensive exploration of its hyperparameters to further

enhance its capabilities. In addition to these BERT-derived models, we also consider the Universal Sentence Encoder (USE) introduced by Cer et al. in 2018. The USE model produces sentence embeddings of fixed length, specifically 512 dimensions, utilizing a pooling layer to compute average embeddings across the constituent words within a sentence. This multifaceted ensemble of state-of-the-art models forms the bedrock of our approach for neural topic modeling. We will also leverage OpenAI embeddings (Neelakantan et al., 2022) as it has been trained on the largest corpora.

In the realm of contemporary natural language processing (NLP) methodologies, the utilization of advanced embedding techniques, such as BERTopic model and Top2vec, has underscored the ascendancy of neural embeddings when compared to conventional word vectors. These pre-trained embeddings offer distinct advantages in the form of contextually rich embeddings that exhibit enhanced semantic accuracy, an attribute particularly attributed to their derivation from models that have undergone extensive training on sizable corpora. Furthermore, empirical evidence substantiates the efficacy of BERT-based embeddings, as demonstrated in the works of Reimers and Gurevych (Reimers and Gurevych, 2020; Thakur et al., 2020). It is important to emphasize that these embeddings find their primary application in the domain of document clustering and are not primarily employed for the explicit purpose of topic representation.

3.5.2 Dimensionality reduction

In the context of this research, it is imperative to address the issue of high-dimensional embeddings, typically characterized by dimensions such as 384, 512, and the like. These dimensions pose a significant computational challenge during the clustering process, necessitating a reduction in dimensionality while preserving essential information and underlying patterns. The task of mitigating the curse of dimensionality is central to the methodology. Traditionally, dimensionality reduction methods such as t-SNE (t-distributed Stochastic Neighbor Embedding), PCA (Principal Component Analysis), and SVD (Singular Value Decomposition) have been widely employed for this purpose. However, recent investigations have revealed that UMAP (Uniform Manifold Approximation and Projection)(McInnes et al., 2018) exhibits superior performance, particularly when dealing with high-dimensional data, such as text embeddings that encompass intricate local and global relationships. As a result, UMAP has been chosen as the primary dimensionality reduction technique for this study, with PCA serving as the baseline for comparative analysis. Additionally, SVD will be employed in conjunction with these methods for a comprehensive

assessment of dimensionality reduction techniques. This selection process is guided by the need to effectively address the computational challenges posed by high-dimensional embeddings while preserving the structural integrity of the data.

3.6 Clustering

The fundamental premise underpinning topic modeling revolves around the concept of Clustering. It constitutes an unsupervised analytical technique wherein data elements are categorized on the basis of their similarity as assessed through quantitative measures. Depending on the specific algorithm employed and the criteria utilized to gauge similarity, Clustering methodologies, as expounded by Xu and Tian(Xu and Tian, 2015) can be classified into four primary categories: centroid-based, density-based, distribution-based, and hierarchical. In a broader context, density-based clustering algorithms have exhibited noteworthy applicability when dealing with the task of embedding data for the purpose of topic modeling. This is chiefly attributed to their reliance on mutual reachability distance as a similarity metric, thus facilitating the preservation of intricate interrelationships within the data. Consequently, DBSCAN emerges as a discernible and pragmatic choice for this endeavor. Nonetheless, within the realm of topic modeling, an additional imperative lies in the establishment of hierarchical relationships that ensure the cohesiveness of documents within a given cluster. In the pursuit of a preliminary proof-of-concept, we opt for the utilization of Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is deemed a compelling selection, as it aptly addresses the presence of outliers within the clustering structure. Furthermore, as a means of establishing a baseline for comparative analysis, the K-means clustering algorithm(Lloyd, 1982) will also be employed. It is noteworthy that, in the specific context of this research, K-means holds the potential to yield comparable performance outcomes.

3.7 Topic extraction and topic representation

In the context of this research methodology, the endeavor is to construct topic representations predicated on the contents of individual clusters, wherein each cluster is to be assigned a singular topic. The primary aim is to discern the distinctive attributes that differentiate one topic from another, based on the word distribution within their respective clusters. To achieve this objective, we propose a modification to the Term Frequency-Inverse Document Frequency (TF-IDF) metric, which traditionally quantifies the significance of a word with respect to a

document. In the BERTopic adaptation, extend TF-IDF is extended to enable the quantification of a term's significance within the context of a specific topic.

$$W_{t,d} = tf_{t,d} * \log\left(\frac{N}{df_t}\right) \quad (3.1)$$

The conventional TF-IDF methodology incorporates two key statistical measures: term frequency and inverse document frequency, as outlined by Joachims(Joachims, 1996) . As in Equation 3.1, term frequency, denoted as tf, encapsulates the frequency of a term 't' within a document 'd,' while inverse document frequency, gauges the information content a term furnishes to a document. IDF is computed as the logarithm of the total number of documents 'N' in a corpus divided by the count of documents containing term 't.' In this research, we extend and generalize the TF-IDF framework to operate within the confines of document clusters. Initially, we treat all documents within a given cluster as a singular entity by concatenating them together. Subsequently, we adapt the TF-IDF formulation to accommodate this new representation in Equation 3.2, effectively shifting our focus from individual documents to clusters.

$$W_{t,c} = tf_{t,c} * \log(1 + \frac{A}{tf_t}) \quad (3.2)$$

Specifically, we introduce cluster-based TF-IDF, denoted as c-TF-IDF, in which term frequency (TF) signifies the frequency of term 't' within a cluster 'c'—now represented as a composite document formed by merging all documents in the cluster. Concurrently, inverse document frequency (IDF) is replaced with inverse class frequency (ICF) to quantify the information content of term 't' with respect to a cluster. ICF is computed by taking the logarithm of the average number of words per cluster 'A' divided by the frequency of term 't' across all clusters, with an additional '1' added to the denominator to ensure positive values.

The adoption of this class-based TF-IDF procedure enables the modeling of term importance within clusters, thus facilitating the generation of topic-word distributions for each cluster of documents. Ultimately, by employing an iterative process, we amalgamate the c-TF-IDF representations of the least common topics with their most similar counterparts. This iterative merger operation serves to reduce the number of topics to a user-specified value, a pivotal step in our analytical framework. To further refine topic modelling representations, BERTopic framework, provides two optional methods – Maximum Marginal Relevance (MMR) and BERT based KeyBERTInspired.

3.8 Evaluation

In the context of algorithmic investigations, the present study embarks upon the exploration of topics generated through unsupervised methodologies, specifically focusing on the distribution of words within each identified topic. The inherent subjectivity entailed in the interpretation of these topics underscores the challenging nature of this endeavor. It is pertinent to note that, in the realm of topic modeling, a universally accepted gold standard for evaluation remains elusive. Nevertheless, the evolution of diverse paradigms in response to various applications has engendered distinct evaluative approaches. When evaluating the efficacy of topic modeling methodologies, two overarching perspectives emerge as paramount. The first perspective, referred to as result-oriented domain-based evaluation, hinges on the interpretability and distinctiveness of the topics generated. This necessitates the engagement of domain experts to undertake the evaluation task, given their specialized knowledge. In this context, the establishment of evaluation metrics capable of gauging the convergence of assessments across experts is imperative. However, it is pertinent to acknowledge that, due to constraints imposed by time considerations within the scope of this research, the feasibility of engaging domain experts for evaluation is regrettably beyond the purview of this study. The second evaluative perspective delves into the inner workings of the topic modeling model itself. Abdelrazek et al. (Abdelrazek et al., 2023) have identified a comprehensive suite of evaluation metrics encompassing dimensions such as quality, interpretability, stability, diversity, efficiency, and flexibility. Quality is measured by is perplexity, which serves as a foundational measure for assessing model performance based on the information gain derived from predicted outcomes. It is crucial to note that perplexity is contingent upon the vocabulary specific to each corpus and the number of topics, rendering it unsuitable for cross-model benchmarking.

The most prevalent evaluative criterion within the domain of topic modeling revolves around the interpretability of latent themes. The fundamental premise is that a well-constructed topic model should yield coherent representations within each topic. Coherence, measured through various techniques, has been shown(Newman et al., 2010; Mimno et al., 2011) to align closely with human judgment in the interpretation of topic models. Analogous to clustering exercises, which examine both intra-topic and inter-topic distances, an analogous perspective on topic coherence is that of topic diversity. This metric gauge the extent to which distinct thematic elements are faithfully represented across topics. Given that each topic can be conceptualized as a vector of words and can subsequently be transformed into topic embeddings, the calculation of similarity and dissimilarity metrics facilitates a nuanced understanding of topic diversity.

Cohesion and diversity evaluations, when considered together, yield a comprehensive evaluation framework that elucidates and interprets the latent themes encapsulated within the generated topics. Noteworthy is the inherent issue of instability and inconsistency in topic modeling outputs, arising from the stochastic nature of these algorithms. Measures of stability are instrumental in assessing the degree of similarity exhibited by generated topics across multiple iterations. Lastly, the computational demands associated with topic modeling, owing to the utilization of distributions and embeddings, render it computationally resource-intensive. Computational time, encompassing model training and inference times, serves as a tangible measure of this computational burden.

In the context of our proof-of-concept research, the principal objective is the identification of key thematic elements. Consequently, the selection of an appropriate evaluation metric gravitates towards interpretability. Hence, the primary evaluation metric adopted for this research is topic coherence, as expounded upon by Newman et al.(Newman et al., 2010), employing the pointwise mutual information (PMI) metric. In this context, topics are construed as sets of the top k words, typically k=10, and pairwise similarity is computed through PMI, subsequently aggregated. In addition to coherence, we place considerable emphasis on the metric of diversity, measured using cosine similarity, to ensure a well-rounded and judicious evaluation of the topic modeling outcomes.

3.9 Summary

The identification of a suitable dataset marks a pivotal milestone in our research endeavour. In conjunction with this, we have meticulously selected an appropriate topic modelling approach, founded upon the BERTopic model and its constituent steps. This discerning choice has been made with utmost consideration, as it forms the bedrock of our methodological framework for the forthcoming investigation.

CHAPTER 4

ANALYSIS, DESIGN, EXPERIMENTS

4.1 Introduction

Chapter 4 provides a comprehensive exposition of the outcomes stemming from the preliminary investigative examination. Subsequently, a meticulous elucidation of the experimental framework and the methodological strategies for optimizing hyperparameters will be proffered.

4.2 Dataset preparation

In the third chapter, Methodology, we delineated the dataset encompassed within the purview of this analysis. An initial scrutiny of the dataset was undertaken to conclude the definitive preprocessing steps to be employed. As illustrated in Figure 4.1, we present a snapshot of the dataset, which consists of seven distinct columns. The original file has 700 verses, each row corresponding to one verse across 18 chapters of Bhagavad Gita.To streamline the dataset for further analysis, an initial subsetting was performed, reducing it to a relevant and manageable format comprising three columns:

- verse_number
- translation_in_english
- meaning_in_english

Upon meticulous examination of the English translation examples, it became evident that the current version of the dataset does not incorporate archaic language. Consequently, there is no imperative need for preprocessing measures aimed at substituting archaic words. Furthermore, it is noteworthy that the translator has thoughtfully maintained the consistency of names, specifically those of Krishna and Arjuna, despite the Sanskrit version utilizing various reference names. This fortuitous alignment simplifies our preprocessing efforts. Another salient observation is the presence of redundancy within the dataset. Notably, certain rows with multiple verses are repetitive, such as those corresponding to 'Chapter 1, verse 4-6.' Given that topic modelling is intrinsically sensitive to word frequencies, we have taken measures to deduplicate the dataset, resulting in the retention of only a single instance where multiple verses were amalgamated into a solitary row. For all subsequent analyses, the English translation shall serve as the primary variable of interest, facilitating both exploratory investigations and topic modelling endeavours. Moreover, for the purpose of conceptual validation, similar analytical procedures will be applied to the meaning and commentary provided by the translator.

verse_number	verse_in_sanskrit	sanskrit_verse_transliteration	translation_in_english	meaning_in_english	translation_in_hindi	meaning_in_hindi
Chapter 1, Verse 1	धृतराष्ट्र उवाच |धर्मक्षेत्रे कुरुक्षेत्रे समवेता युयुत्सवः |मामकाः पाण्डवाश्चैव किमकुर्वत सञ्जय...	dhṛitarāśhtra uvācha dharma-kṣhetre kuru-kṣhetre samavetā yuyutsavaḥ māmakāḥ pāṇḍavāśhchaiva kim...	Dhritarashtra said: O Sanjay, after gathering on the holy field of Kurukshetra, and desiring to ...	The two armies had gathered on the battlefield of Kurukshetra, well prepared to fight a war that...	धृतराष्ट्र ने कहाः हे संजय! कुरुक्षेत्र की पवित्र भूमि पर युद्ध करने की इच्छा से एकत्रित होने के...	राजा धृतराष्ट्र जन्म से नेत्रहीन होने के अतिरिक्त आध्यात्मिक ज्ञान से भी वंचित था। अपने पुत्रों ...
Chapter 1, Verse 2	सञ्जय उवाच ।दृष्ट्वा तु पाण्डवानीकं व्यूढं दुर्योधनस्तदा । आचार्यमुपसङ्गम्य राजा वचनमब्रवीत् ।। 2।।	sañjaya uvācha dṛiṣhṭvā tu pāṇḍavānīkaṁ vyūḍhaṁ duryodhanastadā āchāryamupasaṅgamya rājā vachana...	Sanjay said: On observing the Pandava army standing in military formation, King Duryodhan approa...	Sanjay understood Dhritarashtra's concern, who wanted an assurance that the battle would eventua...	संजय ने कहाः हे राजन्! पाण्डवों की सेना की व्यूहरचना का अवलोकन कर राजा दुर्योधन ने अपने गुरु द्र...	धृतराष्ट्र इस बात की पुष्टि करना चाहता था कि क्या उसके पुत्र अब भी युद्ध करने के उत्तरदायित्व का...
Chapter 1, Verse 3	पश्यैतां पाण्डुपुत्राणामाचार्य महतीं चमूम् ।व्यूढां द्रुपदपुत्रेण तव शिष्येण धीमता ।। 3।।	paśhyaitāṁ pāṇḍu-putrāṇām āchārya mahatīṁ chamūm vyūḍhāṁ drupada-putreṇa tava śhiṣhyeṇa dhīmatā	Duryodhan said: Respected teacher!Behold the mighty army of the sons of Pandu, so expertly array...	Duryodhana asked Dronacharya to look at the skillfully arranged military phalanx of the Pandava ...	दुर्योधन ने कहाः पूज्य आचार्य! पाण्डु पुत्रों की विशाल सेना का अवलोकन करें, जिसे आपके द्वारा प्र...	दुर्योधन एक कुशल कूटनीतिज्ञ के रूप में अपने सेनापति गुरु द्रोणाचार्य द्वारा अतीत में की गई भूल क...
Chapter 1, Verse 4-6	अत्र शूरा महेष्वासा भीमार्जुनसमा युधि युयुधानो विराटश्च द्रुपदश्च महारथ: || 4||धृष्टकेतुश्चेकित...	atra śhūrā maheṣhvāsā bhīmārjuna-samā yudhiyuyudhāno virāṭaśhcha drupadaśhcha mahā-rathaḥdhṛiṣht...	Behold in their ranks are many powerful warriors, like Yuyudhan, Virat, and Drupad, wielding mig...	Due to his anxiety, the Pandava army seemed much larger to Duryodhan than it actually was. He ha...	यहाँ इस सेना में भीम और अर्जुन के समान बलशाली युद्ध करने वाले महारथी युयुधान, विराट और द्रुपद जै...	अपने सम्मख संकट को मंडराते देखकर दुर्योधन को पाण्डवों द्वारा एकत्रित की गयी सेना वास्तविकता से अ...
Chapter 1, Verse 4-6	अत्र शूरा महेष्वासा भीमार्जुनसमा युधि युयुधानो विराटश्च द्रुपदश्च महारथ: || 4||धृष्टकेतुश्चेकित...	atra śhūrā maheṣhvāsā bhīmārjuna-samā yudhiyuyudhāno virāṭaśhcha drupadaśhcha mahā-rathaḥdhṛiṣht...	Behold in their ranks are many powerful warriors, like Yuyudhan, Virat, and Drupad, wielding mig...	Due to his anxiety, the Pandava army seemed much larger to Duryodhan than it actually was. He ha...	यहाँ इस सेना में भीम और अर्जुन के समान बलशाली युद्ध करने वाले महारथी युयुधान, विराट और द्रुपद जै...	अपने सम्मख संकट को मंडराते देखकर दुर्योधन को पाण्डवों द्वारा एकत्रित की गयी सेना वास्तविकता से अ...
Chapter 1, Verse 4-6	अत्र शूरा महेष्वासा भीमार्जुनसमा युधि युयुधानो विराटश्च द्रुपदश्च महारथ: || 4||धृष्टकेतुश्चेकित...	atra śhūrā maheṣhvāsā bhīmārjuna-samā yudhiyuyudhāno virāṭaśhcha drupadaśhcha mahā-rathaḥdhṛiṣht...	Behold in their ranks are many powerful warriors, like Yuyudhan, Virat, and Drupad, wielding mig...	Due to his anxiety, the Pandava army seemed much larger to Duryodhan than it actually was. He ha...	यहाँ इस सेना में भीम और अर्जुन के समान बलशाली युद्ध करने वाले महारथी युयुधान, विराट और द्रुपद जै...	अपने सम्मख संकट को मंडराते देखकर दुर्योधन को पाण्डवों द्वारा एकत्रित की गयी सेना वास्तविकता से अ...
Chapter 1, Verse 7	अस्माकं तु विशिष्टा ये तान्निबोध द्विजोत्तम |नायका मम सैन्यस्य संज्ञार्थं तान्ब्रवीमि ते || 7||	asmākaṁ tu viśhiṣhṭā ye tānnibodha dwijottama nāyakā mama sainyasya sanjñārthaṁ tānbravīmi te	O best of Brahmins, hear too about the principal generals on our side, who are especially qualif...	Dronacharya was a teacher of military science and not really a warrior. However, he was on the b...	हे ब्राह्मण श्रेष्ठ! हमारे पक्ष की ओर के उन सेना नायकों के संबंध में भी सुनो, जो सेना को संचालित...	यहाँ कौरव सेना के प्रधान सेनापति द्रोणाचार्य को दुर्योधन ने 'द्विजोत्तम' (द्विजन्मा में श्रेष्ठ ...

Figure 4.1 Snapshot of original dataset

4.3 Exploratory analysis

The Bhagavad Gita, an ancient philosophical text, represents a dialogic exchange between the revered deities Lord Krishna and the warrior Arjuna. This venerable scripture is structured into 18 chapters and consists of a total of 701 verses. A graphical representation of the verse distribution by chapter is presented in Figure 4.2, elucidating the quantitative disparities among these chapters. Upon careful examination of Figure 4.2, it becomes evident that Chapter 18 boasts the highest verse count, comprising a substantial 78 verses, thereby claiming the distinction of being the most extensive chapter within the Bhagavad Gita. Following closely behind, Chapter 2 emerges as the second most voluminous, incorporating 72 verses, while Chapter 11 contains 55 verses, securing the third position in terms of verse count. In stark contrast, Chapter 12 is the most concise chapter, with a mere 20 verses, signifying its brevity in the context of this profound text. This distribution of verses across chapters unveils significant textual disparities and invites further investigation into potential thematic and structural nuances within the Bhagavad Gita.

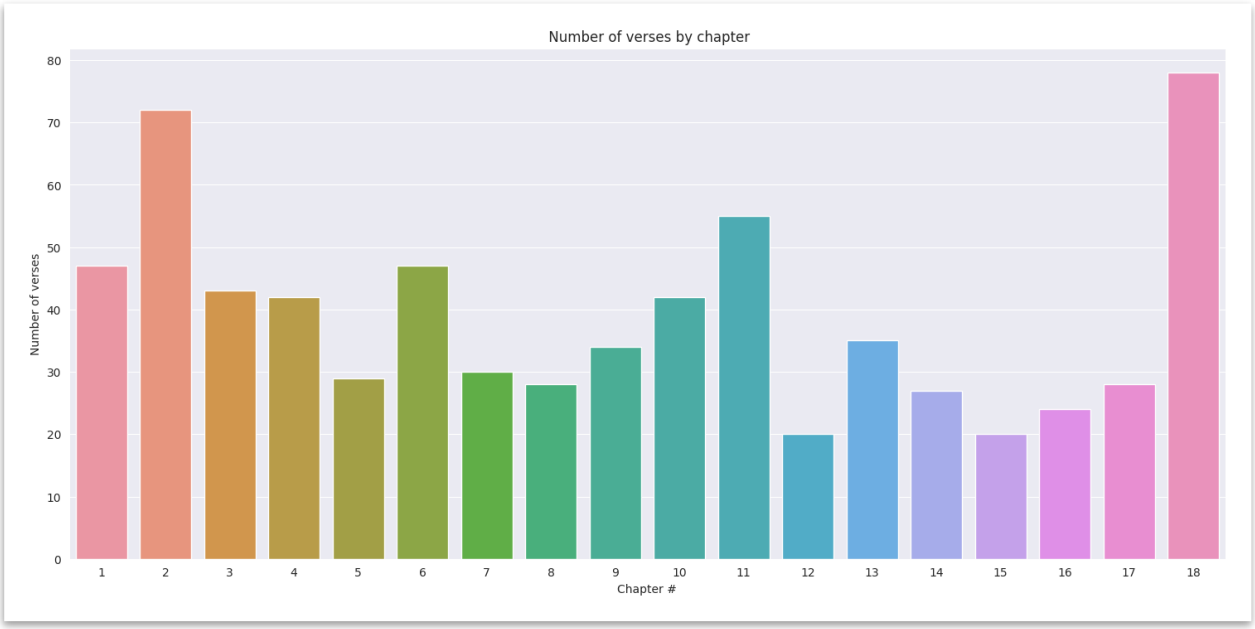

Figure 4.2 Number of verses by chapter

Figure 4.3, displayed herewith, provides a comprehensive analysis of word distribution across the various chapters under investigation. In consonance with our prior observations regarding verse quantities, Chapter 18 prominently emerges as the most extensive in terms of word count, encompassing a substantial 2,223 words. Following closely in second place, Chapter 2 commands a word count of 2,133, while Chapter 11 takes the third position with a total of 1,736 words. In stark contrast, Chapter 12 exhibits the most concise textual composition, comprising a mere 577 words. A noteworthy facet, meriting scrutiny, pertains to the distribution of word counts in the verse commentary associated with each chapter. It is discernible that the pattern deviates from that observed in the case of verse count or overall word count. This intriguing deviation underscores the pivotal importance of Chapters 2 and 18, which boast not only substantial volumes of content but also profound layers of meaning. In stark contrast, the remaining chapters, despite their brevity, convey abundant messages, thereby attesting to the profound depth and significance inherent in this sacred scripture. Indeed, such a nuanced examination of the Bhagavad Gita underscores the magnanimity of its intellectual and spiritual offerings.

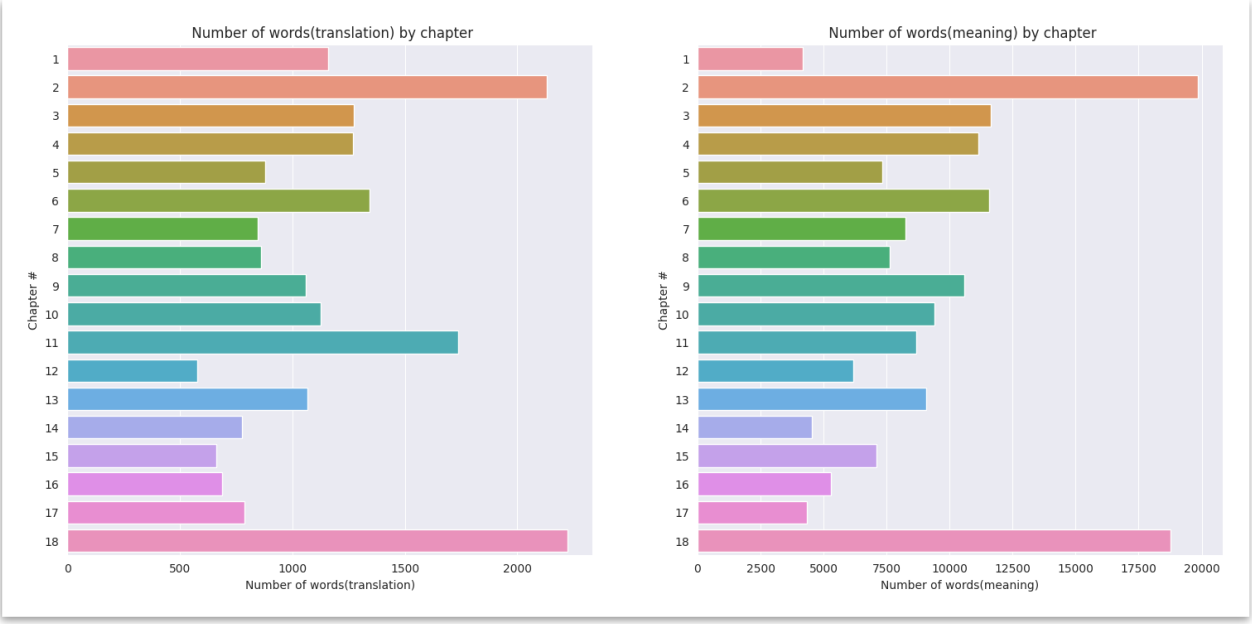

Figure 4.3 Number of words by chapter

N-grams, represent a foundational exploratory analysis technique commonly employed to swiftly gain insights into the distributional characteristics of textual data. This methodology is particularly adept at elucidating the textual landscape, thereby facilitating subsequent endeavours in feature engineering and model development. Figure 4.4 illustrates the top 15 Unigrams, which, on their own, present a challenging interpretive task, given their susceptibility to forming constituents of multi-word contexts. For instance, among these top 15 Unigrams, words such as "arjun," "said," "lord," and "god" fail to convey any discernible thematic focus in isolation. In contrast, words like "supreme," "knowledge," "mind," "beings," "material," "soul," and "nature" exhibit a distinct semantic coherence, as they are intrinsically linked with at least one other word, thereby affording a lucid conceptual framework, as exemplified by phrases like "supreme lord" and "mind & material." Notably, this analogous pattern is conspicuously evident within the unigram distribution of meaning and commentary as well.

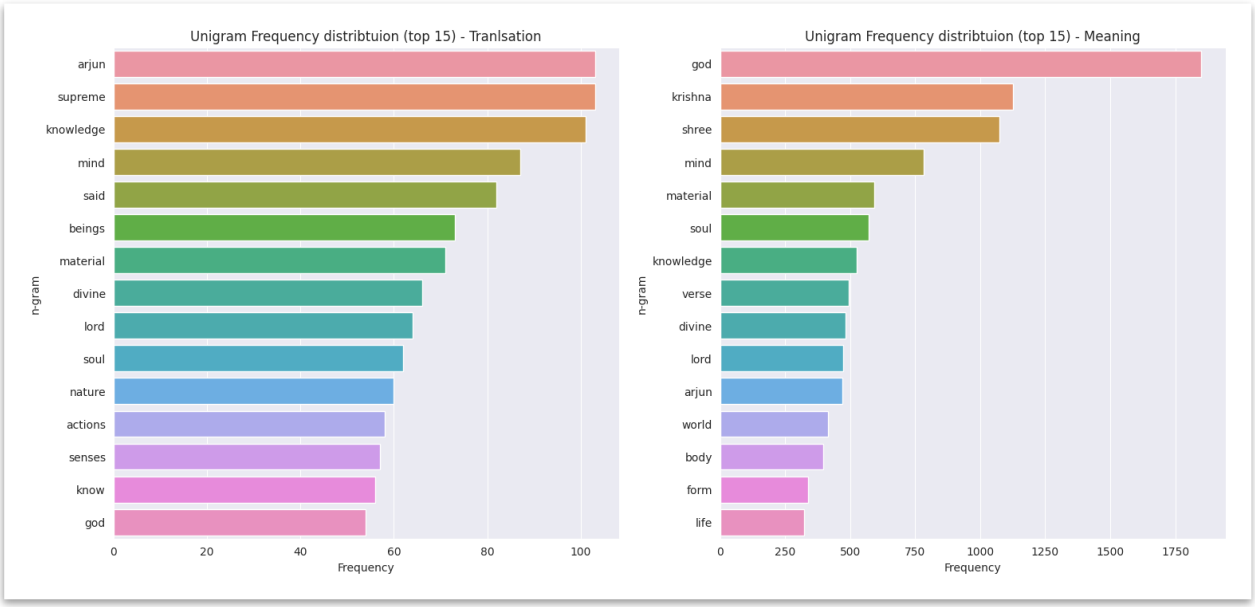

Figure 4.4 Unigram (top 15) frequency

Figure 4.5 portrays the distribution of bigrams, while Figure 4.6 provides insights into the trigram distribution within our dataset. It is within these bi- and trigram structures that we begin to discern meaningful patterns and observations. Nevertheless, it is worth noting that certain data points, such as the occurrences of phrases like "shree krishna," "shree krishna said," and "arjuna said," exhibit limited interpretability in isolation. Conversely, our analysis uncovers compelling combinations such as "passion mode," "goodness mode," and "ignorance mode," which correspond to the three gunas or characteristic modes of sattva, rajas, and tamas. Additionally, a distinct thematic thread emerges with phrases like "fruits action" and "perform prescribed duties," reflecting the principles of karma and dharma as expounded in the Bhagavad Gita. These observations underscore the teaching that individuals should focus on fulfilling their duties and actions without harbouring expectations of outcomes. Another discernible theme revolves around "sacrifice charity penance," further enriching the semantic landscape of our textual analysis.

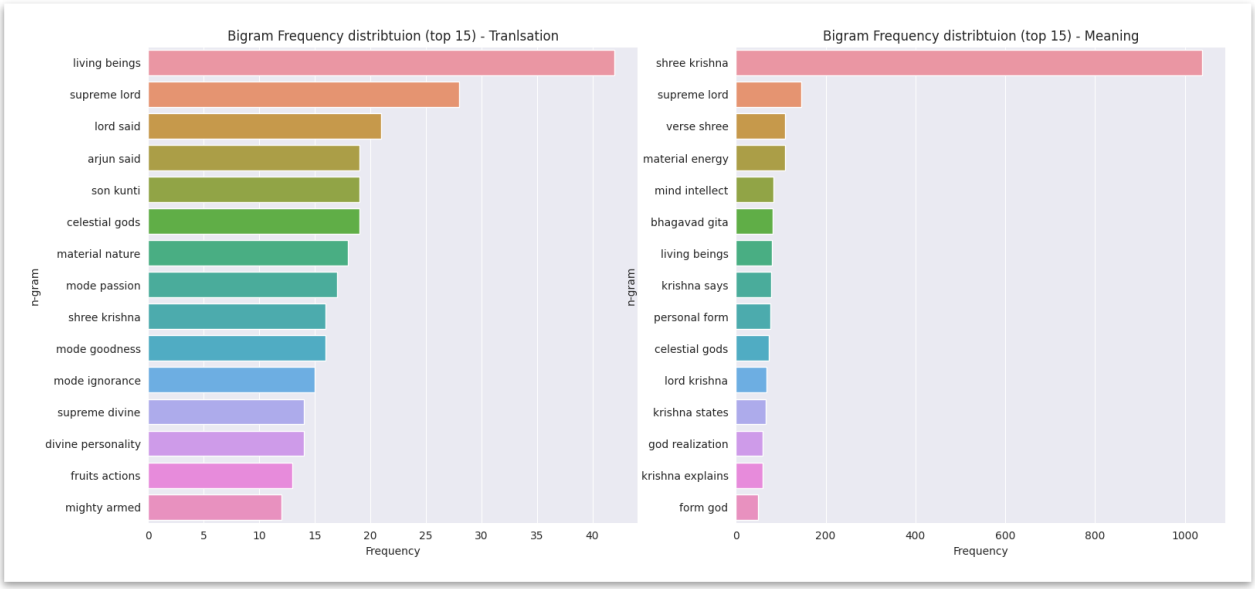

Figure 4.5 Bigram (top 15) frequency

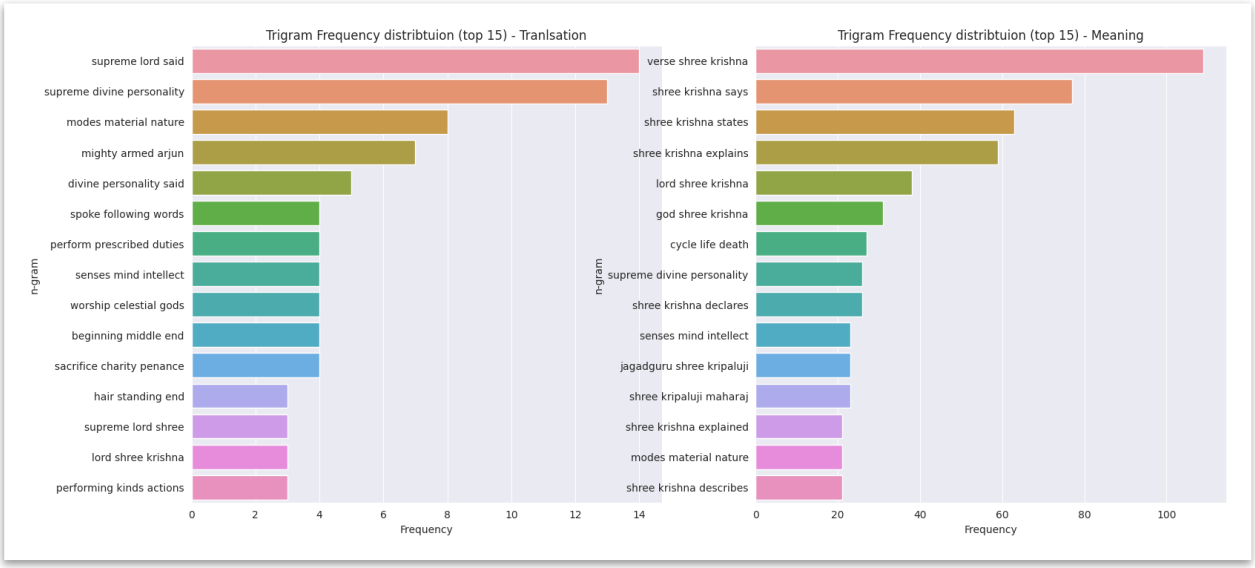

Figure 4.6 Trigram (top 15) frequency

The word clouds in Figure 4.7 and 4.8 signify the context of the Bhagavad Gita.

Figure 4.7 Word cloud - translation

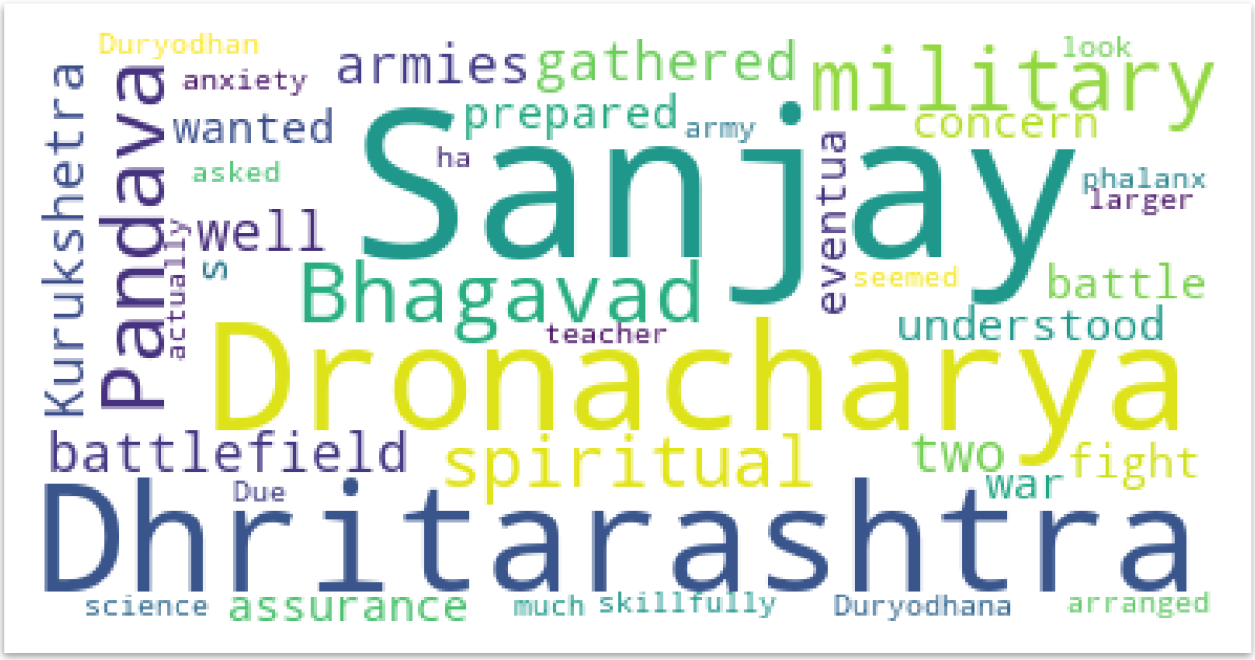

Figure 4.8 Word cloud - meaning

4.4 Experiment design of modelling process

In Chapter 3 of this thesis, Methodology was comprehensively expounded, elucidating the meticulous selection of candidate models tailored to the discrete sub-stages within the overarching topic modelling framework founded upon the BERTopic methodology. It is paramount to underscore the autonomy of each constituent process within this framework.

Nevertheless, the integration of embeddings and dimensionality reduction serves the pivotal purpose of ascertaining the optimal document representation, while the subsequent clustering phase is instrumental in identifying the most suitable document clusters. The ensuing phase, characterized by the topic representation model, deviates from the utilization of embeddings or vectors from the low-dimensional space. Instead, it employs a class-based Term Frequency-Inverse Document Frequency (tf-idf) approach to generate representative words for each topic cluster. Thus, the experimental design meticulously adheres to a systematic multi-step approach, as visually depicted in Figure 4.9.

Step 1 of this methodological cascade involves conducting experiments to discern the optimal document cluster representations. In pursuit of this objective, we execute multiple iterations of the topic model, exploring a plethora of permutations comprising various combinations of embeddings, dimensionality reduction techniques, and clustering methods (from Table 4.1). The aim is to discern and select the most effective document cluster representation. Proceeding to Step 2, we leverage the optimal document cluster representation obtained in the prior step to embark upon a series of experiments aimed at identifying the most efficacious method for each facet of the topic modelling framework. Subsequently, in Step 3, we undertake the crucial task of hyperparameter tuning, fine-tuning the parameters that govern the model's behaviour. The culminating phase entails the execution of the topic model based on the refined hyperparameters, followed by a comprehensive analysis and interpretation of the generated topics. Throughout all phases of experimentation, the evaluation will pivot around the primary metric of topic coherence, complemented by an auxiliary metric, topic diversity, which serves as a supplementary validation tool to address any scenarios that necessitate subjective judgment. For Steps 1 and 2, the selection of hyperparameters will be fixed and adhere to established standards specific to each of the methods, informed by prior work conducted by Chandra(Chandra and Kulkarni, 2022; Chandra and Ranjan, 2022) on analogous textual data derived from spiritual scriptures.

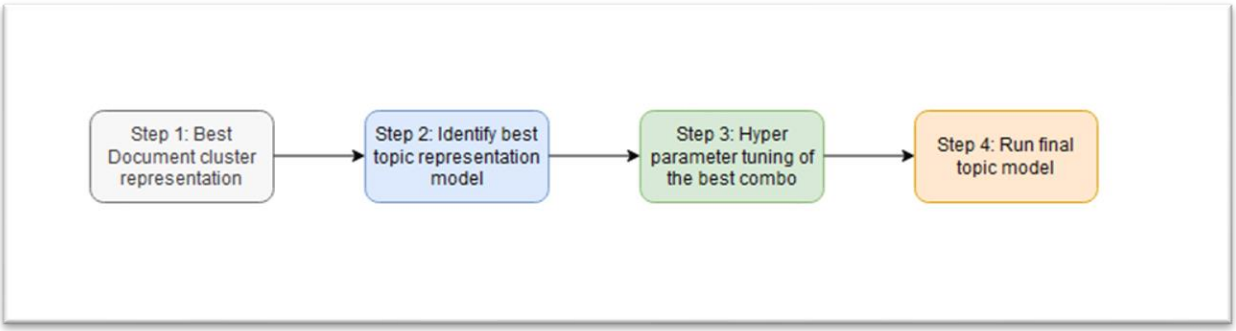

Figure 4.9 Experiment design

Table 4.1 List of Methods

Proposed approach step	Candidate models for experimentation
Extract embedding	Sentence Transformer Universal Sentence Encoder (USE) Roberta base Distilbert OpenAI embeddings (text-embedding-ada-002)
Dimensionality reduction	Principal Component Analysis (PCA) UMAP tSVD
Clustering	HDBSCAN kMeans
Representation model	MMR KeyBERTinspired

4.5 Hyperparameter design

After the determination of the optimal combination of techniques encompassing embeddings, dimensionality reduction, clustering, and topic representation, our research proceeds to the fine-tuning of hyperparameters for these selected methodologies. The ensuing Table 4.2 delineates the requisite hyperparameters under consideration. A comprehensive array of experiments shall be undertaken, encompassing all conceivable combinations, with model performance evaluated primarily based on the metric of topic coherence, supplemented by the secondary metric of diversity.

Table 4.2 List of parameters

Proposed approach step	Candidate models for experimentation	Parameter/Hyperparameter
Overall topic model		Number of topics (5, 7,10,15,20,30)
Extract embedding	Sentence Transformer Universal Sentence Encoder (USE) Roberta base Distilbert OpenAI embeddings (text-embedding-ada-002)	None None None None None
Dimensionality reduction	Principal Component Analysis (PCA) UMAP tSVD	N_componenets (3 to 20) N_componenets (3 to 20, fixed at 15) N_componenets (3 to 20)
Clustering	HDBSCAN kMeans	N_clusters N_clusters
Representation model	MMR KeyBERTinspired	kb_diversity (0.05 to 0.5) none

4.9 Summary

A robust experimental design has been meticulously employed in this study. The ensuing exploratory findings furnish perceptive insights into the Bhagavad Gita, aligning harmoniously with pre-established knowledge in the field.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

Chapter 5 offers a comprehensive exposition of the ultimate outcomes derived from the experimental investigations conducted across various modelling stages. Additionally, it furnishes an in-depth exposition of the final topic model and engages in a thorough discussion of the resultant findings. Furthermore, this chapter diligently presents the identified limitations of the study for critical consideration.

5.2 Results from experiments

In accordance with the experiment design elucidated in Chapter 4, a series of experimental runs were conducted to address the sequential phases of Step 1 and Step 2. During these experimental iterations, particular attention was given to the parameterization of the number of topics, which underwent meticulous tuning, encompassing the discrete values of 5, 7, 10, 15, 20, and 30. Substantive insights have been gleaned from the outcomes of these endeavours. Primarily, a consistent pattern emerged across all permutations and combinations, indicating that within the spectrum of topic representation models under scrutiny, MMR consistently outperformed its counterpart, KeyBERTInspired. In order to maintain a concise and manageable presentation of the findings within the constraints of this document, a strategic decision was made to selectively focus on iterations where MMR served as the designated topic representation model. Consequently, our ensuing analysis is primarily centered on these filtered results.

Upon careful examination of the outcomes, a discernible pattern emerges, highlighting an optimal combination of model components, namely:

- Embedding model: Sentence transformer
- Dimensionality reduction technique: PCA
- Clustering algorithm: kmeans
- Representation mode: MMR

Remarkably, this amalgamation of components yields a coherence score of 0.6, significantly surpassing the lower cutoff thresholds of 0.3 and 0.4, which are commonly considered as benchmarks within the context of coherence evaluation. It is noteworthy to mention that the

compatibility and synergy between PCA and kmeans clustering have been corroborated in the study by Abdelrazek et al.(Abdelrazek et al., 2023), affirming the judicious selection of these components in our model configuration.

Table 5.1 Results from experiments

Embedding	Dimension reduction	Clustering	Coherence	Diversity	Topics
sentence_transformer	pca	kmeans	0.6013	0.3009	10
sentence_transformer	umap	kmeans	0.58	0.3474	20
sentence_transformer	pca	kmeans	0.5716	0.2889	7
use	umap	hdbscan	0.5696	0.352	7
roberta	umap	hdbscan	0.5596	0.0008	30
sentence_transformer	pca	kmeans	0.5565	0.3425	15
use	pca	kmeans	0.5551	0.302	7
sentence_transformer	tSVD	kmeans	0.5543	0.3059	7
use	umap	hdbscan	0.554	0.2416	5
sentence_transformer	pca	kmeans	0.554	0.3479	30
use	tSVD	hdbscan	0.5537	0.1572	7
use	umap	hdbscan	0.5491	0.3417	20
use	umap	hdbscan	0.5458	0.3481	15
sentence_transformer	tSVD	kmeans	0.5418	0.3363	10
sentence_transformer	tSVD	kmeans	0.5382	0.3441	20
sentence_transformer	tSVD	kmeans	0.5378	0.3489	30
use	umap	kmeans	0.5363	0.3575	20
use	umap	hdbscan	0.5363	0.2563	10
sentence_transformer	umap	kmeans	0.5346	0.3242	15
use	umap	kmeans	0.5341	0.3579	15
use	umap	hdbscan	0.534	0.3195	30
use	pca	kmeans	0.5335	0.3569	15
use	umap	hdbscan	0.5332	0.3351	20
sentence_transformer	umap	kmeans	0.532	0.3226	30
use	pca	kmeans	0.5318	0.3013	5
sentence_transformer	tSVD	kmeans	0.5315	0.3086	5
sentence_transformer	tSVD	kmeans	0.5314	0.3488	15
roberta	umap	hdbscan	0.531	0.0008	20
use	pca	kmeans	0.5302	0.3581	15
use	umap	kmeans	0.5299	0.3552	30
use	tSVD	kmeans	0.5294	0.2907	7
use	tSVD	kmeans	0.5285	0.3442	20
use	umap	kmeans	0.5239	0.3061	5
sentence_transformer	tSVD	kmeans	0.5236	0.3497	20
sentence_transformer	umap	kmeans	0.5209	0.3255	20
use	umap	hdbscan	0.5193	0.3714	30

sentence_transformer	pca	kmeans	0.5181	0.3331	30
sentence_transformer	umap	kmeans	0.5144	0.3091	10
sentence_transformer	umap	kmeans	0.5138	0.2712	7
sentence_transformer	tSVD	kmeans	0.5108	0.3453	10
distilbert	tSVD	kmeans	0.5108	0.03	10
sentence_transformer	tSVD	hdbscan	0.51	0.1366	30
use	pca	kmeans	0.5091	0.3211	10
use	pca	kmeans	0.5084	0.3516	20
use	tSVD	kmeans	0.5077	0.3576	30
use	umap	kmeans	0.5075	0.3361	10
use	tSVD	kmeans	0.5063	0.3497	10
roberta	umap	hdbscan	0.5053	0.0008	15
roberta	umap	hdbscan	0.5044	0.0007	10
use	umap	kmeans	0.5009	0.357	20
use	tSVD	kmeans	0.4992	0.346	30
distilbert	tSVD	kmeans	0.4992	0.0358	30
use	umap	hdbscan	0.4982	0.3474	15
sentence_transformer	pca	kmeans	0.4979	0.316	7
roberta	pca	kmeans	0.4978	0.0009	30
roberta	tSVD	kmeans	0.497	0.0011	15
use	tSVD	kmeans	0.4961	0.3577	20
use	pca	kmeans	0.4961	0.334	10
sentence_transformer	pca	kmeans	0.4939	0.301	5
use	pca	kmeans	0.4932	0.3626	30
roberta	tSVD	kmeans	0.4926	0.0011	30
use	umap	kmeans	0.4923	0.3161	7
use	tSVD	kmeans	0.4922	0.3481	15
sentence_transformer	pca	kmeans	0.4916	0.3265	10
sentence_transformer	umap	kmeans	0.4903	0.3231	15
sentence_transformer	umap	kmeans	0.4902	0.3329	30
sentence_transformer	tSVD	kmeans	0.4901	0.3388	15
sentence_transformer	tSVD	hdbscan	0.4874	0.133	30
roberta	tSVD	kmeans	0.4871	0.0009	20
sentence_transformer	tSVD	kmeans	0.4862	0.2795	5
sentence_transformer	tSVD	kmeans	0.4849	0.3438	30
sentence_transformer	umap	hdbscan	0.4848	0.1503	5
roberta	pca	kmeans	0.4843	0.0009	15
roberta	umap	kmeans	0.4839	0.0007	15
roberta	umap	kmeans	0.4835	0.0007	30
roberta	umap	hdbscan	0.483	0.0007	7
roberta	umap	kmeans	0.4828	0.0008	10
roberta	pca	kmeans	0.482	0.0011	7
sentence_transformer	pca	hdbscan	0.4818	0.1573	5
sentence_transformer	umap	hdbscan	0.4811	0.1643	30
roberta	umap	kmeans	0.4804	0.0008	7

sentence_transformer	pca	kmeans	0.4801	0.345	15
sentence_transformer	pca	kmeans	0.479	0.3384	10
roberta	umap	hdbscan	0.4786	0.0007	5
roberta	tSVD	kmeans	0.4773	0.0011	5
sentence_transformer	umap	kmeans	0.4769	0.2798	5
sentence_transformer	umap	kmeans	0.4768	0.3171	7
use	pca	kmeans	0.4755	0.3207	7
use	tSVD	kmeans	0.4746	0.3489	15
roberta	umap	kmeans	0.4743	0.0007	20
use	umap	hdbscan	0.4737	0.2901	10
roberta	tSVD	kmeans	0.4721	0.0011	7
sentence_transformer	pca	kmeans	0.4702	0.3515	20
use	tSVD	kmeans	0.47	0.2403	5
use	umap	hdbscan	0.4698	0.2278	5
distilbert	pca	kmeans	0.4694	0.0345	15
sentence_transformer	tSVD	kmeans	0.4688	0.3027	7
use	umap	kmeans	0.4682	0.3554	30
distilbert	pca	kmeans	0.4669	0.0329	10
use	pca	kmeans	0.4657	0.3499	20
roberta	pca	kmeans	0.4655	0.0011	5
use	umap	kmeans	0.4638	0.3203	7
distilbert	tSVD	kmeans	0.4636	0.0354	15
use	tSVD	hdbscan	0.4597	0.1511	15
use	umap	hdbscan	0.4589	0.2818	7
use	tSVD	kmeans	0.4588	0.3254	10
sentence_transformer	umap	hdbscan	0.4561	0.2507	15
roberta	tSVD	kmeans	0.455	0.001	10
use	umap	kmeans	0.4539	0.355	15
roberta	umap	kmeans	0.453	0.0009	5
distilbert	umap	kmeans	0.4519	0.0338	30
distilbert	umap	kmeans	0.4505	0.0342	15
distilbert	umap	kmeans	0.4504	0.0341	20
use	tSVD	kmeans	0.45	0.2973	7
sentence_transformer	umap	hdbscan	0.4494	0.1561	20
sentence_transformer	umap	hdbscan	0.4455	0.1561	7
use	umap	kmeans	0.4441	0.326	10
sentence_transformer	umap	kmeans	0.4431	0.3306	20
distilbert	pca	kmeans	0.4427	0.0262	7
use	pca	kmeans	0.4422	0.2683	5
roberta	pca	kmeans	0.4412	0.001	10
use	pca	kmeans	0.4407	0.3606	30
distilbert	tSVD	kmeans	0.4365	0.0226	5
distilbert	tSVD	hdbscan	0.4358	0.0105	5
sentence_transformer	umap	hdbscan	0.4341	0.1608	20
distilbert	umap	kmeans	0.4337	0.0323	7

sentence_transformer	pca	kmeans	0.4325	0.301	5
distilbert	tSVD	hdbscan	0.4286	0.0104	20
distilbert	tSVD	hdbscan	0.4286	0.0104	15
distilbert	tSVD	hdbscan	0.4286	0.0104	10
distilbert	tSVD	hdbscan	0.4286	0.0104	7
distilbert	tSVD	hdbscan	0.427	0.0104	30
use	tSVD	kmeans	0.4269	0.252	5
sentence_transformer	umap	kmeans	0.4258	0.2277	5
distilbert	umap	kmeans	0.4236	0.0348	10
distilbert	tSVD	kmeans	0.4229	0.0351	15
distilbert	pca	kmeans	0.4205	0.0341	30
distilbert	umap	kmeans	0.4187	0.0349	5
sentence_transformer	umap	hdbscan	0.4142	0.1585	10
roberta	pca	kmeans	0.4139	0.0009	20
distilbert	tSVD	kmeans	0.4127	0.034	20
roberta	umap	hdbscan	0.4121	0.0008	30
sentence_transformer	umap	hdbscan	0.4093	0.2569	15
distilbert	pca	kmeans	0.4063	0.0219	5
distilbert	umap	hdbscan	0.4042	0.0188	15
distilbert	tSVD	hdbscan	0.4021	0.0104	7
distilbert	tSVD	hdbscan	0.4018	0.0104	5
distilbert	tSVD	hdbscan	0.4018	0.0104	10
distilbert	tSVD	hdbscan	0.4012	0.0104	20
distilbert	umap	hdbscan	0.4012	0.0251	10
distilbert	tSVD	hdbscan	0.4012	0.0104	15
distilbert	tSVD	kmeans	0.3998	0.0227	5
sentence_transformer	umap	hdbscan	0.3993	0.1514	10
roberta	pca	kmeans	0.3991	0.001	5
roberta	umap	kmeans	0.3946	0.0007	15
distilbert	pca	kmeans	0.3929	0.0359	20
roberta	tSVD	kmeans	0.3912	0.0009	15
distilbert	umap	kmeans	0.3897	0.0319	5
distilbert	umap	hdbscan	0.3884	0.0165	20
roberta	umap	kmeans	0.3869	0.0007	20
distilbert	umap	hdbscan	0.3839	0.0188	30
roberta	tSVD	kmeans	0.3834	0.0009	20
sentence_transformer	umap	hdbscan	0.3834	0.1693	30
distilbert	tSVD	kmeans	0.3829	0.0293	10
distilbert	umap	kmeans	0.3822	0.0351	7
use	umap	kmeans	0.382	0.2935	5
roberta	pca	kmeans	0.3817	0.0011	15
distilbert	tSVD	kmeans	0.3798	0.0275	7
distilbert	umap	hdbscan	0.3791	0.0129	5
distilbert	tSVD	kmeans	0.3781	0.0343	30
roberta	umap	kmeans	0.3771	0.0007	30

roberta	tSVD	kmeans	0.3759	0.001	5
distilbert	umap	hdbscan	0.3737	0.0252	20
distilbert	pca	kmeans	0.372	0.0353	30
roberta	pca	kmeans	0.3687	0.0013	10
distilbert	umap	kmeans	0.3676	0.0336	10
sentence_transformer	umap	hdbscan	0.3651	0.1627	7
distilbert	tSVD	kmeans	0.3629	0.0282	7
distilbert	umap	hdbscan	0.3608	0.0223	7
distilbert	umap	hdbscan	0.3592	0.0186	30
distilbert	umap	kmeans	0.3585	0.034	30
roberta	umap	hdbscan	0.3585	0.0008	20
distilbert	umap	hdbscan	0.3578	0.0185	10
distilbert	umap	hdbscan	0.3576	0.0198	15
distilbert	umap	kmeans	0.3573	0.0342	20
distilbert	pca	kmeans	0.3554	0.0359	15
distilbert	umap	hdbscan	0.3541	0.0188	5
roberta	tSVD	kmeans	0.3533	0.001	7
sentence_transformer	umap	hdbscan	0.3522	0.1718	5
distilbert	pca	kmeans	0.3504	0.0312	10
distilbert	tSVD	hdbscan	0.3454	0.01	30
roberta	tSVD	kmeans	0.341	0.0009	30
roberta	pca	kmeans	0.3407	0.0011	7
roberta	umap	hdbscan	0.3394	0.0007	5
roberta	pca	kmeans	0.3389	0.0009	30
roberta	tSVD	kmeans	0.3337	0.0009	10
roberta	umap	kmeans	0.3336	0.0008	10
roberta	umap	hdbscan	0.3325	0.0008	15
distilbert	pca	hdbscan	0.3309	0.01	20
distilbert	pca	hdbscan	0.3309	0.01	15
distilbert	pca	hdbscan	0.3309	0.01	10
distilbert	pca	hdbscan	0.3305	0.0098	5
distilbert	pca	hdbscan	0.3305	0.0098	30
distilbert	pca	hdbscan	0.3305	0.0098	7
roberta	umap	hdbscan	0.328	0.0007	7
roberta	pca	kmeans	0.3261	0.0009	20
roberta	umap	kmeans	0.3223	0.0008	7
roberta	umap	hdbscan	0.3208	0.0008	10
distilbert	pca	kmeans	0.3194	0.0269	5
distilbert	pca	kmeans	0.3182	0.0271	7
distilbert	pca	kmeans	0.315	0.0353	20
roberta	umap	kmeans	0.3079	0.0009	5
distilbert	pca	hdbscan	0.3053	0.0098	5
distilbert	pca	hdbscan	0.3053	0.0098	30
distilbert	pca	hdbscan	0.3053	0.0098	20
distilbert	pca	hdbscan	0.3053	0.0098	15

distilbert	pca	hdbscan	0.3053	0.0098	10
distilbert	pca	hdbscan	0.3053	0.0098	7
distilbert	umap	hdbscan	0.2995	0.0102	7
distilbert	umap	kmeans	0.2988	0.0343	15
distilbert	tSVD	kmeans	0.2748	0.035	20
use	pca	hdbscan	0	0	5
use	tSVD	hdbscan	0	0	5
sentence_transformer	tSVD	hdbscan	0	0	5
roberta	pca	hdbscan	0	0	5
roberta	tSVD	hdbscan	0	0	5
use	pca	hdbscan	0	0	30
use	tSVD	hdbscan	0	0	30
sentence_transformer	pca	hdbscan	0	0	30
roberta	pca	hdbscan	0	0	30
roberta	tSVD	hdbscan	0	0	30
use	pca	hdbscan	0	0	20
use	tSVD	hdbscan	0	0	20
sentence_transformer	pca	hdbscan	0	0	20
sentence_transformer	tSVD	hdbscan	0	0	20
roberta	pca	hdbscan	0	0	20
roberta	tSVD	hdbscan	0	0	20
use	pca	hdbscan	0	0	15
sentence_transformer	pca	hdbscan	0	0	15
sentence_transformer	tSVD	hdbscan	0	0	15
roberta	pca	hdbscan	0	0	15
roberta	tSVD	hdbscan	0	0	15
use	pca	hdbscan	0	0	10
use	tSVD	hdbscan	0	0	10
sentence_transformer	pca	hdbscan	0	0	10
sentence_transformer	tSVD	hdbscan	0	0	10
use	pca	hdbscan	0	0	5
use	tSVD	hdbscan	0	0	5
roberta	pca	hdbscan	0	0	10
sentence_transformer	pca	hdbscan	0	0	5
sentence_transformer	tSVD	hdbscan	0	0	5
roberta	tSVD	hdbscan	0	0	10
roberta	pca	hdbscan	0	0	5
roberta	tSVD	hdbscan	0	0	5
use	pca	hdbscan	0	0	30
use	tSVD	hdbscan	0	0	30
sentence_transformer	pca	hdbscan	0	0	30
roberta	pca	hdbscan	0	0	30
roberta	tSVD	hdbscan	0	0	30
use	pca	hdbscan	0	0	20
use	tSVD	hdbscan	0	0	20

sentence_transformer	pca	hdbscan	0	0	20
use	pca	hdbscan	0	0	7
sentence_transformer	tSVD	hdbscan	0	0	20
roberta	pca	hdbscan	0	0	20
roberta	tSVD	hdbscan	0	0	20
use	pca	hdbscan	0	0	15
use	tSVD	hdbscan	0	0	15
sentence_transformer	pca	hdbscan	0	0	15
sentence_transformer	pca	hdbscan	0	0	7
sentence_transformer	tSVD	hdbscan	0	0	15
roberta	pca	hdbscan	0	0	15
sentence_transformer	tSVD	hdbscan	0	0	7
roberta	tSVD	hdbscan	0	0	15
use	pca	hdbscan	0	0	10
use	tSVD	hdbscan	0	0	10
sentence_transformer	pca	hdbscan	0	0	10
roberta	pca	hdbscan	0	0	7
sentence_transformer	tSVD	hdbscan	0	0	10
roberta	pca	hdbscan	0	0	10
roberta	tSVD	hdbscan	0	0	7
roberta	tSVD	hdbscan	0	0	10
use	pca	hdbscan	0	0	7
use	tSVD	hdbscan	0	0	7
sentence_transformer	pca	hdbscan	0	0	7
sentence_transformer	tSVD	hdbscan	0	0	7
roberta	pca	hdbscan	0	0	7
roberta	tSVD	hdbscan	0	0	7

5.3 Results from hyper parameters

Optimal combination of model components identified was:

- Embedding model: Sentence transformer
- Dimensionality reduction technique: PCA
- Clustering algorithm: kmeans
- Representation mode: MMR

Based on these, hyper parameter tuning was performed on the following hyper parameters:

- PCA – n_components [3, 5,7, 10,12,15,20]
- Kmean – n_clusters [3, 5,7, 10,12,15,20]
- MMR – kb_diversity [0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45, 0.5]

In the course of this study, a total of 440 iterations were executed. Presented in Table 5.2 are the foremost 20 outcomes derived from the comprehensive hyperparameter optimization process. The evaluation criteria employed for assessing model performance were grounded in the measurement of topic coherence. Upon close examination of the tabulated results in Table 5.2, it is evident that the model exhibiting the highest level of topic coherence attains a score of 0.63. It is worth noting, however, that this optimal model was characterized by a topic count of merely 3, which significantly deviates from the initially designated parameter value of 10. Consequently, for the primary objectives of our research study, which revolve around the identification of thematic elements, a topic count of 3 is deemed inadequate. Subsequently, in pursuit of a more suitable model configuration, a careful consideration of the ensuing results led us to choose the third-ranked outcome. The final set of hyperparameters deemed most appropriate for our research objectives is as follows:

- PCA n_components = 12
- Kmeans k =15
- Diversity = 0.2
- Number of topics = 10

Table 5.2 Top 20 results of hyperparameter tuning

PCA(n)	Kmeans k	diversity	Coherence	Diversity	Number of topics resulted	Number of topics parameter
12	3	0.1	0.6306	0.1681	3	10
15	3	0.1	0.6229	0.1682	3	10
12	15	0.2	0.6086	0.3601	10	10
20	3	0.1	0.6056	0.1677	3	10
20	10	0.1	0.6046	0.3108	10	10
15	10	0.15	0.602	0.3149	10	10
10	15	0.1	0.5969	0.3456	10	10
5	20	0.25	0.5965	0.3486	10	10
12	12	0.2	0.5962	0.3216	10	10
12	10	0.1	0.5914	0.3138	10	10
5	15	0.2	0.5885	0.3302	10	10
20	10	0.2	0.586	0.308	10	10

15	5	0.4	0.5855	0.2427	5	10
7	12	0.2	0.5833	0.3274	10	10
5	15	0.1	0.5828	0.3394	10	10
5	20	0.2	0.5809	0.3598	10	10
10	3	0.15	0.5793	0.1681	3	10
5	20	0.15	0.5787	0.3233	10	10
12	10	0.15	0.5784	0.3096	10	10
15	10	0.1	0.5775	0.3094	10	10

5.4 Final topic model

In accordance with a structured framework, a systematic methodology, and meticulous hyperparameter optimization, the resultant model for the purpose of topic modelling is as follows:

- Embedding Model: Employed herein is the 'Sentence Transformer' architecture denoted as 'all-MiniLM-L6-v2' to encode textual data efficiently.
- Dimensionality Reduction: Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data, retaining only the top 12 principal components for feature representation.
- Clustering Algorithm: The K-Means clustering algorithm is chosen with a configuration of 15 clusters, thereby partitioning the data into distinct topics based on embedded features.
- Topic Representation: A Multi-MMR (Maximal Marginal Relevance) approach is utilized for topic representation, introducing a diversity factor of 0.2 to ensure a varied and informative selection of representative documents within each topic.
- Number of Topics: The final model identifies and delineates a total of 10 distinct topics within the corpus, each characterized by its own unique set of representative documents and thematic content.

Within the ambit of this study's Results chapter, we present a comprehensive overview of key thematic elements derived from our analytical framework. Figure 5.1 offers an exhaustive tabulation, elucidating an array of topics conjoined with their respective top ten pivotal keywords. Fig. 5.2, in congruence with this paradigm, provides a more concise depiction, narrowing its focus to the quintessential top five keywords. In Fig. 5.3, we employ document

embeddings to project these topics onto a two-dimensional space, effectively mapping their interrelationships across various documents. This unique visualization affords the discernment of discernible clusters, underscored by spatial segregation. Drawing insights from the trifecta of Figs. 5.1, 5.2, and 5.3, we proceed to expound upon emergent amalgamations of themes as follows:

- Topic 4, visibly detached within Fig. 5.3, emanates as an emblematic embodiment of the fundamental concepts intrinsic to yoga and the yogi, an inseparable facet of Hindu philosophy and the Bhagavad Gita.

- Positioned at the epicenter of Fig. 5.3, Topic 5 embodies the quintessential teachings of the Bhagavad Gita, emphasizing the primacy of duty and action over the allure of outcomes.

- Topics 2 and 7, clustering harmoniously within Fig. 5.3, encapsulate the discourses and narratives articulated by Lord Krishna, centering around diverse characters in the presence of whom Arjuna grappled with profound existential anguish.

- Topic 3 undertakes the elucidation of the qualities and characteristics germane to the modes of rajas, tamas, and sattva.

- Topic 6 adumbrates the concept of exercising control over sensory perceptions and the objects that stimulate them.

- Topics 1 and 8 converge around the theme of the pursuit of knowledge, resolutely resisting the allure of ignorance.

- Topic 9 confronts the cyclical phenomena of birth, death, and rebirth, while concurrently exalting the concept of an indestructible Brahman.

These synthesized thematic amalgamations, deduced from the orchestrated interplay of Figures 5.1, 5.2, and 5.3, constitute a pivotal facet of our research findings, reinforcing the centrality of the Bhagavad Gita's multifaceted teachings within the purview of our investigation.

Topic	Count	Name	Representation
0	104	0_divine_beings_worship_devotion	[divine, beings, worship, devotion, gods, universe, devotees, celestial, creator, devotee]
1	93	1_intellect_soul_knowledge_ignorance	[intellect, soul, knowledge, ignorance, senses, demoniac, nature, self, persons, possess]
2	92	2_arjun_said_krishna_shree	[arjun, said, krishna, shree, knowledge, sanjay, guṇas, chariot, spoke, soul]
3	90	3_sacrifice_passion_vedas_duties	[sacrifice, passion, vedas, duties, charity, actions, scriptures, qualities, worship, tamas]
4	62	4_yog_yogis_yogi_sanyās	[yog, yogis, yogi, sanyās, devotion, senses, attain, desires, purified, meditation]
5	44	5_duty_actions_renunciation_duties	[duty, actions, renunciation, duties, inaction, surrender, peace, attain, fighting, instruction]
6	40	6_senses_objects_soul_knowledge	[senses, objects, soul, knowledge, nature, embodied, intellect, modes, matter, control]
7	39	7_sons_bheeshma_battle_dhritarasthra	[sons, bheeshma, battle, dhritarasthra, dronacharya, blew, duryodhan, generals, mighty, kingdom]
8	39	8_knowledge_divine_faith_ignorance	[knowledge, divine, faith, ignorance, knower, attain, intellect, darkness, beings, dog]
9	37	9_beings_creation_soul_moon	[beings, creation, soul, moon, sun, nature, eternal, indestructible, realm, brahman]

Figure 5.1 List of topic and top 10 words

Figure 5.2 List of topic and top 5 words

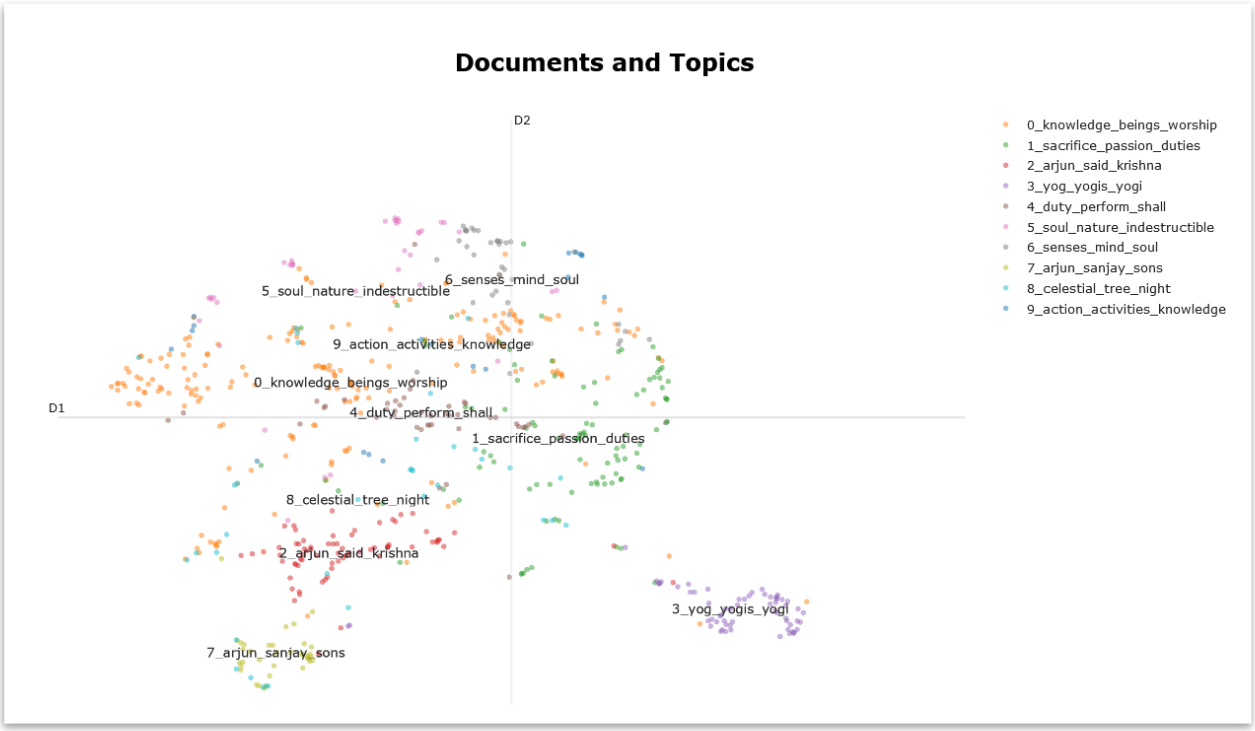

Figure 5.3 Topic clusters over documents

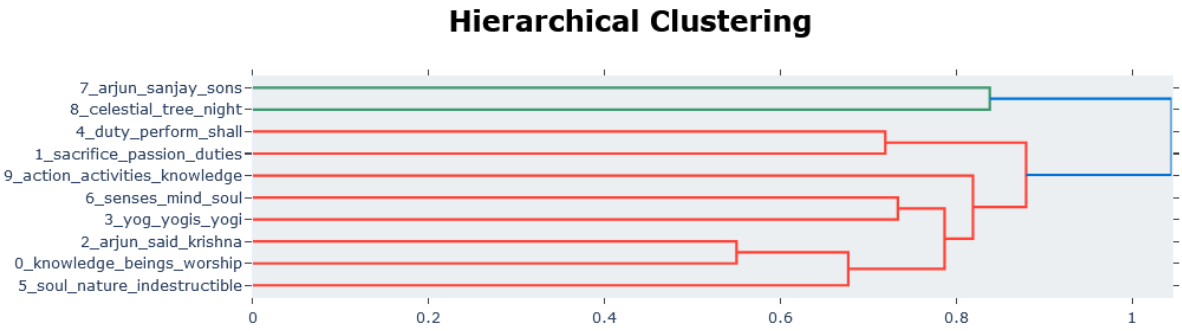

Figure 5.4 Hierarchical clusters of topics

An evaluative dimension in assessing the diversity of topic models is achieved through the utilization of an intertopic distance map, as illustrated in Figure 5.4. This visual representation serves as a valuable tool for discerning the extent of dispersion among the various models under scrutiny. It is noteworthy that the absence of overlapping regions within this map signifies the potential viability of a given topic model for subsequent analytical endeavours. Similar conclusions are offered from the topic similarity matrix shown in Fig 5.6.

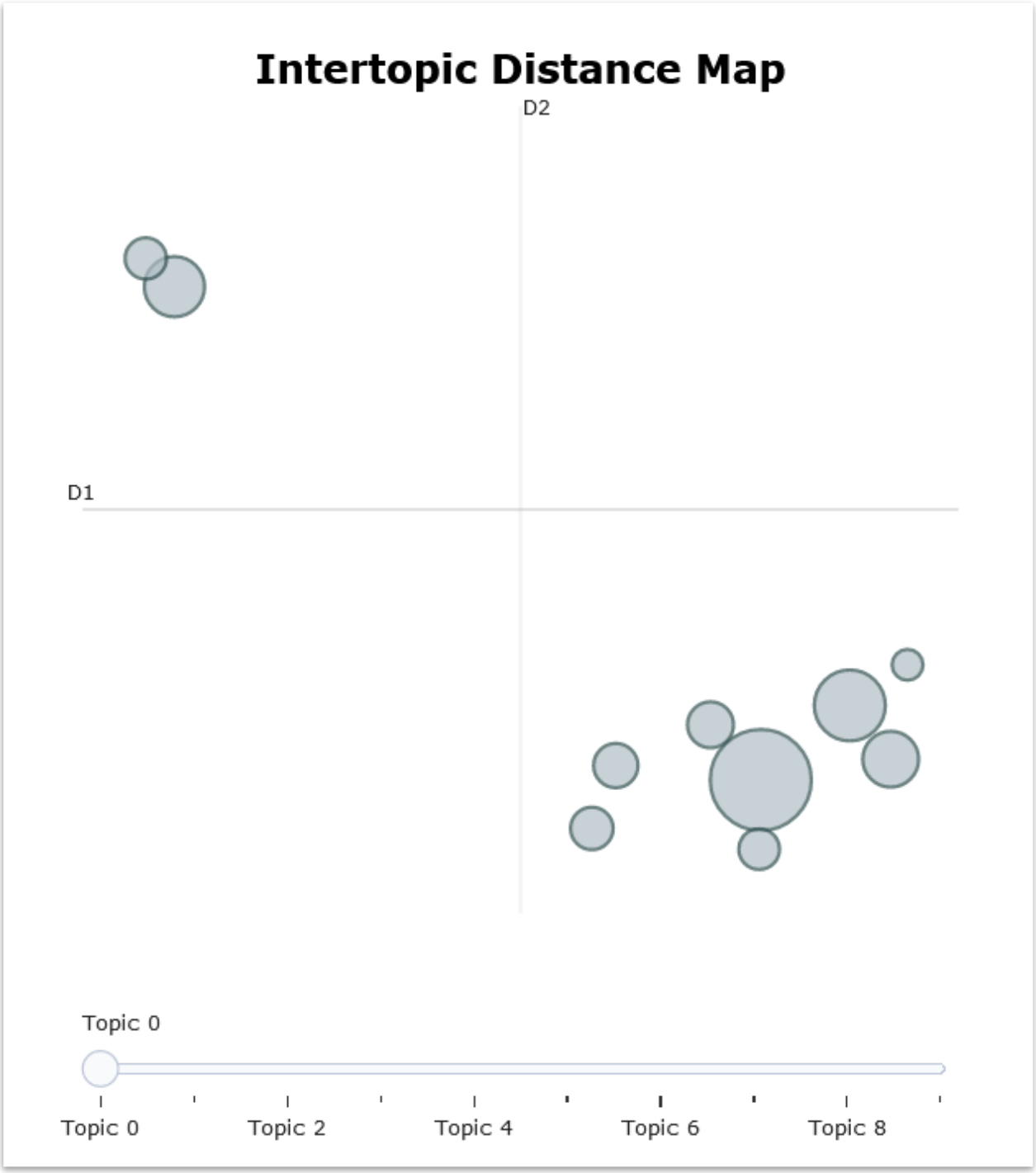

Figure 5.5 Intertopic distance map

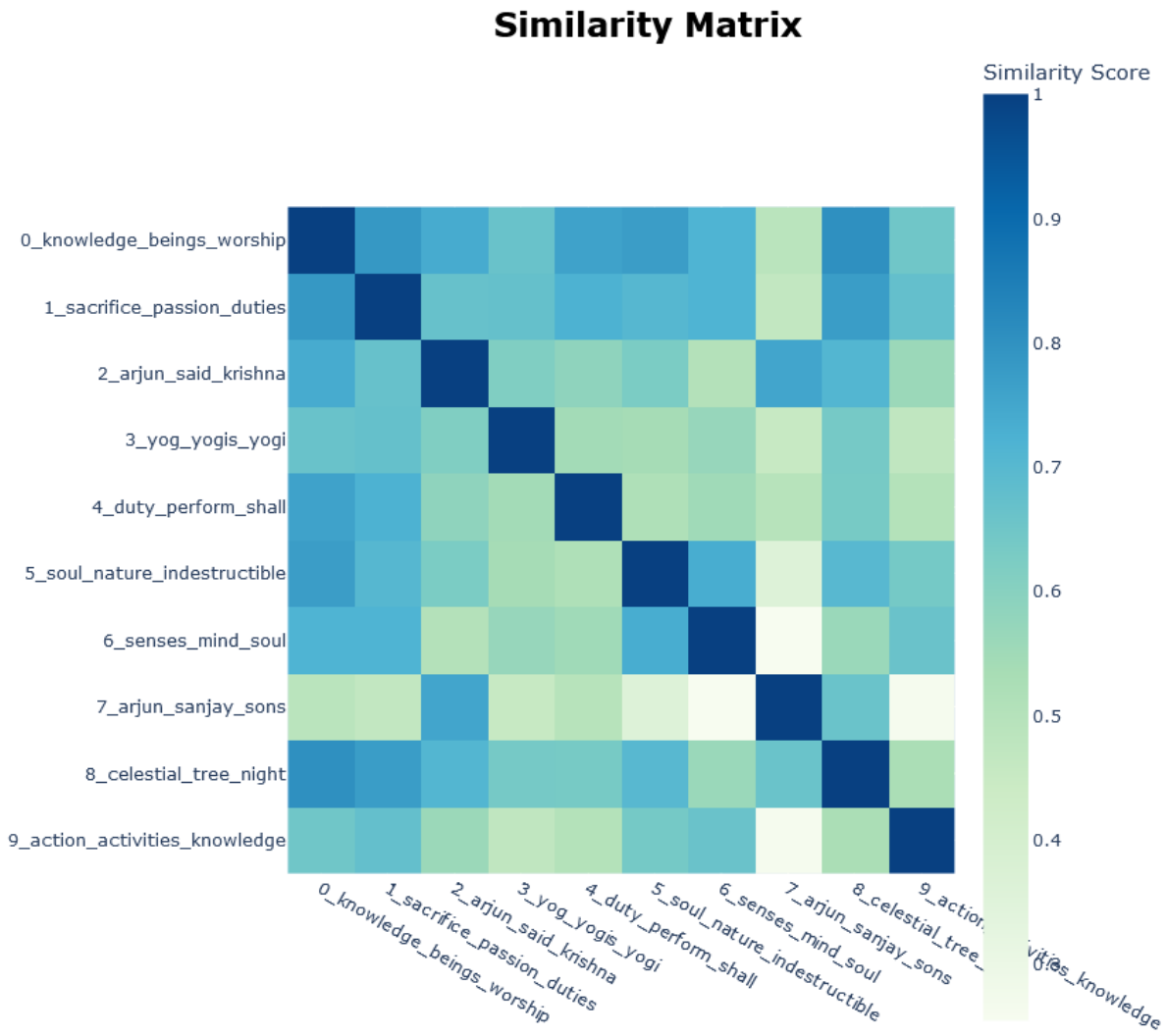

Figure 5.6 Similarity matrix across topics

5.5 Discussion of results

The performance evaluation of the models in terms of both coherence and diversity yielded noteworthy results, aligning with our expectations for a dataset as rich and complex as the Bhagavad Gita. It is imperative to emphasize that certain models exhibited zero results, indicative of their inability to extract coherent topics. This phenomenon was particularly conspicuous in the case of HDBSCAN, which can be attributed to the intricacies of the algorithm and its compatibility with upstream and downstream data processing techniques. For instance, our experiments revealed that the combination of Principal Component Analysis (PCA) and K-means clustering exhibited a commendable synergy, consistent with prior literature. Furthermore, BERTopic analysis suggested that HDBSCAN performed more favourably when paired with Uniform Manifold Approximation and Projection (UMAP).

The anticipation that HDBSCAN would outperform K-means was not substantiated by our empirical findings, as K-means demonstrated superior topic extraction capabilities. This outcome is not entirely surprising, given that HDBSCAN is merely a recommendation, and the choice between these clustering methods hinges on the global and local relationships inherent to the data. In the context of our study, K-means proved to be the more suitable choice. Notably, our observations point to a nuanced scenario where the effectiveness of a specific model is contingent upon the interplay of multiple steps within the framework, rather than a straightforward preference of one model over another. This parallelism can also be drawn in the context of UMAP and PCA, wherein PCA displayed remarkable robustness and resilience. Contrary to initial expectations, the fine-tuning of hyperparameters for PCA yielded competitive results, affirming its efficacy alongside UMAP.

Diversity in topic representation, as discerned through the Maximum Marginal Relevance (MMR) algorithm, emerged as a pivotal consideration for generating a spectrum of diverse topics. The hyperparameter, kb_diversity, proved to be instrumental in optimizing this aspect, particularly when working with datasets characterized by a wide variety of contextual nuances. This nuanced hyperparameter adjustment becomes indispensable when dealing with data that exhibits a plethora of diverse contexts, as it facilitates the extraction of topics that are both coherent and diverse. Overall, the strength of the results lies in its state-of-the-art performing topic modelling approach.

5.6 Limitations

In the context of our research, it is imperative to acknowledge and delineate the inherent limitations of our proposed methodology. Primarily, our approach, which is based on BERTopic, operates under the assumption that each document can be unequivocally assigned to a singular topic. However, this assumption may inadequately capture the nuanced complexities present in textual content, especially in instances of intricate messaging, such as religious scriptures. It should be noted that topic modelling techniques, like BERTopic, may exhibit inherent instability. Variability arises from the stochastic nature of topic modelling algorithms, where different initializations (e.g., random seeds) can yield disparate sets of topic clusters, thereby impeding the interpretability and consistency of results. Moreover, our approach encounters limitations regarding the exhaustiveness of topic representation within a text. The introduction of a higher number of topics may introduce excessive noise and undermine the interpretability of the results. Furthermore, it is worth noting that the embedding

models predominantly employed in our research are primarily based on BERT and similar models. These embedding models, while proficient in handling general tasks, may fall short when confronted with the intricate nuances present in religious scripture. Consequently, the generalizability of our approach may be restricted in such contexts. Additionally, our methodology retains a reliance on TF-IDF-based representations for topic modelling, rather than adopting more contemporary embedding-based techniques. This choice may introduce limitations in capturing the semantic relationships and contextual nuances present in the text.

The dataset employed in our study inherently bears biases introduced during the translation process, stemming from the stylistic choices, personal views, and vocabulary of the translators. It is important to acknowledge that our proof of concept is based on a single example translation and may not comprehensively reflect the insights attainable through topic modelling when applied to a complex text like the Bhagavad Gita, a revered scripture. Furthermore, the translation aspect introduces a temporal dimension to our limitations. Over time, interpretations of the Bhagavad Gita have evolved, and different translations have emerged. Therefore, conducting a similar analysis on archaic interpretations may yield distinct results compared to more recent translations, underscoring the dynamic nature of the text. Lastly, it is essential to acknowledge that the Bhagavad Gita, originally composed in Sanskrit as a poetic work, inherently possesses layers of meaning that may be lost in translation into English or other languages. The richness and subtlety of its poetic form may not be fully captured in our analytical approach, marking a notable limitation.

These limitations, collectively, warrant a cautious interpretation of our findings and suggest avenues for future research to refine and expand upon the methodology for a more comprehensive understanding of complex textual content, particularly in the realm of religious scripture analysis.

5.7 Summary

Within the realm of Natural Language Processing (NLP) applied to the Bhagavad Gita, there exists a significant gap in advanced research efforts. To address this lacuna and contribute to the field, this thesis represents pioneering work, marking the inception of advanced NLP investigations into the Bhagavad Gita. Prior research endeavours have predominantly adhered to rudimentary and linguistically oriented approaches, thus warranting a novel perspective. The present study serves as a trailblazing demonstration, affirming the feasibility and applicability

of NLP techniques to the Bhagavad Gita, a feat hitherto unexplored in the academic discourse. In doing so, this research not only signifies a pioneering achievement but also establishes a crucial foundational framework for future investigations. A pivotal contribution of this research lies in the identification and validation of the optimal framework for NLP analysis in this context, which is based on the innovative BERTopic methodology. Additionally, this study has meticulously ascertained the combination of methods that holds substantial promise for adoption by the broader research community, thereby setting a precedent for future inquiries into this domain.

CHAPTER 6

CONCLUSIONS & RECOMMENDATIONS

6.1 Introduction

Chapter 6 represents the culmination of this thesis, drawing to a close with a detailed exposition of the conclusions drawn from the research findings. Furthermore, it provides a set of prospective recommendations for future research directions while elucidating the distinct contributions made by this study to the existing body of knowledge.

6.2 Discussion & Conclusion

This study has successfully demonstrated the implementation of a robust topic modelling framework applied to the Bhagavad Gita text within the context of the current research. The resultant topics have been meticulously interpreted and substantiated by drawing upon the insights gleaned from the comprehensive literature review conducted in this study. Moreover, this research has undertaken an exhaustive approach in formulating the sub-components of a state-of-the-art topic modelling methodology, specifically integrating BERTtopic. This comprehensive approach has yielded not only the optimal combination of methodologies but also the judicious selection of hyperparameters.

In summary, a total of ten distinct topics have been identified through the application of this framework. The evaluation of the model's performance, from both a coherence perspective (achieving a score of 0.61) and diversity (scoring 0.30), underscores its effectiveness. The most suitable document representation for this dataset has been determined to be based on Sentence Transformer embeddings in conjunction with Principal Component Analysis (PCA) featuring twelve components. The clustering algorithm employed in this context is K-Means. Furthermore, the incorporation of Maximal Marginal Relevance (MMR) has been instrumental in ensuring the requisite diversity in the representation of topics. It is noteworthy that the themes that have emerged from the final set of topics exhibit a remarkable alignment with the teachings and insights discerned from the Bhagavad Gita, thus reinforcing the efficacy of this topic modelling approach within the chosen domain.

6.3 Contribution to knowledge

Within the realm of Natural Language Processing (NLP) applied to the Bhagavad Gita, there exists a significant gap in advanced research efforts. To address this lacuna and contribute to the field, this thesis represents pioneering work, marking the inception of advanced NLP investigations into the Bhagavad Gita. Prior research endeavours have predominantly adhered to rudimentary and linguistically oriented approaches, thus warranting a novel perspective. The present study serves as a trailblazing demonstration, affirming the feasibility and applicability of NLP techniques to the Bhagavad Gita, a feat hitherto unexplored in the academic discourse. In doing so, this research not only signifies a pioneering achievement but also establishes a crucial foundational framework for future investigations. A pivotal contribution of this research lies in the identification and validation of the optimal framework for NLP analysis in this context, which is based on the innovative BERTopic methodology. Additionally, this study has meticulously ascertained the combination of methods that holds substantial promise for adoption by the broader research community, thereby setting a precedent for future inquiries into this domain.

6.4 Future Recommendations

In the context of future recommendations for research, several avenues present themselves for further exploration and enhancement of the study. These recommendations are imperative for advancing our understanding of the Bhagavad Gita and its interpretation through the lens of Natural Language Processing (NLP). The following suggestions are made to guide future investigations in this domain:

- Exploration of Verses: Delve deeper into the verses of the Bhagavad Gita, focusing on the interactions between Krishna, Arjuna, and Sanjaya. This analysis can provide valuable insights into the dynamics of communication and discourse within the text.
- Translation Variations: Consider incorporating a wider array of translations, including those obtained through Optical Character Recognition (OCR). Comparative analysis of translations can unveil nuances in interpretation and reveal potential biases introduced by different authors.
- Chapter-wise Analysis: Conduct a systematic chapter-wise analysis of the Bhagavad Gita, examining linguistic patterns, thematic evolution, and variations in sentiment. This approach can yield a comprehensive understanding of the text's structure and narrative progression.

- Hindi Translation and Meaning: Utilize Hindi translations and meanings to enrich the analysis, ensuring a comprehensive exploration of linguistic and cultural dimensions.
- Sentiment Analysis: Introduce classification tags such as sentiment analysis to augment the interpretive framework, enabling a more nuanced understanding of emotional and contextual nuances within the text.
- Translator Bias and Trends: Investigate translator biases and trends over time to elucidate evolving perspectives and interpretations of the Bhagavad Gita, shedding light on the text's evolving comprehension.
- Summarization of Religious Texts: Extend the analysis to include the summarization of other religious texts, fostering comparative insights and coherence assessment across different religious scriptures.
- Comparative Analysis: Continuously compare the findings of this study with other research results, both within the context of the Bhagavad Gita and across various NLP techniques and linguistic aspects.
- Exploration of Other Techniques: Consider the integration of other advanced techniques and preprocessing methodologies, such as GPT-based models and chatbots, to enhance the depth of analysis.
- Cross-Scriptural Analysis: Extend the research to encompass analyses of other scriptures from different religious traditions, facilitating cross-cultural comparisons and enriching the scholarly discourse.
- Sanskrit Models: Investigate the feasibility of employing advanced Sanskrit NLP models to improve the accuracy and depth of analysis when dealing with the original Sanskrit text.
- Various NLP Formulations: Explore diverse NLP formulations and methodologies to refine and expand the scope of textual analysis.
- Linguistic Aspects: Delve into various linguistic aspects of the Bhagavad Gita, such as syntax, semantics, and discourse structure, to extract deeper insights into the text's linguistic characteristics.

By pursuing these recommendations, future research endeavours can build upon the foundation laid by this thesis, contributing significantly to the understanding of the Bhagavad Gita and advancing the field of NLP applied to religious and philosophical texts.

REFERENCES

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W. and Hassan, A., (2023) *Topic modelling algorithms and applications: A survey*. *Information Systems*,

Ambrosino, A., Cedrini, M., Davis, J.B., Fiori, S., Guerzoni, M. and Nuccio, M., (2018) What topic modelling could reveal about the evolution of economics*. *Journal of Economic Methodology*, 254, pp.329–348.

Anon (2011) *2011 26th IEEE/ACM International Conference on Automated Software Engineering*. IEEE.

Bahdanau, D., Cho, K. and Bengio, Y., (2014) Neural Machine Translation by Jointly Learning to Align and Translate. [online] Available at: http://arxiv.org/abs/1409.0473.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., Ca, J.U., Kandola, J., Hofmann, T., Poggio, T. and Shawe-Taylor, J., (2003) *A Neural Probabilistic Language Model*. *Journal of Machine Learning Research*, .

Bhatia, S.C., Madabushi, J., Kolli, V., Bhatia, S.K. and Madaan, V., (2013) *The Bhagavad Gita and contemporary psychotherapies*. *Indian Journal of Psychiatry*, .

Blei, D.M. and Lafferty, J.D., (n.d.) *Correlated Topic Models*. [online] Available at: www.jstor.org.

Blei, D.M., Ng, A.Y. and Edu, J.B., (2003a) *Latent Dirichlet Allocation Michael I. Jordan*. *Journal of Machine Learning Research*, .

Blei, D.M., Ng, A.Y. and Edu, J.B., (2003b) *Latent Dirichlet Allocation Michael I. Jordan*. *Journal of Machine Learning Research*, .

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B. and Kurzweil, R., (2018) Universal Sentence Encoder. [online] Available at: http://arxiv.org/abs/1803.11175.

Chandra, R. and Kulkarni, V., (2022) Semantic and Sentiment Analysis of Selected Bhagavad Gita Translations Using BERT-Based Language Framework. *IEEE Access*, 10, pp.21291–21315.

Chandra, R. and Ranjan, M., (2022) Artificial intelligence for topic modelling in Hindu philosophy: Mapping themes between the Upanishads and the Bhagavad Gita. *PLoS ONE*, 179 September.

Chowdhury, G.G., (2003) Natural language processing. *Annual Review of Information Science and Technology*, 37, pp.51–89.

Collobert, R. and Weston, J., (2008) *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*. [online] Available at: http://wordnet.princeton.edu.

Das, N. and Behura, A.K., (2021) *Relevance of Bhagavad Gita for healthcare workers amidst the COVID -19 crisis*. *Asian Journal of Psychiatry*, .

Das, R., Zaheer, M. and Dyer, C., (n.d.) *Gaussian LDA for Topic Models with Word Embeddings*. [online] Available at: https://code.google.com/p/word2vec/.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 416, pp.391–407.

Devlin, J., Chang, M.-W., Lee, K., Google, K.T. and Language, A.I., (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] Available at: https://github.com/tensorflow/tensor2tensor.

Dhillon, M., (2023) Weaving Together the Ancient and the Contemporary: Intersections of the Bhagavad Gita with Modern Psychology. *Pastoral Psychology*.

Dieng, A.B., Ruiz, F.J.R. and Blei, D.M., (n.d.) Topic Modeling in Embedding Spaces. [online] Available at: https://doi.org/10.1162/tacl.

Dit, B., Revelle, M., Gethers, M. and Poshyvanyk, D., (2013) Feature location in source code: A taxonomy and survey. *Journal of software: Evolution and Process*, 251, pp.53–95.

Easwaran, E., (1985) *Bhagavad Gita Translated for Modern Readers*.

Gandhi, M., (2010) *The bhagavad gita according to Gandhi*. North Atlantic Books.

Grootendorst, M., (2022) BERTopic: Neural topic modelling with a class-based TF-IDF procedure. [online] Available at: http://arxiv.org/abs/2203.05794.

Haghighi, A. and Vanderwende, L., (n.d.) *Exploring Content Models for Multi-Document Summarization*. [online] Available at: http://www-nlpir.nist.gov/projects/duc/data.html.

Hemmati, H., Fang, Z., Mäntylä, M. V. and Adams, B., (2017) Prioritizing manual test cases in rapid release environments. In: *Software Testing Verification and Reliability*. John Wiley and Sons Ltd.

Hochreiter, S. and Urgen Schmidhuber, J.¨, (1997) *Long Short-Term Memory*. [online] Available at: http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

Hofmann, T., (1999) *Probabilistic Latent Semantic Indexing*.

Institute of Electrical and Electronics Engineers, (2012) *ICSM 2012 : Proceedings of the 28th IEEE International Conference on Software Maintenance : Riva Del Garda, Trento, Italy : [23-28 Sept. 2012]*. IEEE.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L., (2019) Latent Dirichlet allocation (LDA) and topic modelling: models, applications, a survey. *Multimedia Tools and Applications*, 7811, pp.15169–15211.

Jeong, B., Yoon, J. and Lee, J.M., (2019) Social media mining for product planning: A product opportunity mining approach based on topic modelling and sentiment analysis. *International Journal of Information Management*, 48, pp.280–290.

Jeste, D. V. and Vahia, I. V., (2008) *Comparison of the conceptualization of wisdom in ancient Indian literature with modern views: Focus on the Bhagavad Gita*. *Psychiatry*, .

Joachims, T., (1996) *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*.

Kalra, B., Joshi, A., Kalra, S., Shanbhag, V.G., Kunwar, J., Balhara, Y.P.S., Chaudhary, S., Khandelwal, D., Aggarwal, S., Priya, G., Verma, K., Baruah, M.P., Sahay, R., Bajaj, S., Agrawal, N., Pathmanathan, S., Prasad, I., Chakraborty, A. and Ram, N., (2018) *Coping with illness: Insight from the Bhagavad Gita*. *Indian Journal of Endocrinology and Metabolism*, .

Kalra, S., Joshi, A., Kalra, B., Shanbhag, V.G., Bhattacharya, R., Verma, K., Baruah, M.P., Sahay, R., Bajaj, S., Agrawal, N., Chakraborty, A., Balhara, Y.P.S., Chaudhary, S., Khandelwal, D., Aggarwal, S., Ram, N., Jacob, J., Julka, S., Priya, G., Bhattacharya, S. and Dalal, K., (2017) *Bhagavad gita for the physician*. *Indian Journal of Endocrinology and Metabolism*, .

Karekar, A., Limaye, S., Nara, A. and Panchal, S., (2023) Bhagavad Geeta Based ChatBot. In: *2023 3rd International Conference on Intelligent Technologies (CONIT)*. [online] IEEE, pp.1–6. Available at: https://ieeexplore.ieee.org/document/10205716/.

Keshavan, M.S., (2020) *Building resilience in the COVID-19 era: Three paths in the bhagavad gita*. *Indian Journal of Psychiatry*, .

Khurana, D., Koli, A., Khatter, K. and Singh, S., (2023) Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 823, pp.3713–3744.

Lee D.D and Seung H.S, (1999) 44565. *Nature*.

Lee, D.D. and Seung, H.S., (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 4016755, pp.788–791.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W., (2016) *An overview of topic modelling and its current applications in bioinformatics*. *SpringerPlus*, .

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. [online] Available at: http://arxiv.org/abs/1907.11692.

Lloyd, S.P., (1982) *Least Squares Quantization in PCM*. *IEEE TRANSACTIONS ON INFORMATION THEORY*, .

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O. and Zaremba, W., (2014) Addressing the Rare Word Problem in Neural Machine Translation. [online] Available at: http://arxiv.org/abs/1410.8206.

McInnes, L. and Healy, J., (2017) Accelerated Hierarchical Density Based Clustering. In: *IEEE International Conference on Data Mining Workshops, ICDMW*. IEEE Computer Society, pp.33–42.

McInnes, L., Healy, J. and Melville, J., (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. [online] Available at: http://arxiv.org/abs/1802.03426.

Meister, C., (2009) *Introducing philosophy of religion*. Routledge.

Menon, B., Narayan, S.K. and Bhade, S., (2021) COVID-19, Moral Injury and the Bhagvad Gita. *Journal of Religion and Health*, 602, pp.654–662.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., (2013a) *Distributed Representations of Words and Phrases and their Compositionality*.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., (2013b) Efficient Estimation of Word Representations in Vector Space. [online] Available at: http://arxiv.org/abs/1301.3781.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and Mccallum, A., (2011) *Optimizing Semantic Coherence in Topic Models*. Association for Computational Linguistics.

Moody, C.E., (2016) Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. [online] Available at: http://arxiv.org/abs/1605.02019.

Mukherjee, S., (2017) Bhagavad Gita: The key source of modern management. *Asian Journal of Management*, 81, p.68.

Muniapan, B. and Satpathy, B., (2013) The ‘dharma’ and ‘karma’ of CSR from the Bhagavad-Gita. *Journal of Human Values*, 192, pp.173–187.

Murray, M.J. and Rea, M.C., (2008) *An introduction to the philosophy of religion*. Cambridge University Press.

Nagappa, S.P., (2023) *LIFE MANAGEMENT PRINCIPLES FROM THE BHAGAVAD GITA*. [online] Available at: www.ijcrt.org.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T.E., Sastry, G., Krueger, G., Schnurr, D., Such, F.P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P. and Weng, L., (2022) Text and Code Embeddings by Contrastive Pre-Training. [online] Available at: http://arxiv.org/abs/2201.10005.

Newman, D., Jey, ♠ ♣, Lau, H., Grieser, K. and Baldwin, T., (n.d.) *Automatic Evaluation of Topic Coherence*.

Newman, D., Jey, Lau, H., Grieser, K. and Baldwin, T., (2010) *Automatic Evaluation of Topic Coherence*.

Nguyen, D.Q., Billingsley, R., Du, L. and Johnson, M., (n.d.) Improving Topic Models with Latent Feature Word Representations. [online] Available at: http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00140/1566790/tacl_a_00140.pdf.

Osuntoki, S., Odumuyiwa, V. and Sennaike, O., (2022) Understanding Document Thematic Structure: A Systematic Review of Topic Modeling Algorithms. *Journal of Information and Organizational Sciences*, 462, pp.305–322.

Pandey, Y.R., (2017) *Economic Interpretation of Philosophy of Bhagavad Gita: A Descriptive Analysis*. *Economic Journal of Development Issues*, .

Pandurangi, A.K., Shenoy, S. and Keshavan, M.S., (2014) *Psychotherapy in the Bhagavad Gita, the Hindu Scriptural Text*. *American Journal of Psychiatry*, .

Pennington, J., Socher, R. and Manning, C.D., (2014) *GloVe: Global Vectors for Word Representation*. [online] Available at: http://nlp.

Prabhupada, A.C.B.S. and Swami, B., (1972) *Bhagavad-Gita as it is*. Bhaktivedanta Book Trust Los Angeles.

Purna Sudhakar, G., (2014) *Project Management Insights from Bhagavad Gita*. [online] *PM World Journal Project Management Insights from Bhagavad Gita*, Available at: http://ssrn.com/abstract=2475474www.pmworldjournal.net.

Rajagopalachari, C., (1970) *Mahabharata (Vol. 1)*.

Rajandran, K., (2017) From matter to spirit: Metaphors of enlightenment in Bhagavad-gītā. *GEMA Online Journal of Language Studies*, 172, pp.163–176.

Rajput, N.K., Ahuja, B. and Riyal, M.K., (2019) A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita. *Pramana - Journal of Physics*, 924.

Reimers, N. and Gurevych, I., (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [online] Available at: http://arxiv.org/abs/1908.10084.

Reimers, N. and Gurevych, I., (2020) *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. [online] Available at: https://github.com/facebookresearch/.

Renard, P., (1995) Historical bibliography of upani ads in translation. *Journal of Indian philosophy*, 232, pp.223–246.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T., (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. [online] Available at: http://arxiv.org/abs/1910.01108.

Santoro, A., Faulkner, R., Raposo, D., Rae αβ, J., Chrzanowski α, M., Weber α, T., Wierstra α, D., Vinyals α, O., Pascanu α, R. and Lillicrap αβ, T., (n.d.) *Relational recurrent neural networks*.

Satpathy, B., Muniapan, B. and Dass, M., (2013) *UNESCAP's characteristics of good governance from the philosophy of Bhagavad-Gita and its contemporary relevance in the Indian context*. [online] *Int. J. Indian Culture and Business Management*, Available at: http://www.gdrc.org/u-gov/.

Sia, S., Dalmia, A. and Mielke, S.J., (2020) Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! [online] Available at: http://arxiv.org/abs/2004.14914.

Simpson, A. V. and Pina e Cunha, M., (2021) A Bhagavad Gita-inspired Linked Leadership Model. *Journal of Leadership Studies*, 153, pp.43–48.

Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C., (n.d.) *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. [online] Association for Computational Linguistics. Available at: http://nlp.stanford.edu/.

Srinivas, S. and Ramachandiran, S., (2020) Discovering Airline-Specific Business Intelligence from Online Passenger Reviews: An Unsupervised Text Analytics Approach. [online] Available at: http://arxiv.org/abs/2012.08000.

Stein, D., (2012) *Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes*. [online] Available at: https://hal.science/hal-02879329.

Sun, L. and Yin, Y., (2017) Discovering themes and trends in transportation research using topic modelling. *Transportation Research Part C: Emerging Technologies*, 77, pp.49–66.

Sun, Q., Li, R. and Wu, X., (n.d.) *Text Segmentation with LDA-Based Fisher Kernel*.

Swami Mukundananda, (n.d.) *Holy Bhagavad Gita*. https://www.holy-bhagavad-gita.org/.

Swami, S.P., (2003) *Bhagavad Gita*. Jaico Publishing House.

Tan, K.L., Lee, C.P., Anbananthen, K.S.M. and Lim, K.M., (2022) RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access*, 10, pp.21517–21525.

Thakur, N., Reimers, N., Daxenberger, J. and Gurevych, I., (2020) Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. [online] Available at: http://arxiv.org/abs/2010.08240.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., (2017) *Attention Is All You Need*.

Vayansky, I. and Kumar, S.A.P., (2020) A review of topic modelling methods. *Information Systems*, 94.

Verma, M., (2017) Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools. *International Journal of Computer Applications*, 1688, pp.39–45.

Wei Wang and Jianxun Gang, (2018) *Proceedings of 2018 International Conference on Information Systems and Computer Aided Education : ICISCAE 2018 : July 6-8, 2018, Changchun, China*.

Wiese, G., Weissenborn, D. and Neves, M., (2017) Neural Domain Adaptation for Biomedical Question Answering. [online] Available at: http://arxiv.org/abs/1706.03610.

Xu, D. and Tian, Y., (2015) A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 22, pp.165–193.

Yamunathangam, D., Priya, C.B., Shobana, G. and Latha, L., (2021) An Overview of Topic Representation and Topic Modelling Methods for Short Texts and Long Corpus. In: *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2021*. Institute of Electrical and Electronics Engineers Inc.

Yang, M.-S., (1993) *A Survey of Fuzzy Clustering. Mathl. Comput. Modelling,* .

Yash Narnaware, (2023) *Bhagavad Gita verse-wise (English, Hindi, Sanskrit)*. https://www.kaggle.com/datasets/yashnarnaware/bhagavad-gita-versewise.

Yu, S., Indurthi, S., Back, S. and Lee, H., (2018) *A Multi-Stage Memory Augmented Neural Network for Machine Reading Comprehension*.