Dissertation/Report submitted to the

UNIVERSITY OF MOHAMED BOUDIAF - M'SILA, ALGERIA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

in partial fulfillment of the requirements for the degree of

Master's in Computer Science

Specialization: Information System and Software Engineering

Presented by

HADJER MEGHICHE
KHADIDJA NOUR ELHOUDA GANA

Entitled

Document Clustering and Visualization for Large Text Collections

Under the supervision of

DR.BILAL LOUNNAS

Jury Members

Dr. Marouane Kihal	University of M'sila	President
Dr. BILAL LOUNNAS	University of M'sila	Reporter
Dr. Nour Elhouda Chalabi	University of M'sila	Examiner

June, 2026

الملخص

أفرز النمو الهائل للبيانات النصية الرقمية في المجالات العلمية والصحفية ووسائل التواصل الاجتماعي حاجةً ملحّة إلى أدوات آلية لتنظيم المستندات واستكشافها. تقدّم هذه الدراسة منظومة متكاملة من البداية إلى النهاية مخصصة لتجميع المستندات وتمثيلها المرئي على مجموعات نصية واسعة النطاق. يتم تقييم ثلاثة مقاربات تكاملية بصورة منهجية: خوارزمية K-Means مقرونةً بتمثيلات TF-IDF، وخوارزمية التخصيص الديريكليتي الكامن (LDA)، وإطار BERTopic العصبي المتقدم الذي يجمع بين التضمينات السياقية المستندة إلى المحوّلات وتقليل الأبعاد عبر UMAP والتجميع الكثافي HDBSCAN. أُجريت التجارب على ثلاثة مجموعات مرجعية 20 Newsgroups وReuters-21578 ومجموعة arXiv منتقاة وتمثّل مستويات متنوعة من التعقيد الدلالي وتوزيع الفئات. قُيّم الأداء بستة مقاييس تكاملية تشمل الجودة الداخلية للتجمعات والاتساق مع التسميات المرجعية والتماسك الموضوعي والتنوع والكفاءة الحسابية. تُبيّن النتائج أن BERTopic يحقق تماسكاً دلالياً وتنوعاً موضوعياً متميزاً، في حين يوفر K-Means أعلى دقة هيكلية وكفاءة حسابية عند إقرانه بتضمينات الجمل الكثيفة، بينما يُسجّل LDA أداءً أدنى في استيعاب البنية الدلالية العميقة رغم قابليته للتفسير. تتيح هذه النتائج إرشادات عملية لاختيار أساليب تجميع المستندات في تطبيقات استرجاع المعلومات الفعلية.

الكلمات المفتاحية-أمثلة: تجميع المستندات ، BERTopic، K-Means، LDA، TF-IDF، Sentence-BERT، UMAP، تقليل الأبعاد، النمذجة الموضوعية، تنقيب النصوص.

Abstract

The exponential growth of digital textual data across scientific, journalistic, and social media domains has created a critical need for automated document organisation and exploration tools. This study presents a comprehensive end-to-end pipeline for document clustering and visualisation applied to large text collections. Three complementary approaches are systematically evaluated: K-Means clustering with TF-IDF representations, Latent Dirichlet Allocation (LDA), and BERTopic, a state-of-the-art neural framework combining transformer-based embeddings with UMAP dimensionality reduction and HDBSCAN density-based clustering. Experiments are conducted on three benchmark corpora 20 Newsgroups, Reuters-21578, and a curated arXiv dataset representing diverse levels of semantic complexity and class distribution. Performance is assessed using six complementary metrics covering internal cluster quality, external label agreement, topic coherence and diversity, and computational efficiency. Results demonstrate that BERTopic achieves superior semantic coherence and topic diversity, while K-Means delivers the highest structural accuracy and computational efficiency when paired with dense sentence embeddings. LDA, although interpretable, consistently underperforms in capturing deep semantic structure. These findings provide practical guidelines

for selecting document clustering methods in real-world information retrieval applications.

Keywords: Document Clustering, BERTopic, K-Means, LDA, TF-IDF, Sentence-BERT, UMAP, Dimensionality Reduction, Topic Modelling, Text Mining.

Dedication

This work is respectfully dedicated to all individuals who have supported and encouraged the completion of this research. Their guidance, assistance, and continuous motivation have been invaluable throughout this academic journey.

Hadjer Meghiche

"*To my beloved parents,*

whose unconditional love, endless sacrifices, and unwavering support
have been our greatest strength throughout this journey.

To my family,

for their patience, encouragement, and constant presence
in every moment of difficulty and joy.

And to everyone who believed me
and supported me along the way,

this work is dedicated to you with all our gratitude and love.

Khadidja Nour Elhouda Gana

Acknowledgements

We begin by thanking ALLAH for granting us the strength and patience to complete this journey.

We would like to express our profound gratitude to our supervisor, Dr. Bilal Lounnas, for his continuous guidance, support, and encouragement. His expertise and dedication have been invaluable to us throughout this project.

Our deepest and most heartfelt thanks go to our parents, for their unconditional love, their endless sacrifices, and their prayers that have never ceased. We dedicate this achievement to them above all.

We are also grateful to our families and friendsfor standing by us, for their patience and support, and for making this journey easier and more meaningful.

Finally, we thank all our professors and everyone at the University of Mohamed Boudiaf,M'sila who contributed to our academic growth.

To all of you,thank you from the bottom of our hearts.

Contents

List of Tables

List of Figures

Introduction

The exponential proliferation of digital textual data across diverse domains has fundamentally transformed how information is generated, processed, and consumed in the modern era. From scientific repositories and financial news archives to social media platforms and enterprise document management systems, the sheer volume of unstructured text produced every day has far outpaced the capacity of human analysts to organise, interpret, and extract value from it. Traditional document management approaches, which relied on manually curated taxonomies, keyword-based indexing, and domain-expert classification, are no longer scalable or economically viable in this data-rich environment. As a consequence, automated computational techniques capable of organising, exploring, and discovering knowledge from large text collections have become indispensable components of modern information systems, driving significant research investment across the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML).

Among the most fundamental of these techniques, document clustering the task of automatically partitioning a corpus of texts into coherent, semantically meaningful groups without relying on predefined labels or human supervision has attracted sustained research attention for several decades. By uncovering latent thematic structures embedded within document collections, clustering enables a broad spectrum of high-value downstream applications, including search engine result diversification, automated topic tracking in real-time news streams, personalised recommendation and content filtering systems, scientific literature mapping and citation analysis, and large-scale enterprise knowledge management. Despite this progress, the field continues to face substantive and largely unresolved challenges rooted in the inherent complexity of natural language:extreme feature-space dimensionality, pervasive semantic ambiguity, distributional imbalance across latent topics, vocabulary mismatch between semantically related documents, and the fundamental inability of classical statistical models to capture contextual word meaning beyond surface-level lexical overlap.

- **Key Research Aspects:**
  - Current state of document clustering methodologies in NLP and text mining, spanning classical partition-based, probabilistic, and modern neural approaches.

  - Critical challenges in high-dimensional semantic text representation, including the curse of dimensionality, sparsity of lexical feature spaces, and the semantic gap between syntactic form and conceptual meaning.
  - Emerging needs for scalable, interpretable, and domain-agnostic clustering systems capable of generalising across heterogeneous corpora with varying structural and distributional properties.
  - The growing role of transformer-based language models in enabling context-aware document representations that substantially close the semantic gap left unaddressed by classical approaches.

- **Core Problems Identified:**
  - *Semantic ambiguity across heterogeneous document collections:* documents covering the same underlying topic may employ entirely different vocabulary, making purely lexical similarity measures unreliable and leading to fragmented or incoherent cluster formations.
  - *Scalability limitations of classical clustering algorithms:* methods such as hierarchical clustering and density-based approaches exhibit quadratic or super-linear time complexities that render them impractical for large-scale corpora containing tens of thousands of documents or more.
  - *Absence of unified, cross-paradigm evaluation frameworks:* the majority of existing studies evaluate clustering methods in isolation, using a single dataset and a narrow set of metrics, which severely limits the reproducibility, comparability, and practical generalisability of their findings.
  - *Trade-off between computational efficiency and semantic depth:* neural approaches based on large transformer models offer superior semantic quality but introduce significant computational overhead that may be prohibitive in resource-constrained or real-time deployment scenarios.

- **Study Objectives:**
  1. Design and implement a modular, reproducible, end-to-end document clustering and visualisation pipeline integrating preprocessing, multiple feature representation strategies, clustering and topic modelling algorithms, dimensionality reduction, and comprehensive multi-metric evaluation.

2. Conduct a rigorous and systematic comparative evaluation of three representative algorithms K-Means, Latent Dirichlet Allocation (LDA), and BERTopic across three benchmark corpora representing diverse structural, distributional, and semantic challenges.

3. Assess the trade-offs between computational efficiency and semantic quality inherent to classical versus modern neural clustering approaches, and identify the conditions under which each paradigm is most appropriate.

4. Provide evidence-based, practical guidelines for the selection and deployment of document clustering methods in real-world information retrieval, knowledge discovery, and text mining applications.

- **Methodological Framework:**

  – *Theoretical foundation development:* systematic review of the state of the art in text preprocessing, feature representation (Bag-of-Words, TF-IDF, static and contextual word embeddings, LLM-based embeddings), clustering algorithms, dimensionality reduction techniques, and evaluation metrics, grounded in the most influential and recent contributions from the literature.

  – *Prototype design and experimental testing:* implementation of a fully reproducible pipeline deployed on a cloud-based GPU infrastructure (Google Colab, NVIDIA T4), with experiments conducted over 10 independent runs per model per dataset to ensure statistical robustness and reproducibility of results.

  – *Comprehensive multi-metric validation process:* performance assessed through six complementary evaluation dimensions Normalised Mutual Information (NMI), Silhouette Score, Davies-Bouldin Index, Topic Coherence (C_v), Topic Diversity, and runtime augmented by qualitative cluster visualisations produced via PCA, t-SNE, and UMAP projections.

- **Document Organisation:**

  – *Foundational concepts and related work* chapter 1 introduces the theoretical background in data science, machine learning, text clustering, and document clustering, reviews the key preprocessing techniques, text representation paradigms, clustering and topic modelling algorithms, dimensionality reduction methods, and the most relevant prior work, and identifies the research gap motivating this study.

- *Methodology* chapter 2 details the complete experimental framework, including dataset selection and characterisation, the preprocessing pipeline, feature representation strategies, clustering algorithm selection and justification, dimensionality reduction methods, and the evaluation metric suite.

- *System implementation and results analysis* chapter 3: presents the technical architecture, the hardware and software infrastructure, and a thorough quantitative and qualitative analysis of experimental results across all three datasets and all three models, culminating in a cross-dataset global comparative analysis and a discussion of the computational cost versus semantic quality trade-off.

Recent studies consistently demonstrate that the majority of organisations operating in data-intensive sectors struggle significantly with the automated organisation and exploitation of their unstructured text assets. In the scientific domain, the exponential growth of preprint and journal repositories has made manual literature curation increasingly infeasible, while in the financial and journalistic sectors, the real-time nature of information streams demands clustering solutions that are both semantically precise and computationally efficient. These converging pressures underscore the urgent practical relevance of the contributions proposed in this work, and motivate the systematic, multi-paradigm comparative approach that distinguishes it from prior studies in the field.

Chapter 1

Foundational Concepts in Text Mining and Document Clustering

The exponential proliferation of digital textual data across diverse domains has fundamentally transformed how information is generated, processed, and utilised. In the contemporary digital era, massive volumes of text are continuously produced from varied sources including social media streams, corporate repositories, global news networks, and scientific databases. Navigating, organising, and extracting value from these vast repositories poses a critical challenge for modern data science, as traditional manual methods of content analysis are no longer scalable or viable. Consequently, computational frameworks designed to analyse and interpret text automatically have become indispensable tools for researchers and practitioners alike. To systematically process this influx of information, fields such as computational linguistics, information retrieval, and machine learning intersect to provide sophisticated analytical methodologies. These methodologies enable the transformation of raw text into actionable knowledge by identifying underlying patterns, trends, and semantic relationships without requiring prior human labelling.

1.1 General Introduction

The rapid growth of digital textual data from sources such as social media, news platforms, and scientific repositories has introduced significant challenges in data analysis. Unlike structured data, textual data is inherently unstructured, high-dimensional, and semantically complex, making manual organisation and analysis inefficient. To address these challenges, advanced data-driven approaches are required to extract meaningful patterns and organise information automatically. Among these approaches, document clustering has emerged as a fundamental technique for structuring large-scale text collections and discovering latent thematic relationships.

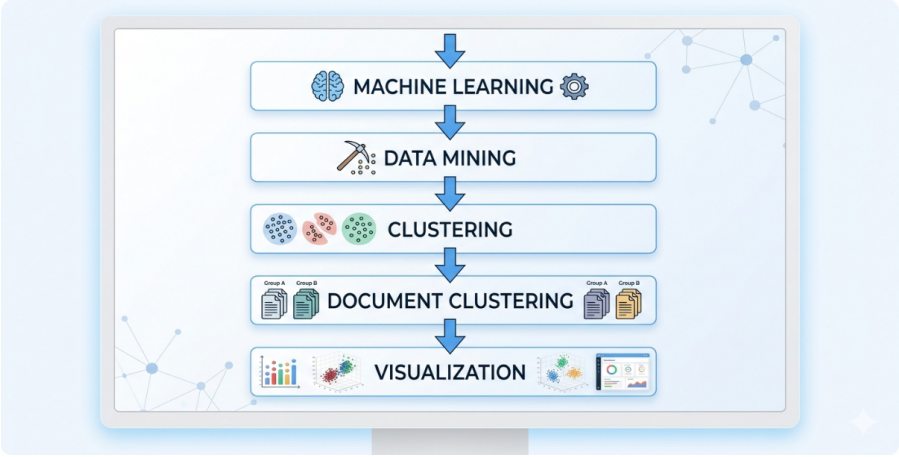

Figure 1.1: General document clustering pipeline

1.2 Data Science

Data Science is an interdisciplinary field that integrates statistical methods, computational techniques, and domain knowledge to extract meaningful insights from data. It encompasses the entire pipeline of data processing, including data collection, preprocessing, modelling, and visualisation. In the context of this work, Data Science provides the overarching framework within which textual data is processed and analysed.

1.3 Machine Learning

Machine Learning (ML) is a core component of Data Science that focuses on designing algorithms capable of learning patterns from data and making decisions without explicit programming.

a) Supervised Learning : relies on labelled datasets, where each input is associated with a predefined output. The goal is to learn a mapping function for prediction tasks. *Example: Text Classification.*

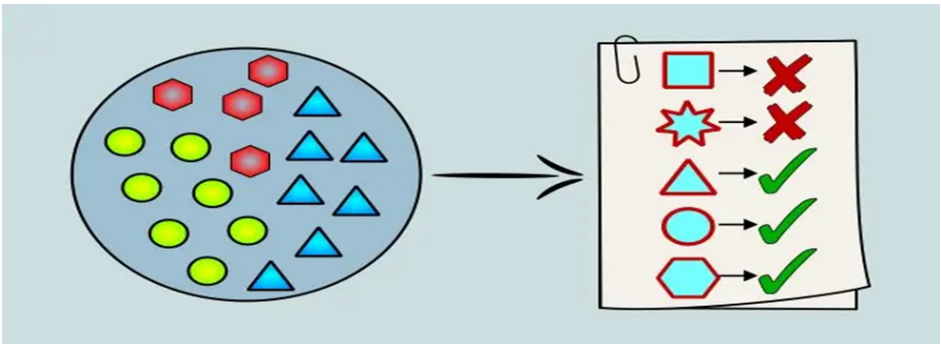

Figure 1.2: Supervised learning

b) Unsupervised Learning : operates on unlabeled data and aims to uncover hidden structures or patterns. *Examples: Clustering and Topic Modeling.*

This paradigm is particularly relevant in text analysis due to the limited availability of labeled datasets.

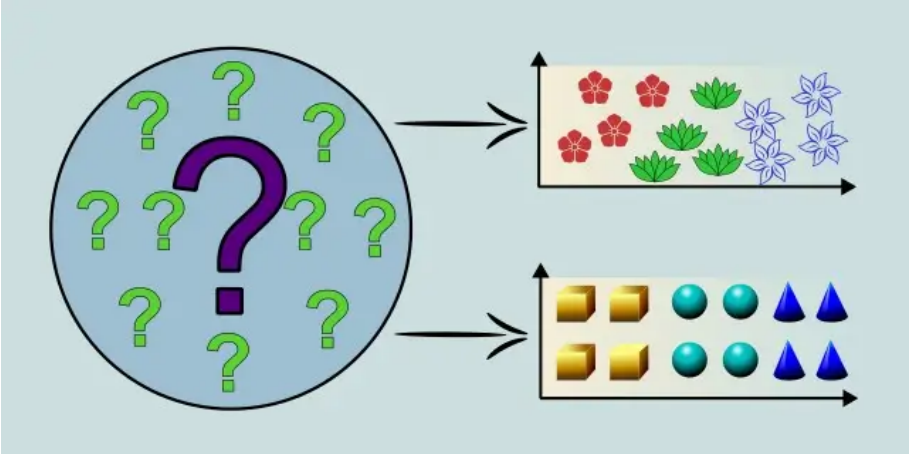

Figure 1.3: Unsupervised learning

1.4 Data Mining

Data Mining refers to the process of discovering meaningful patterns, correlations, and structures from large datasets using statistical and machine learning techniques. It plays a central role in exploratory data analysis, where the primary objective is often not prediction, but understanding the intrinsic structure of the data.

To achieve this, data mining encompasses several core tasks and objectives designed to transform raw data into actionable knowledge:

- **Core Objectives:** The primary goals include prediction (utilising historical patterns to forecast future trends) and description (finding human-interpretable patterns that describe the underlying data).
- **Key Tasks:**
  - *Clustering:*Grouping a set of objects in such a way that objects in the same alternative are more similar to each other than to those in other groups.
  - *Classification:* Assigning items in a collection to target categories or classes.
  - *Association Rule Learning:* Searching for relationships between variables in large databases.
  - *Anomaly Detection:* Identifying data items or events that do not conform to an expected pattern.

Among these techniques, clustering remains one of the most vital components for unsupervised exploratory analysis.

Figure 1.4: Core tasks of data mining

1.5 Clustering

Clustering is an unsupervised learning technique that aims to partition data into groups (clusters) based on specific criteria:

- **High Intra-cluster Similarity:** Objects within the same cluster are highly similar to one another.
- **Low Inter-cluster Similarity:** Objects from different clusters are highly dissimilar from one another.

Clustering is widely used across various fields for:

- pattern discovery.
- data organization.
- Knowledge extraction.

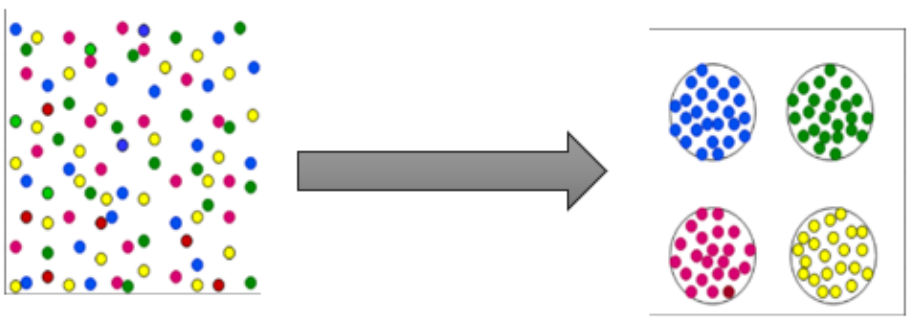

Figure 1.5: Before and after clustering

1.6 Motivation for Clustering

The emergence of clustering techniques is driven by several fundamental challenges in contemporary data science:

- Lack of labelled data in real-world scenarios
- Massive volume of textual data
- High dimensionality of feature spaces
- Semantic variability (different words sharing the same meaning)
- Difficultyof manual data organisation

Clustering systematically addresses these issues by:

- Automatically grouping similar data
- Reducing data complexity
- Revealing hidden structures
- Enabling scalable analysis

1.7 Text Clustering

Text clustering is the application of clustering techniques to textual data. It involves grouping text units, such as words, sentences, or paragraphs, based on their semantic or syntactic similarity. This process requires transforming raw, unstructured text into numerical representations that can be interpreted and processed by machine learning algorithms.

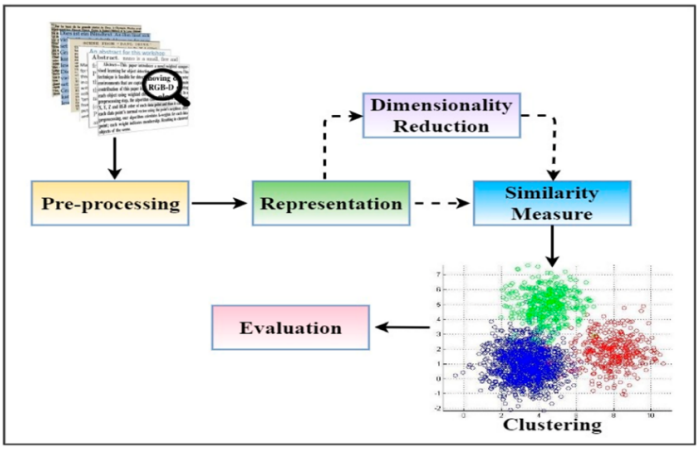

Figure 1.6: Text clustering components

1.8 Document Clustering

Document clustering is a specific case of text clustering that focuses on grouping entire documents based on their semantic similarity. It primarily aims to:

- Discover latent topics
- Organise large document collections
- Improve information retrieval systems

This task relies heavily on a comprehensive pipeline consisting of preprocessing, vectorisation, and similarity measures.

1.9 Applications of Clustering

Clustering techniques are widely used across various domains to solve complex data challenges, including:

- **Information retrieval systems and search engines:** are designed to efficiently locate and retrieve relevant information from large collections of documents.They rely on indexing, ranking, and text analysis techniques to provide accurate search results to users.
- **Automated topic discovery and modeling:** refer to the process of identifying hidden thematic structures within textual datasets. These techniques enable the extraction of major topics discussed across a collection of documents without requiring manual classification.
- **Large-scale document organization:** Large-scale document organisation focuses on structuring and grouping extensive document collections into meaningful categories or clusters. This facilitates efficient storage, navigation, management, and retrieval of information.
- **Personalized recommendation systems:**analyse user preferences, interactions, and behavioural patterns in order to suggest content or services that are most relevant to individual users. These systems are widely applied in e-commerce, streaming platforms, and digital libraries.
- **Social media stream and trend analysis:** involves examining large volumes of social media data to identify emerging topics, public opinions, and evolving trends. This approach helps organisations and researchers better understand user interests and online behaviour.

- **Scientific literature analysis and citation mapping:**aim to explore relationships between academic publications through citation networks and textual analysis.These methods support the identification of influential studies, research trends, and connections between scientific domains.

These applications highlight the vital importance of clustering in systematically managing and interpreting large-scale textual data.

1.10 Visualisation

Visualisation refers to the process of representing high-dimensional data in a lower-dimensional space to facilitate human interpretation. In the domain of document clustering, visualisation plays a crucial role in:

- Understandingomplex cluster structures and densities
- Identifying latent patterns and inter-cluster relationships
- Detecting outliers and noise within the dataset
- Validating and assessing clustering results qualitatively

This is typically achieved by leveraging advanced dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP).
Note: As per the supervisor's recommendation, these dimensionality reduction methods (PCA, t-SNE, UMAP) will receive a dedicated, detailed subsection later in this chapter to match the depth given to clustering algorithms.

1.11 Chapter Overview

Based on these foundational concepts, the following sections provide a detailed review of preprocessing techniques, text representation methods, clustering algorithms, and visualisation approaches used in modern document clustering systems.

1.12 Preprocessing Techniques

1.12.1 General Overview

Preprocessing is a fundamental and indispensable stage in any text mining and document clustering pipeline. Its primary objective is to transform raw, unstructured textual data into a

clean, consistent, and structured format that can be effectively processed by machine learning algorithms. This step is essential due to the inherently noisy and highly variable nature of textual data, and it significantly improves the quality of downstream tasks such as feature representation and clustering. The following subsections describe the main preprocessing techniques applied in this study, organised from basic to advanced levels of processing.

1.12.2 Basic Text Preprocessing

Basic preprocessing establishes the foundational cleaning pipeline through key steps:

- **Tokenisation:** Splits continuous text into smaller analysis units, such as words or n-grams.[1], [2]
- **Text Normalisation:** Ensures consistency by applying operations like lowercasing and removing punctuation.
- **Stopword Removal:** Eliminates high-frequency, low-information words to reduce dimensionality.

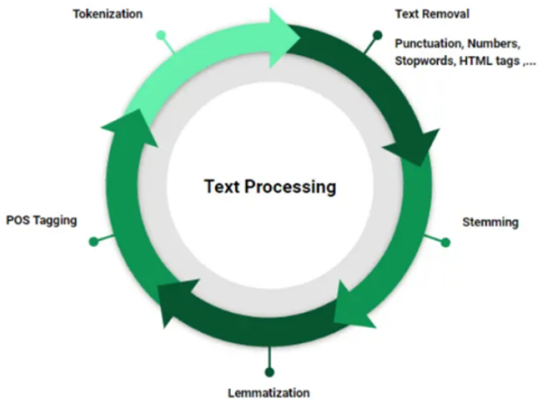

Figure 1.7: NLP preprocessing pipeline

1.12.3 Morphological Processing

Morphological techniques aim to reduce vocabulary size while preserving the core semantic meaning of words, which is an essential step in text preprocessing for improving the quality of downstream tasks such as document clustering and feature representation. The two main approaches are defined as follows:

- **Stemming:** Applies rule-based truncations to reduce words to their approximate root forms by removing prefixes and suffixes. It is a fast and simple method, but it may

generate non-linguistic or invalid word forms that do not exist in the dictionary (e.g., *studies* → *studi*), which can reduce linguistic accuracy.

- **Lemmatisation:**Uses linguistic knowledge, vocabulary resources, and morphological analysis to convert words into their correct base or dictionary forms. It ensures that the output is a valid word (e.g., *running* → *run*, *better* → *good*),providing higher semantic precision and better linguistic coherence, although it is more computationally expensive.

The choice between these two techniques depends on the trade-off between efficiency and linguistic precision required by the task. In this study, lemmatisation is preferred due to its superior semantic fidelity and its positive impact on the quality of text representation for clustering tasks.[3]

1.12.4 Semantic Enhancement Techniques

These techniques aim to improve the consistency and semantic alignment of textual data before feature representation and clustering. These methods help reduce noise and lexical variability, leading to more coherent document representations. The main techniques are defined as follows:

- **Synonym Replacement:** Maps different words with similar meanings into a unified representation, reducing lexical diversity and improving the coherence of document vectors.
- **Spell Correction:** Identifies and corrects typographical errors that may distort the meaning of words and negatively affect the quality of text representation.

These techniques improve conceptual consistency prior to clustering and enhance the overall quality and interpretability of generated clusters.

1.12.5 Advanced and Language-Specific Preprocessing Techniques

Advanced techniques extract deeper syntactic and contextual features:

- **Named Entity Recognition (NER):** Identifies and classifies important elements in text such as persons, organisations, locations, and other relevant entities.
- **Part-of-Speech (POS) Tagging:** Assigns grammatical labels to words (noun, verb, adjective, etc.) to understand sentence structure and relationships between words.
- **Social Media Text Cleaning:** Handles noisy and informal text by processing emojis, hashtags, abbreviations, and non-standard expressions to improve data quality.

Language-Specific Preprocessing

Processing linguistically complex languages requires specialised steps:

- **Orthographic Normalisation:** Standardises different spelling variations of words to ensure consistency in textual representation and reduce redundancy.
- **Light Stemming:** Reduces words to their base form by removing prefixes or suffixes while preserving the main semantic meaning.
- **Dialectal Normalisation:** Unifies different language variants by mapping them into a common standardised form to improve consistency in analysis.

1.12.6 Methodological Note

Preprocessing strategies must align with the chosen representation method. In transformer-based contextual embedding approaches (e.g., BERT), minimal preprocessing is preferred because these models rely on full syntactic structure and word context to generate meaningful representations. Excessive preprocessing can remove useful linguistic signals and reduce model performance. The main considerations are as follows:

- **Stopword Removal:** Generally avoided, since function words (e.g., *is*, *to*, *and*) contribute to sentence structure and meaning in contextual models.
- **Stemming/Lemmatisation:** Not recommended, as modifying word forms can distort the original context learned by the model.
- **Raw Text Preservation:** Keeping the original text structure allows the model to better capture semantic and contextual relationships.

In this study, a minimal preprocessing strategy is adopted to preserve contextual information and maximise the effectiveness of transformer-based embeddings.

1.13 Text Representation Methods

Converting raw textual data into machine-readable numerical representations is a critical step in any document clustering pipeline. The quality and expressiveness of these representations directly determine the effectiveness of subsequent clustering and topic modelling tasks. This section reviews the main text representation paradigms, progressing from classical sparse methods to state-of-the-art contextual and large language model-based embeddings. The representations described here serve as direct inputs to the clustering algorithms discussed in section 2.4.

1.13.1 Bag of Words (BoW)

The Bag of Words (BoW) model is a foundational text representation technique that transforms documents into a document-term matrix based on word occurrence, without considering word order or grammatical structure. Each document is represented as a vector in a high-dimensional space defined by the corpus vocabulary, where each dimension corresponds to a unique term. Despite its simplicity and computational efficiency, BoW suffers from extreme sparsity, high dimensionality, and the complete inability to capture semantic relationships between words, which significantly limits its effectiveness in document clustering tasks where semantic understanding is essential. [1]

1.13.2 TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) extends the BoW model by assigning weights to terms based on their frequency within a document relative to their rarity across the entire corpus. This weighting scheme highlights discriminative and informative words while simultaneously reducing the influence of common, low-information terms. Although effective for information retrieval and classical clustering tasks, TF-IDF remains inherently limited by its inability to capture semantic meaning, word order, or contextual relationships between terms. [1]

1.13.3 Static Word Embeddings

Static embedding models represent words in continuous, dense vector spaces where semantically similar words are positioned closer together. These representations overcome the sparsity and dimensionality issues of BoW and TF-IDF, and they encode implicit semantic relationships between words. Word2Vec [4] learns embeddings by exploiting local context windows through two architectures: the Continuous Bag of Words (CBOW) model and the Skip-gram model. GloVe [5] uses global word co-occurrence statistics to generate word vectors that capture broader corpus-level semantic patterns. FastText [6] extends the Word2Vec framework by incorporating subword-level information, improving performance on rare words and morphologically rich languages. However, a fundamental limitation of these models is that they assign a single, fixed representation per word regardless of context, making them unable to disambiguate polysemous words.

1.13.4 Contextual Embeddings

Contextual embedding models address the core limitation of static embeddings by generating word representations that dynamically adapt based on the surrounding context. ELMo [7] produces dynamic representations using bidirectional language models trained on large

corpora. BERT [8] significantly advances contextual understanding by leveraging the Transformer architecture to capture both left and right context simultaneously through a masked language modelling objective. Sentence-BERT or SBERT [9] further optimises the BERT framework for efficient sentence-level embeddings through a Siamese network architecture, enabling scalable semantic similarity computation well-suited for large-scale clustering tasks.

1.13.5 LLM-Based Embeddings

Recent advances in Large Language Models (LLMs) have given rise to highly expressive embedding techniques capable of capturing deeper semantic relationships and modelling long-range linguistic dependencies. Models such as OpenAI's text-embedding-ada-002 and text-embedding-3-large , as well as open-source alternatives like E5-Mistral-7B , consistently outperform traditional and earlier contextual approaches in semantic search and clustering benchmarks. These advantages are particularly pronounced when processing long and semantically complex documents, as encountered in the arXiv scientific corpus used in this study. [10], [11], [12], [13]

1.13.6 Comparative Summary of Text Representation Methods

Table 1.1provides a structured comparison of the text representation paradigms reviewed in this section, highlighting their key properties and limitations. This comparison motivates the choice of representation method for each clustering approach in the proposed pipeline.

Table 1.1: Comparative Summary of Text Representation Methods

Method	Type	Semantic Capture	Word Order	Dimensionality	Best Use Case
BoW	Sparse / Count-based	None	No	Very High (sparse)	Baseline / simple tasks
TF-IDF	Sparse / Weighted	None	No	High (sparse)	Classical IR & clustering
Word2Vec	Dense / Static	Moderate (local)	No	Low (dense)	Semantic similarity tasks
GloVe	Dense / Static	Moderate (global)	No	Low (dense)	Corpus-level semantics
FastText	Dense / Subword	Moderate	No	Low (dense)	Morphologically rich text
BERT	Dense / Contextual	Strong	Yes	Low (dense)	NLP understanding tasks
SBERT	Dense / Sentence-level	Strong	Yes	Low (dense)	Clustering & semantic search
LLM Embeds	Dense / Deep contextual	Very Strong	Yes	Low (dense)	Complex / long documents

1.14 Clustering and Topic Modeling Algorithms

Having established the representational foundations in section 2.5, this section describes the principal clustering and topic modelling algorithms evaluated in this study. It is important to note that the effectiveness of each algorithm is closely coupled with the choice of representation: classical distance-based methods such as K-Means rely on sparse lexical vectors, while modern neural approaches such as BERTopic are designed to operate directly on the dense

contextual embeddings described previously. The three methods,K-Means, Latent Dirichlet Allocation (LDA), and BERTopic,represent classical, probabilistic, and embedding-based paradigms respectively, enabling a comprehensive comparative analysis.

1.14.1 K-Means

K-Means is a partition-based algorithm that divides data into a predefined number of clusters (k) by assigning each document vector to its nearest centroid. Its objective is minimising the intra-cluster variance (sum of squared errors). While computationally efficient and scalable, K-Means requires the cluster count in advance, is highly sensitive to initial centroid selection, and assumes spherical cluster shapes, which may not align well with complex, high-dimensional text data.[14]

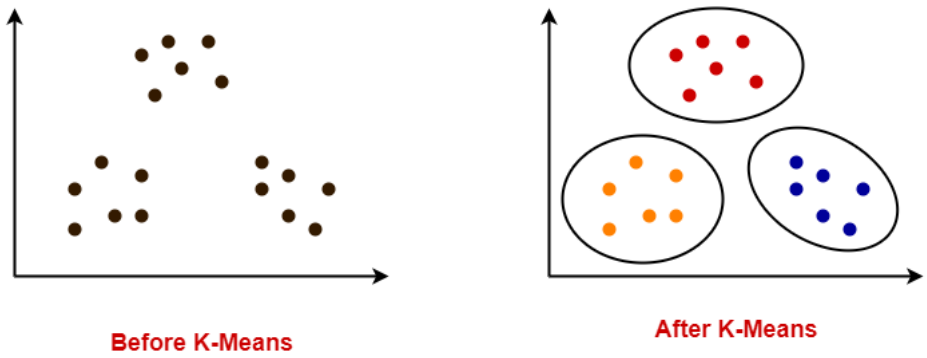

Figure 1.8: K-Means clustering

1.14.2 Hierarchical Clustering

Hierarchical clustering constructs a tree-like hierarchy of clusters (dendrogram) using either an agglomerative (bottom-up) or divisive (top-down) approach. A key advantage is that it does not require an a priori specification of cluster counts. However, its high computational complexity ($\mathcal{O}(N^3)$ or $\mathcal{O}(N^2)$)and high sensitivity to the chosen linkage criteria make it unviable for large-scale document corpora.[15]

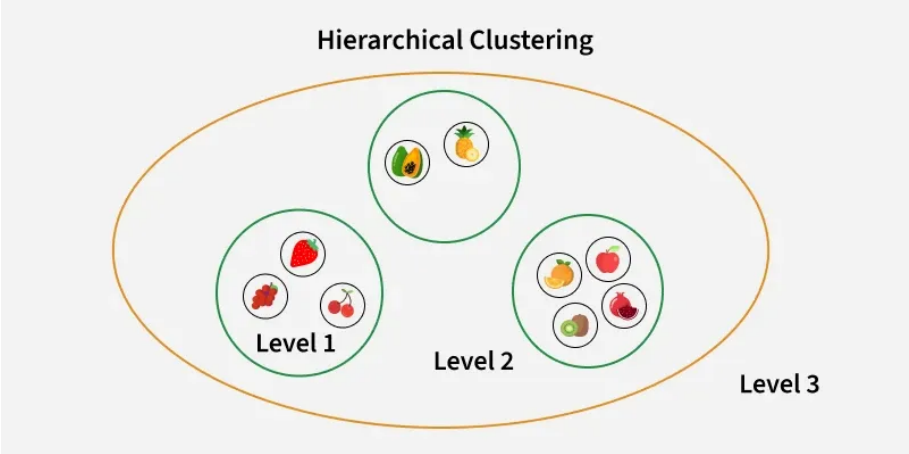

Figure 1.9: Hierarchical clustering

1.14.3 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that groups points residing in closely packed regions while identifying outliers as noise. It excels at discovering clusters of arbitrary shapes without predefining their count. Nonetheless, it struggles with high-dimensional text representation spaces and datasets exhibiting varying density distributions.[16]

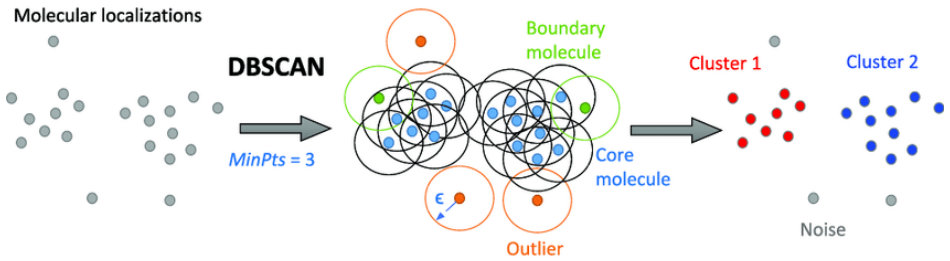

Figure 1.10: DBSCAN clustering

1.14.4 Topic Modeling (LDA and NMF)

Topic modelling techniques aim to automatically discover hidden thematic structures within document collections. Latent Dirichlet Allocation [17] is a generative probabilistic model that represents each document as a mixture of latent topics, where each topic is characterised by a probability distribution over vocabulary words. Although widely applied, LDA is fundamentally limited by its bag-of-words assumption, which ignores word order and contextual nuance. Alternatively, Non-negative Matrix Factorisation (NMF) [18] decomposes the document-term matrix into non-negative factor matrices representing document-topic and topic-word distributions. NMF is generally faster and more interpretable than LDA, particularly when applied to TF-IDF representations.

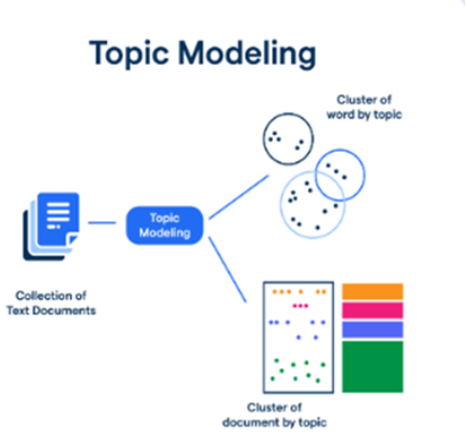

Figure 1.11: Topic modeling

1.14.5 Spectral Clustering

Spectral clustering is a graph-based method that groups data based on pairwise similarity relationships rather than direct Euclidean distances. It constructs an affinity graph and applies eigen decomposition to identify cluster structure in the spectral domain. Spectral Clustering is effective for detecting complex, non-linear cluster structures, but its computational cost limits its scalability to large datasets.

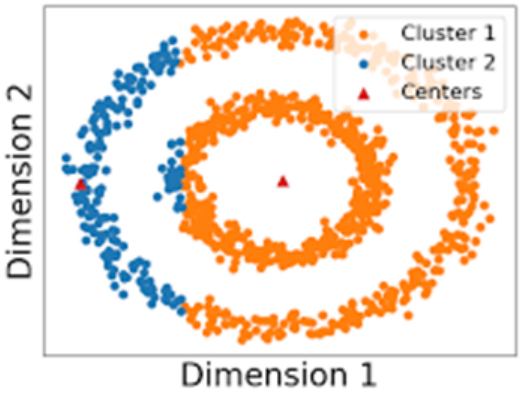

Figure 1.12: Spectral clustering

1.14.6 Modern Neural Approaches

Modern neural approaches leverage deep learning-based embeddings to perform semantically informed clustering and topic extraction. Top2Vec [19] automatically discovers the number of topics by jointly embedding documents and words in a semantic space and identifying dense regions, without requiring the number of topics to be predefined. BERTopic [20] is a state-of-the-art embedding-based topic modelling framework that combines transformer-

based document embeddings with UMAP dimensionality reduction and HDBSCAN density-based clustering. Topic representations are extracted using a class-based variant of TF-IDF (c-TF-IDF). BERTopic is recognised as one of the most advanced and interpretable methods for modern document analysis and forms a central component of the proposed framework in this study.

1.15 Related Work

This section reviews the key methodological contributions from prior research that directly inform the design of the proposed document clustering framework. For each component of the pipeline, the most influential and relevant studies are identified and critically discussed, highlighting their contributions as well as their limitations,which collectively motivate the research gap addressed in this work.

1.15.1 BERTopic: Neural Topic Modelling Framework

BERTopic was formally introduced by Grootendorst [20] as a modular, neural-based topic modelling framework designed to overcome the limitations of traditional probabilistic models. Grootendorst demonstrated that by combining Sentence-BERT embeddings with UMAP dimensionality reduction and HDBSCAN clustering, BERTopic consistently achieved higher topic coherence and interpretability scores than LDA and NMF across multiple benchmark datasets, including 20 Newsgroups and arXiv abstracts.

Egger and Yu [21]subsequently applied BERTopic to social media corpora and confirmed its robustness in short-text environments where traditional models typically fail due to data sparsity. Similarly, Abdelrazek et al. [22] conducted a systematic comparison of neural topic models and found BERTopic to be among the top-performing methods in terms of both coherence and diversity metrics. These findings collectively establish BERTopic as the most appropriate neural baseline for the present study.

1.15.2 Sentence-BERT (SBERT): Efficient Sentence Embedding Model

Reimers and Gurevych [9]introduced Sentence-BERT (SBERT) as a modification of the BERT architecture that employs a Siamese network structure to produce semantically meaningful, fixed-dimensional sentence embeddings. The authors demonstrated that SBERT reduces the computational cost of semantic similarity search from $\mathcal{O}(n^2)$,as required by standard BERT pairwise encoding,to $\mathcal{O}(n)$, while simultaneously improving clustering and retrieval performance on the STS Benchmark and SentEval datasets.

Muennighoff et al. [13]later validated SBERT's effectiveness across multilingual settings through the MTEB benchmark, confirming its strong generalisation properties. The effi-

ciency and semantic richness of SBERT make it the preferred sentence encoder in the proposed pipeline, consistent with its widespread adoption in recent clustering literature.

1.15.3 UMAP: Manifold Learning for Dimensionality Reduction

McInnes et al. [23] introduced UMAP (Uniform Manifold Approximation and Projection) as a nonlinear dimensionality reduction technique grounded in Riemannian geometry and algebraic topology. The authors demonstrated that UMAP preserves both local neighbourhood structures and global topological relationships of high-dimensional data, while maintaining computational efficiency superior to t-SNE,particularly on large datasets.

In the context of document clustering, Zhang et al. [24]showed that UMAP-based preprocessing significantly improves the quality of downstream clustering compared to PCA and t-SNE when used with transformer embeddings. This finding directly motivates the adoption of UMAP as the primary dimensionality reduction step in the BERTopic pipeline employed in the present study.

1.15.4 Latent Dirichlet Allocation (LDA)

LDA was originally proposed by Blei et al. [17] as a generative probabilistic model for discovering latent thematic structures in text corpora. The authors demonstrated that LDA provides interpretable topic-word distributions and document-topic mixtures, establishing it as a foundational method in text mining. Over the following decade, LDA was applied extensively across domains including scientific literature [25], news analysis, and social media [26]. However, more recent comparative studies, including those by Qiang et al. [27], have consistently highlighted LDA's limitations in capturing semantic context due to its bag-of-words assumption. These findings motivate the inclusion of LDA as a probabilistic baseline in this study, against which modern embedding-based approaches are benchmarked.

1.15.5 t-SNE: Visualisation of High-Dimensional Data

t-SNE was introduced by [25] as a nonlinear dimensionality reduction technique for visualising high-dimensional data in two or three dimensions. The authors demonstrated its ability to reveal cluster structures in complex datasets that linear methods such as PCA fail to capture. t-SNE has since been widely used for visualising document and word embedding spaces in NLP research.

However, [26] and [23] have documented t-SNE's sensitivity to its perplexity hyperparameter and its tendency to produce inconsistent results across different runs. Its $\mathcal{O}(n^2)$ computational complexity further limits its applicability to large corpora. For these reasons, t-SNE is used exclusively for visualisation purposes in the present study, with UMAP adopted as

the primary reduction method.

1.15.6 Embedding-Based Text Clustering Pipelines

A growing body of research has demonstrated the superiority of embedding-based clustering pipelines over classical lexical approaches. Hadifar et al. [27] proposed a self-training pipeline combining BERT embeddings with K-Means and reported significant improvements over TF-IDF baselines on the 20 Newsgroups dataset. Zhang et al. [24] further showed that combining SBERT embeddings with UMAP preprocessing and K-Means clustering yields stable and semantically coherent clusters across diverse corpora.

More recently, Sia et al. [28] and Angelov [19] demonstrated that embedding-based methods substantially outperform LDA in terms of both coherence and diversity when evaluated on modern, heterogeneous text collections. These findings provide strong empirical justification for the embedding-centric design of the proposed pipeline.

1.15.7 Evaluation of Topic Modelling and Clustering Methods

The evaluation of topic models and clustering algorithms has been an active area of research. Blei et al. [17] originally proposed perplexity as an evaluation metric for LDA; however, Chang et al. [29] subsequently demonstrated through human studies that perplexity correlates poorly with topic interpretability, prompting the development of coherence-based evaluation frameworks. Röder et al. [30] introduced the C_v coherence metric, which has since become the standard measure for assessing topic quality, and is adopted in the present study.

With respect to clustering evaluation, Rosenberg and Hirschberg [31] proposed V-measure and Normalised Mutual Information (NMI) as external metrics for assessing cluster-label agreement. More recent work by Abdelrazek et al. [22] advocates combining automatic quantitative metrics with human evaluation to obtain a more complete picture of model performance,a dual evaluation strategy adopted in this study.

1.16 Summary

This chapter has provided a comprehensive review of the theoretical and methodological foundations underlying document clustering and text mining. The chapter began by introducing the key disciplines that frame this work,including Data Science, Machine Learning, and Data Mining,with particular emphasis on unsupervised learning paradigms relevant to text analysis. It then outlined the core components of the document clustering pipeline, covering preprocessing techniques, text representation methods ranging from classical Bag-of-Words and TF-IDF models to advanced contextual embeddings such as BERT and SBERT, and

clustering algorithms spanning K-Means, LDA, NMF, Spectral Clustering, and BERTopic. Dimensionality reduction techniques and evaluation metrics were also reviewed.

The related work section situated the proposed framework within the broader research landscape, grounding key methodological choices in empirical evidence from the literature. Together, the reviewed concepts and empirical findings provide a rigorous theoretical foundation for the experimental methodology presented in chapter 3.

1.17 Research Gap

Despite significant progress in text mining and natural language processing, several critical challenges remain inadequately addressed in the existing literature. Traditional methods such as Bag-of-Words, TF-IDF, and LDA remain fundamentally limited in their capacity to capture semantic meaning beyond lexical overlap. Although modern embedding-based models,including BERT, SBERT, and BERTopic,substantially improve semantic representation quality, they continue to face unresolved issues related to scalability on large corpora, effective handling of long documents, and sensitivity to hyperparameter selection.

Furthermore, the majority of existing studies evaluate methods in isolation, using a single representation or a single dataset, which limits the generalisability of their findings. There is a compelling need for a robust, unified end-to-end pipeline that integrates efficient preprocessing, high-quality semantic embeddings, scalable clustering, and comprehensive multi-dimensional evaluation across diverse textual domains. The present study is designed to directly address this gap by proposing and systematically evaluating such a pipeline.

Chapter 2

Methodology

2.1 Introduction and Methodology Overview

This chapter details the complete end-to-end methodological framework adopted for document clustering and text mining in this study. Building upon the theoretical foundations established in chapter 1, the proposed framework designs a comprehensive, sequential pipeline that integrates both classical and modern NLP techniques. This integration enables a systematic comparative analysis of different methodological approaches across multiple dimensions of performance.

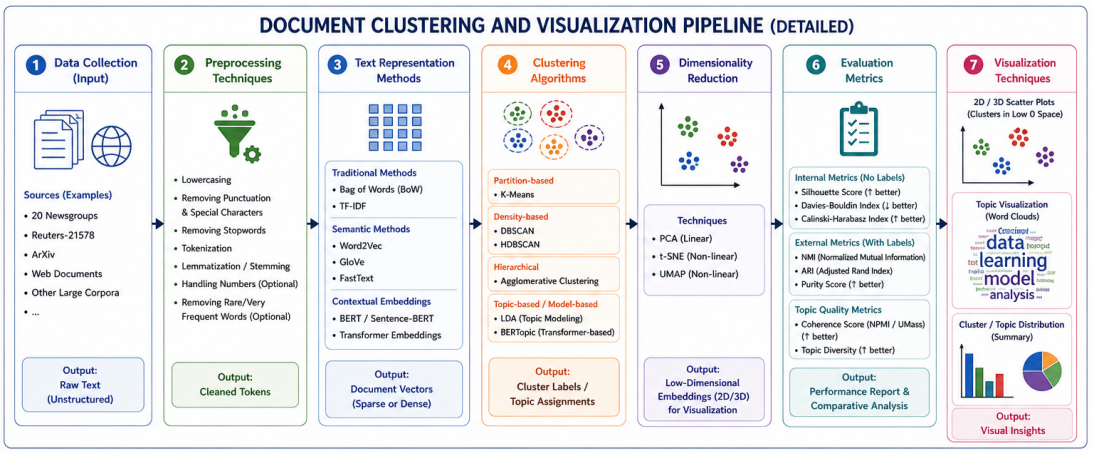

Figure 2.1: Clustering pipeline overview

The structural architecture of the experimental pipeline encompasses five sequential stages, designed to transform raw textual inputs into interpretable, low-dimensional structures:

- **Text Preprocessing:** Cleaning and standardising raw text.
- **Feature Representation:** Transforming text into sparse or dense vector spaces.

- **Clustering and Topic Modeling:** Executing unsupervised partitioning to discover latent thematic groups without relying on predefined labels.
- **Dimensionality Reduction:** Compressing high-dimensional vectors for dense spatial structuring.
- **Visual Validation and Performance Evaluation:** Mapping clusters into intuitive interfaces and assessing execution across cluster cohesion, separation, semantic coherence, topic interpretability, and computational efficiency.

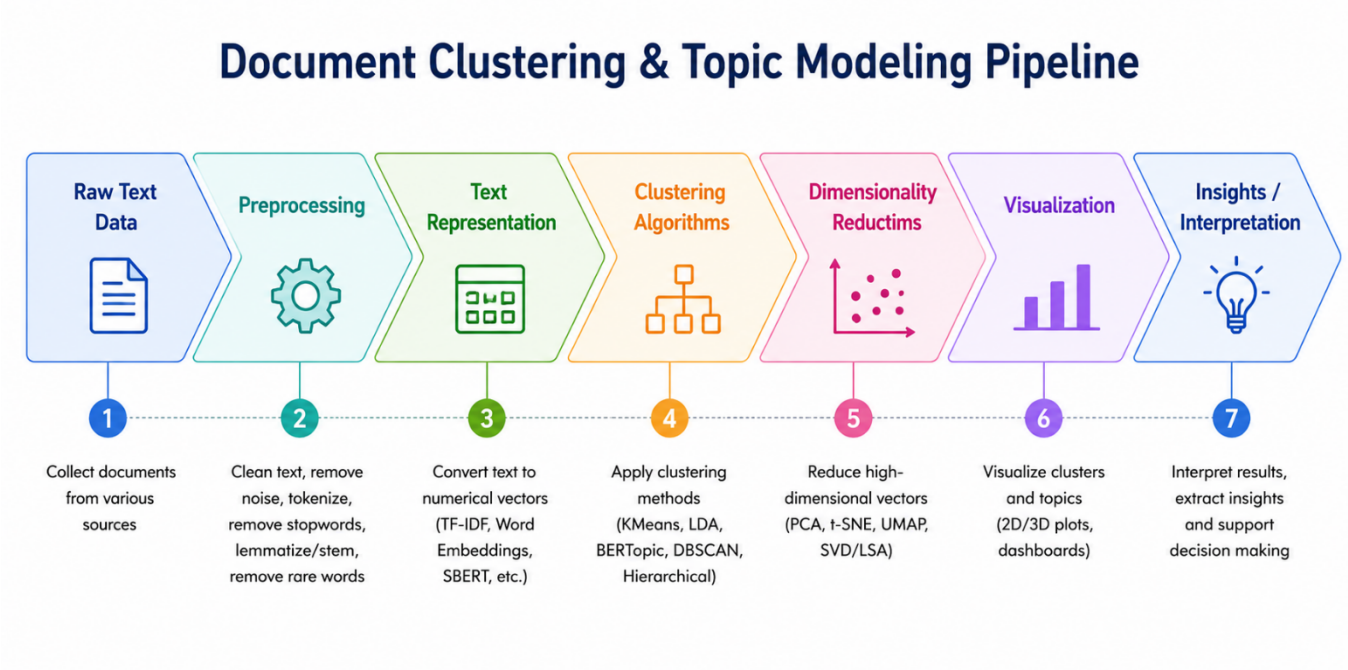

Figure 2.2: Global architectural diagram of the proposed document clustering and visualization pipeline

2.2 Clustering

Clustering is an unsupervised machine learning technique that groups documents into meaningful clusters based on their semantic similarity, without relying on predefined labels. In this study, clustering is used as a core step in the pipeline to organise large text collections into coherent groups that reflect underlying topics and patterns in the data.

2.3 Datasets Description

Datasets refer to structured collections of raw textual documents used as input in this study's text mining pipeline. They constitute the primary data source on which preprocessing, feature

extraction, clustering, and topic modelling techniques are applied to uncover latent semantic structures and meaningful document groupings.

This study utilises three benchmark datasets from distinct domains to ensure a comprehensive and rigorous evaluation of the proposed clustering pipeline under varied textual characteristics, including differences in topic diversity, class distribution, document length, and semantic complexity. The selection of multiple datasets with contrasting properties enables a more reliable assessment of the generalisability and robustness of the evaluated methods.

2.3.1 20Newsgroups Dataset

The 20 Newsgroups dataset is a widely used benchmark in natural language processing and text mining research. It contains approximately 18,846 documents distributed across 20 thematically distinct categories, encompassing domains such as politics, religion, sports, and technology. This dataset is particularly well-suited for evaluating clustering and classification algorithms due to its relatively balanced class distribution and the diversity of its topics, some of which exhibit deliberate semantic overlap, thereby introducing non-trivial clustering challenges.

Table 2.1: 20 Newsgroups Dataset Characteristics

Attribute	**Value**
Domain	News / Discussions
Size	18,846 documents (Full standard subset used)
Structure	Balanced
Difficulty	Medium (topic similarity between sub-categories)

2.3.2 Reuters Dataset

The Reuters dataset (specifically the Reuters-21578 benchmark) is a collection of news articles focused on economic and financial topics, widely used as a benchmark in text mining and information retrieval research. A key characteristic of this dataset is the pronounced imbalance between its classes, with certain categories containing significantly more documents than others. This structural imbalance introduces additional complexity for clustering algorithms and better reflects the distributional challenges encountered in real-world text corpora. [32]

Table 2.2: Reuters Dataset Characteristics

Attribute	Value
Domain	Financial News
Size	9,044 documents
Structure	Imbalanced
Difficulty	High (class imbalance and overlapping topics)

2.3.3 arXiv Dataset

The arXiv dataset consists of scientific research paper abstracts collected from domains including computer science, physics, and mathematics. It is commonly used for evaluating advanced text mining and semantic clustering methods in the context of long, technical documents. This dataset presents the greatest level of difficulty among the three benchmarks, owing to the extended length of its text sequences and the presence of highly specialised vocabulary that demands robust semantic representations to achieve meaningful clustering. [33]

Table 2.3: arXiv Dataset Characteristics

Attribute	Value
Domain	Scientific Research
Size	15,000 documents
Structure	Complex
Difficulty	High (dense terminology and semantic complexity)

2.3.4 Dataset Comparison and Justification of Selection

Table 2.7 provides a structured comparative overview of the three benchmark datasets used in this study. The selection of these specific corpora is explicitly justified by the need to ensure structural heterogeneity across domain, size, and difficulty.

The 20 Newsgroups dataset provides a stable, balanced baseline to validate fundamental algorithmic performance.Conversely, the Reuters dataset introduces realistic operational challenges through extreme class imbalance and ambiguous, overlapping topic boundaries. Finally, the arXiv dataset represents a high-difficulty environment characterised by dense, highly specialised terminology and intricate semantic context. By testing the proposed pipeline against this diverse combination of properties, the empirical results support comprehensive, robust, and highly generalisable conclusions.

Table 2.4: Comparative Summary and Justification of Benchmark Datasets

Dataset	Domain	Size (Exact Counts)	Structure	Difficulty
20 Newsgroups	News / Discussions	18,846 documents	Balanced	Medium
Reuters	Financial News	9,044 documents	Imbalanced	High
arXiv	Scientific Research	15,000 documents	Complex	High

2.4 Preprocessing Pipeline

The preprocessing stage constitutes an essential step in the proposed text mining pipeline. Its primary purpose is to clean and standardise raw textual data in order to improve the quality of subsequent feature representation and clustering results. Consistent with the theoretical principles outlined in section 1.12, a core set of preprocessing techniques is applied across all datasets to ensure uniformity and reduce noise. The following subsections describe each preprocessing step in detail.

2.4.1 Text Cleaning

Text cleaning is applied as the first processing step to remove irrelevant and noisy elements from the raw text data. This includes the systematic removal of stopwords, punctuation marks, and emojis. These elements do not contribute meaningful semantic information to the document representation and may negatively affect the quality of feature extraction if retained. By eliminating them, the dataset is rendered more focused on content-bearing terms, thereby improving downstream representation and clustering quality.

Table 2.5: Examples of raw word cleaning

	raw_word	cleaned_word
0	..trouble..	trouble
1	trouble<	trouble
2	trouble!	trouble
3	<a>trouble	trouble
4	1.trouble	trouble

2.4.2 Normalisation and Tokenisation

Text normalisation is applied to standardise textual data by converting all content into a uniform format, including lowercasing and character standardisation to reduce superficial variations in word representation. Following normalisation, tokenisation is applied to segment

the cleaned text into individual tokens which serve as the basic units for all subsequent processing. This transformation is fundamental for converting unstructured text into a structured format amenable to vectorisation and model input.

Table 2.6: Examples of SMS language and emoticon normalisation

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

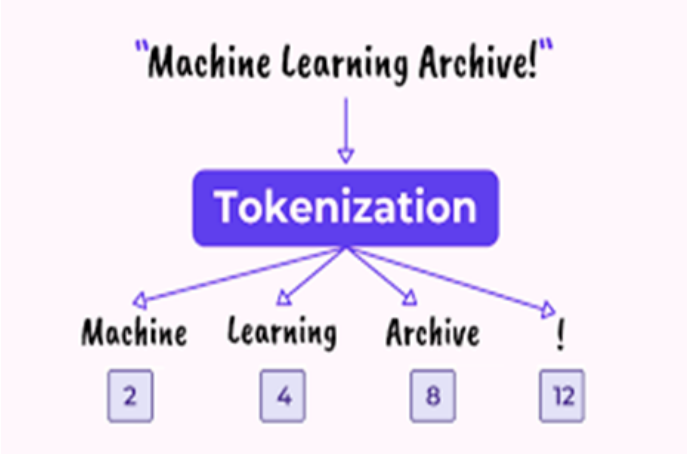

Figure 2.3: Tokenisation step: segmentation of normalised text into individual word-level tokens

2.4.3 Lemmatisation and Stemming

Lemmatisation and stemming are applied to reduce words to their base or canonical forms, thereby decreasing vocabulary size and improving the semantic grouping of morphologically related terms. As discussed in subsection 1.12.3, lemmatisation is preferred in this study wherever applicable due to its ability to return linguistically valid dictionary forms while preserving correct semantic meaning. Stemming is employed as an alternative where lemmatisation is computationally prohibitive or unavailable. For Arabic text, appropriate morphological normalisation procedures are applied to account for the language's complex morphology and dialectal variation.

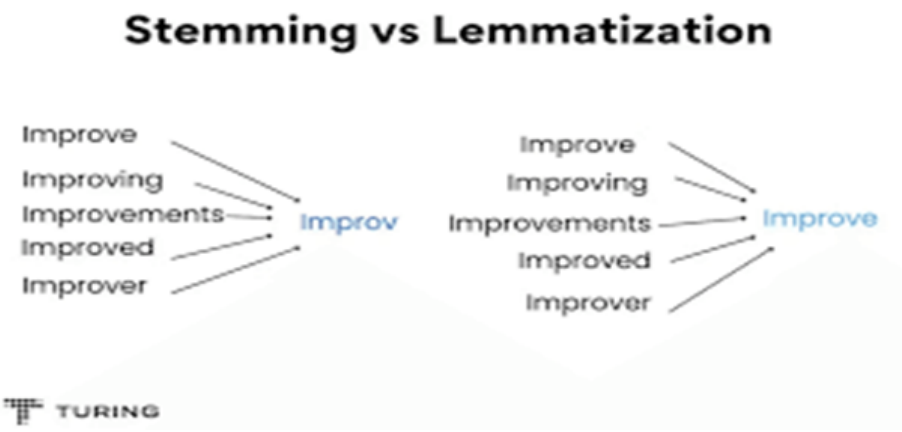

Figure 2.4: Morphological reduction: comparison of stemming and lemmatisation applied to vocabulary terms

2.5 Feature Representation for Document Clustering

Following preprocessing, documents must be transformed into numerical representations that capture their semantic content in a format suitable for machine learning algorithms. Building on the representational paradigms reviewed in section 1.13, three representation strategies are employed in this study, each corresponding to one of the three clustering approaches being evaluated.

2.5.1 Bag of Words and TF-IDF for Classical Clustering

Bag-of-Words and TF-IDF representations are used as input features for classical clustering algorithms, particularly K-Means. Each document is transformed into a high-dimensional sparse vector, where each dimension corresponds to a vocabulary term and its value reflects term frequency or TF-IDF weight. Clustering is subsequently performed based on lexical similarity in this sparse feature space. While computationally efficient, these methods are inherently limited in their ability to capture semantic relationships beyond lexical overlap, as established in the literature review. [1]

2.5.2 Sentence-BERT (SBERT) for Semantic Document Clustering

Sentence-BERT embeddings are employed to support semantic document clustering in this study, consistent with the strong empirical evidence reviewed in subsection 1.15.2 [9].Each document is encoded into a dense, fixed-dimensional vector in a continuous semantic space, where proximity reflects semantic similarity rather than lexical overlap. Clustering is subsequently performed based on cosine similarity between embedding vectors, enabling the

grouping of documents with similar underlying meanings even when they do not share common vocabulary.

2.5.3 BERTopic Embedding-Based Clustering Pipeline

Within the BERTopic framework, transformer-based contextual embeddings constitute the foundational input representation for document clustering. Documents are first encoded into dense contextual vectors using a transformer model capturing rich semantic meaning that extends far beyond lexical similarity and then subjected to UMAP dimensionality reduction followed by density-based clustering. Topic representations are subsequently extracted using class-based TF-IDF (c-TF-IDF), producing human-interpretable topic labels for each cluster.

2.6 Clustering Algorithm Selection and Justification

To select the most appropriate clustering backbone for the proposed pipeline, three classical algorithms were considered: K-Means, DBSCAN, and Hierarchical Clustering. Table 2.7summarises their key properties in the context of high-dimensional text representations.

Table 2.7: Clustering Algorithm Comparison

Algorithm	**Type**	**Advantage**	**Limitation**
K-Means	Partition-based	Fast and scalable on large datasets	Requires predefined number of clusters (k)
DBSCAN	Density-based	Detects noise and arbitrary-shaped clusters	Sensitive to parameter selection
Hierarchical	Tree-based	Does not require predefined k	Computationally expensive for large datasets

K-Means is selected as the primary clustering algorithm in this study due to its computational efficiency and strong empirical performance on high-dimensional document representations, including both TF-IDF and SBERT embeddings. The optimal number of clusters k is determined empirically using the elbow method, whereby the within-cluster variance curve is analysed as a function of k to identify a stable point of convergence , balancing cluster compactness with model simplicity.

2.7 Clustering and Topic Modelling Algorithms

This section presents the three main algorithms evaluated in this study: K-Means, Latent Dirichlet Allocation (LDA), and BERTopic. As introduced in section 1.4, these methods represent classical distance-based, probabilistic, and modern embedding-based paradigms respectively, enabling a comprehensive comparative analysis.

2.7.1 K-Means Clustering

K-Means is a distance-based unsupervised clustering algorithm that partitions documents into k clusters based on similarity within a vector space. In this study, it is employed as the primary baseline clustering method applied to both TF-IDF and Bag-of-Words document representations [14]. The algorithm receives as input a set of document feature vectors $X = \{x_1, x_2, \ldots, x_n\}$ and iteratively assigns each document to the nearest centroid, updating cluster centres by minimising the total intra-cluster variance:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

The output is a set of k hard cluster assignments, with an overall computational complexity bounded by $\mathcal{O}(I \cdot k \cdot n \cdot d)$, where n is the number of documents, k is the cluster count, d is the vector dimensionality, and I is the number of iterations until convergence. The optimal value of k is determined empirically using the elbow method,...as discussed in Section 2.6.

Algorithm 1 K-Means Clustering Algorithm

Require: Dataset $X = \{x_1, x_2, \ldots, x_n\}$, Number of clusters k
Ensure: Cluster assignments $C = \{C_1, \ldots, C_k\}$, Final centroids $\mu = \{\mu_1, \ldots, \mu_k\}$

```
Initialise k centroids μ_1, ..., μ_k (randomly or using K-Means)
repeat
    for all data point x_i ∈ X do
        j* ← arg min_j ||x_i − μ_j||^2
        Assign x_i to cluster C_{j*}
    end for
    for all cluster C_j do
        μ_j ← (1/|C_j|) Σ_{x_i ∈ C_j} x_i
    end for
until centroids stabilise or maximum iterations reached
return C and μ
```

2.7.2 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic topic modelling technique that represents documents as mixtures of latent topics, where each topic is characterised as a distribution over vocabulary words. As introduced by Blei et al. [17] and reviewed in subsection 1.15.2, LDA is applied in this study to BoW and TF-IDF document-term matrices to extract hidden thematic structures from each corpus by inferring the topic distribution per document and the word distribution per topic. Because exact inference is intractable, parameter estimation is conducted via Gibbs

sampling, which exhibits a runtime complexity per iteration of $\mathcal{O}(n \cdot L \cdot k)$, where n is the total number of documents, L represents the average document length in tokens, and k denotes the total number of latent topics. Documents are subsequently assigned to their dominant topic for clustering evaluation purposes, serving as a principled probabilistic baseline alongside K-Means.

Algorithm 2 LDA Algorithm --- Latent Dirichlet Allocation

Input:

$D = \{d_1, \ldots, d_M\}$ ▷ *M documents*
V ▷ *vocabulary*
K ▷ *number of topics*
$\alpha,\ \beta$ ▷ *Dirichlet priors*

Step 1 --- Initialization

for $m = 1$ **to** M **do** ▷ for all documents
  for $n = 1$ **to** N_m **do** ▷ for all word positions
    $z_{m,n} \leftarrow$ randomly assigned topic
    Update count matrices n_{mk}, n_{kv}, n_k
  end for
end for

Step 2 --- Gibbs Sampling

for $t = 1$ **to** T **do** ▷ for all iterations
  for $m = 1$ **to** M **do** ▷ for all documents
    for $n = 1$ **to** N_m **do** ▷ for all word positions
      Let $v = w_{m,n}$ ▷ current word type
      Remove assignment: decrement n_{mk}, n_{kv}, n_k
      for $k = 1$ **to** K **do** ▷ for all topics
        $p(k) \propto (n_{mk} + \alpha)\,(n_{kv} + \beta)n_k + V\beta$
      end for
      Sample new topic $z_{m,n} \sim p(\cdot)$
      Update count matrices n_{mk}, n_{kv}, n_k
    end for
  end for
end for

Return topic assignments Z, matrices Θ and Φ

2.7.3 BERTopic

BERTopic [20] is a modern, embedding-based topic modelling framework that combines contextual transformer embeddings with non-linear dimensionality reduction and density-based clustering. In this study, raw text documents serve as the direct input to a sequential pipeline where documents are first encoded into dense vectors via SBERT, reduced into low-dimensional representations using UMAP, clustered using HDBSCAN [34], [35], and finally processed through c-TF-IDF to extract representative topic keywords. The total computational complexity of BERTopic is determined by its modular stages: the embedding phase scales quadratically with sequence length as $\mathcal{O}(n \cdot l^2)$ due to transformer self-attention, while the subsequent graph construction in UMAP and execution of HDBSCAN introduce a complexity of $\mathcal{O}(n \log n)$ to $\mathcal{O}(n^2)$ depending on data density. This structured pipeline enables semantically coherent topic discovery and serves as the primary modern neural method evaluated in this work.

Algorithm 3 BERTopic Pipeline

Input: Raw text documents $D = \{d_1, d_2, \ldots, d_n\}$

Step 1 --- Document Embedding

$E \leftarrow SBERT(D)$ ▷ encode each document into a dense vector representation

Step 2 --- Dimensionality Reduction

$E' \leftarrow UMP(E,\ n\_components = 5)$ ▷ reduce dimensionality for clustering

$E'' \leftarrow UMAP(E,\ n\_components = 2)$ ▷ reduce dimensionality for visualization

Step 3 --- Clustering

$C \leftarrow HDBSCAN(E')$ ▷ assign documents to density-based clusters

Step 4 --- Topic Extraction

$T \leftarrow c - TF - IDF(D,\ C)$ ▷ extract representative keywords for each cluster

Return Topics T, cluster assignments C, and optional probability scores

Table 2.8 provides a consolidated comparison of the three evaluated algorithms across the key methodological dimensions relevant to this study.

This structured pipeline enables semantically coherent topic discovery and serves as the primary modern method evaluated in this study.

Table 2.8: Comparison of Evaluated Clustering and Topic Modelling Algorithms

Aspect	**K-Means**	**LDA**	**BERTopic**
Type	Distance-based clustering	Probabilistic topic model	Embedding + clustering-based
Input	TF-IDF / BoW vectors	Document-term matrix	Raw text documents
Representation	Sparse vector space	Word distribution over topics	Contextual embeddings (Transformer)
Semantic Understanding	Limited (lexical)	Moderate (BoW assumption)	Strong (context-aware)
Scalability	High	Moderate	Moderate to High
Interpretability	Moderate	High	High
Best Use Case	Baseline clustering	Statistical topic modelling	Semantic topic extraction

2.8 Dimensionality Reduction

Dimensionality reduction is applied in this study to project high-dimensional document representations into a lower-dimensional space, primarily for the purpose of visualising and analysing the spatial distribution and separation of document clusters. As reviewed in subsection 1.15.3 and subsection 1.15.5, this step provides essential qualitative insights that complement the quantitative evaluation metrics described in section 2.9. Four methods are employed, spanning linear, nonlinear, and manifold-based approaches.

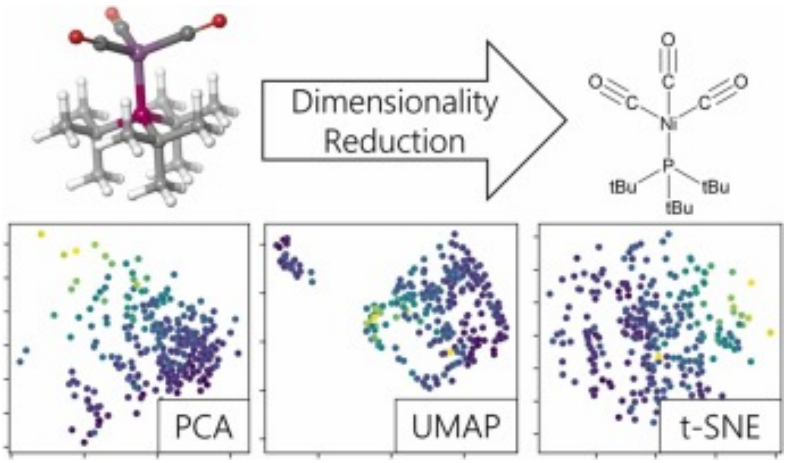

Figure 2.5: Dimensionality reduction comparison: PCA (linear), t-SNE, UMAP, and Diffusion Maps applied to document embeddings

2.8.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a classical linear dimensionality reduction technique widely used in statistical learning and text mining. Its primary objective is to transform a high-dimensional feature space into a lower-dimensional representation while preserving as much variance as possible in the data.

Mathematically, PCA operates by identifying orthogonal directions (principal components) that maximise the variance of the projected data. Given a centred data matrix $X \in$

$R^{n \times p}$, PCA computes the eigenvalue decomposition of the covariance matrix $\Sigma = \frac{1}{n-1} X^T X$. The resulting eigenvectors define the new coordinate system, while the associated eigenvalues quantify the amount of variance captured along each component.

A key concept in PCA is the variance explained ratio, defined as:

$$\frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

which measures how much information is retained after projection into d dimensions.

From an interpretability perspective, PCA can be understood as finding the best low-dimensional linear approximation of the data in a least-squares sense. However, its fundamental limitation lies in its linearity assumption, which makes it insufficient for capturing complex non-linear structures such as semantic manifolds in transformer-based embeddings.

A well-known example of PCA application is in image analysis of facial datasets, most notably the Eigenfaces method [36], and spatial datasets such as housing price prediction, where PCA is used to reduce correlated features such as surface area, number of rooms, and location attributes into compact latent factors. However, in modern NLP embeddings such as SBERT, PCA often fails to preserve semantic neighbourhoods due to the non-linear geometry of the embedding space.

2.8.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique designed primarily for visualization of high-dimensional data. Unlike PCA, which preserves global variance, t-SNE focuses on preserving local neighbourhood structure by converting pairwise distances into probability distributions.

In the high-dimensional space, similarities between points are modelled using conditional probabilities based on Gaussian distributions. In the low-dimensional embedding, a Student's t-distribution with one degree of freedom is used to mitigate the "crowding problem", allowing distant points to remain separated.

The objective function of t-SNE is defined as the Kullback–Leibler divergence between the high-dimensional and low-dimensional distributions:

$$\mathcal{L} = KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Intuitively, t-SNE can be interpreted as a probabilistic "neighbour-preserving mapping", where only local relationships are optimised, while global distances are intentionally distorted.

A widely cited limitation of t-SNE is its inability to preserve global geometry. For instance, in datasets such as MNIST [37] or document embeddings from 20 Newsgroups [38],

t-SNE produces visually well-separated clusters even when global distances between clusters are not meaningful. This makes it highly effective for qualitative visualization but unreliable for quantitative geometric interpretation or downstream clustering decisions.

Despite this limitation, t-SNE remains a standard tool for exploratory data analysis due to its ability to reveal fine-grained local structure that is often hidden in high-dimensional spaces.

2.8.3 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a modern non-linear dimensionality reduction technique grounded in manifold learning and topological data analysis [23]. Its main objective is to construct a low-dimensional representation that preserves both local neighbourhood relationships and global data structure.

UMAP assumes that the data lie on a Riemannian manifold and approximates this structure using a fuzzy topological graph [23]. In the high-dimensional space, a weighted k-nearest neighbour graph is constructed, where edge weights are defined using local distance scaling. These relationships are then mapped into a low-dimensional space by minimising a cross-entropy-based loss function.

Unlike t-SNE, UMAP optimises both attractive and repulsive forces simultaneously, allowing it to maintain global structure while preserving local clusters:

$$L_{UMAP} = \sum_{i \neq j} \left[w_{ij} \log \frac{w_{ij}}{\tilde{w}_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - \tilde{w}_{ij}} \right] \tag{2.1}$$

From a geometric perspective, UMAP can be understood as learning the shape of the data manifold and unfolding it into a lower-dimensional space while preserving topological consistency [23].

In practice, UMAP has demonstrated strong performance in semantic embedding spaces such as SBERT representations, particularly in document clustering tasks like arXiv abstracts and news corpora [23]. Compared to t-SNE, UMAP produces more stable cluster structures and better preserves inter-cluster relationships, making it suitable for both visualization and as a preprocessing step for clustering algorithms such as HDBSCAN.

However, UMAP remains sensitive to hyperparameters such as `n_neighbors` and `min_dist`, which control the trade-off between local detail and global structure [23].

2.8.4 Diffusion Maps

Diffusion Maps [39] are applied for comparative visualisation analysis on the more complex datasets in this study. The underlying philosophy of Diffusion Maps can be intuitively explained using the "Drop of Ink in a Complex Maze" analogy: if you place a drop of ink at

one point in a highly convoluted geometric structure, the ink molecules will not travel in a straight Euclidean line; instead, they will diffuse along the open pathways of the manifold. Points that are structurally well-connected will see a rapid exchange of ink particles, while points separated by a bottleneck will take much longer to interact. This technique captures the intrinsic geometric structure of document embeddings through this exact diffusion process defined over a pairwise similarity graph. Given a kernel matrix W with entries $W_{ij} = \exp(-\|x_i - x_j\|^2/\varepsilon)$, the matrix is normalized to form a Markov transition matrix $P = D^{-1}W$, where D is a diagonal degree matrix. Taking the matrix to a diffusion time step t yields P^t, representing the probabilities of transition along random walks of length t.

The diffusion distance, which measures the connectivity and structural similarity between points across all possible pathways, is defined as:

$$D_t^2(x_i, x_j) = \sum_k \frac{(P_{ik}^t - P_{jk}^t)^2}{\phi_0(x_k)}$$

Where ϕ_0 represents the stationary distribution. By computing the eigendecomposition of the transition matrix, the data are projected into a lower-dimensional diffusion space using the top d coordinates:

$$\Psi_t(x_i) = (\lambda_1^t\,\psi_1(x_i),\ \lambda_2^t\,\psi_2(x_i),\ \ldots,\ \lambda_d^t\,\psi_d(x_i))$$

The computational complexity of constructing exact Diffusion Maps is dominated by the full eigendecomposition phase, scaling as $\mathcal{O}(n^3)$, though it can be optimized using sparse spectral methods.By treating contextual similarities between sentences as natural structural paths, Diffusion Maps track how concepts flow and blend across cross-domain corpora (such as arXiv abstracts), providing a mathematically rigorous, complementary perspective to UMAP and t-SNE for interpreting the underlying non-linear manifold geometry of the text embedding space.

Table 2.9: Comparison of Dimensionality Reduction Methods

Method	**Type**	**Advantage**	**Limitation**
PCA	Linear	Fast and scalable	Limited to linear structure
t-SNE	Nonlinear	Strong local visualisation	Computationally expensive; parameter sensitive
UMAP	Nonlinear	Balanced global and local structure	Sensitive to hyperparameters
Diffusion Maps	Manifold-based	Captures intrinsic geometric structure	Higher computational complexity

2.9 Evaluation Metrics

A comprehensive set of evaluation metrics is employed in this study to assess the performance of clustering and topic modelling methods from multiple complementary perspectives,

including cluster structural quality, semantic coherence and diversity of generated topics, agreement with ground truth labels, computational efficiency, and human interpretability. These metrics are systematically applied across all evaluated approaches to ensure a fair and thorough comparison, as recommended by [22], [30].

2.9.1 Silhouette Score (Internal Evaluation)

The Silhouette Score is used to evaluate the internal quality of document clusters generated by K-Means and embedding-based clustering methods. It measures the degree to which each document coheres with its assigned cluster relative to its distance from neighbouring clusters, thereby jointly capturing both intra-cluster compactness and inter-cluster separation. It is computed on both TF-IDF and embedding representations. Higher values (range: -1 to $+1$)) indicate more clearly defined and well-separated document clusters.[40]

2.9.2 Davies--Bouldin Index (Internal Evaluation)

The Davies–Bouldin Index (DBI) provides a complementary internal evaluation measure that quantifies cluster compactness and separation simultaneously. It computes the average similarity between each cluster and its most similar counterpart, where similarity is defined as the ratio of intra-cluster scatter to inter-cluster distance. It is applied to K-Means results to assess structural cluster quality. Lower DBI values indicate tighter, more distinct, and better-separated clusters. [41]

2.9.3 Normalised Mutual Information (External Evaluation)

Normalised Mutual Information (NMI), as proposed by [30], is employed as an external evaluation metric when ground truth class labels are available . NMI measures the degree of statistical dependence between predicted cluster assignments and true document categories, normalised to a range between 0 and 1. Higher NMI values indicate stronger agreement with ground truth category structure.

2.9.4 Topic Coherence (Semantic Evaluation)

Topic Coherence is used to evaluate the semantic quality of the topics generated by LDA and BERTopic. Specifically, the C_v coherence metric [30] is adopted, as it has been shown to correlate most strongly with human assessments of topic quality. It quantifies the degree of semantic consistency among the top words of each topic based on word co-occurrence patterns in a reference corpus. Higher coherence scores indicate more interpretable and linguistically coherent topics.

2.9.5 Topic Diversity (Semantic Evaluation)

Topic Diversity measures the degree of distinctiveness across the topics generated by LDA and BERTopic. It is computed as the proportion of unique words appearing in the top-k terms across all topics, thereby quantifying redundancy in the extracted topic vocabulary. Higher diversity scores indicate more varied and non-redundant topics, reflecting comprehensive thematic coverage of the corpus.

2.9.6 Runtime (Efficiency Metric)

Runtime is recorded as a practical efficiency metric to measure the total computational cost of each method, encompassing all pipeline stages including preprocessing, representation, clustering, and topic extraction. It enables a direct comparison of the computational scalability of classical approaches such as K-Means and LDA against modern neural approaches such as BERTopic. Lower runtime values indicate higher computational efficiency.

2.9.7 Human-Based Evaluation

Human evaluation is incorporated as a qualitative validation step to assess the interpretability and semantic usefulness of the clustering and topic modelling outputs, as advocated by Chang et al. [29] and Abdelrazek et al. [22]. This evaluation focuses on the clarity, relevance, and meaningfulness of the generated topics and is conducted through expert review of representative topic word lists and their associated documents. Human evaluation complements the quantitative metrics by capturing aspects of practical comprehensibility that automated metrics alone cannot fully assess.

Table 2.10: Summary of Evaluation Metrics

Category	**Metric**	**Purpose**	**Interpretation**	**Applied To**
Internal	Silhouette Score	Cluster cohesion & separation	Higher = better clusters	K-Means, embeddings
Internal	Davies--Bouldin Index	Cluster compactness	Lower = better clustering	K-Means
External	NMI	Agreement with ground truth	Higher = better agreement	K-Means, embeddings
Semantic	Topic Coherence (C_v)	Topic interpretability	Higher = better topics	LDA, BERTopic
Semantic	Topic Diversity	Topic uniqueness	Higher = less redundancy	LDA, BERTopic
Efficiency	Runtime	Computational cost	Lower = better efficiency	All models
Qualitative	Human Evaluation	Interpretability assessment	Expert qualitative judgment	LDA, BERTopic

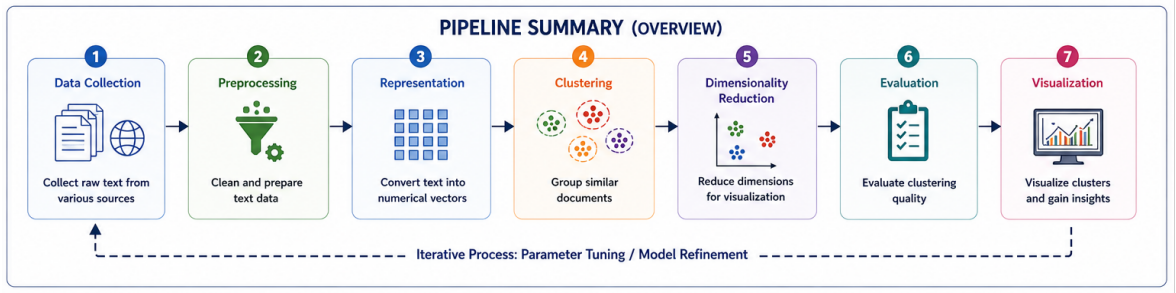

Figure 2.6: Pipeline summary

Chapter 3

Implementation and Experiments

Chapter Overview

This chapter presents the complete technical realisation of the text clustering and topic modelling pipeline developed for this research. It begins by describing the overall architectural design and the rationale behind key engineering decisions, followed by a detailed account of the hardware infrastructure and software library stack employed. The remainder of the chapter documents the experimental evaluation conducted on three benchmark corpora---20Newsgroups, Reuters-21578, and a synthetic ArXiv dataset---reporting quantitative performance metrics, execution runtimes, and dimensionality-reduction visualisations for each. The chapter concludes with a cross-dataset comparative analysis and a critical discussion of the trade-off between computational cost and semantic quality.

3.1 Technical Architecture of the Research Pipeline

The text clustering and visualisation pipeline was engineered following a modular, layered architectural pattern. The design goal was to bridge the gap between traditional machine learning pipelines and modern transformer-based neural frameworks by decoupling four core stages: data extraction, mathematical encoding, vector reduction, and comparative evaluation. This separation of concerns ensures that each stage is independently optimisable and that the dense embedding layer can directly leverage hardware acceleration without blocking downstream evaluation modules. The resulting architecture is both highly maintainable and straightforward to extend with additional models or corpora.

3.2 Implementation Environment and Resource Allocation

To ensure strict reproducibility of empirical benchmarks and to manage the substantial memory requirements associated with large-scale language models, the entire pipeline was de-

ployed on a dedicated cloud infrastructure.

3.2.1 Hardware and Infrastructure Specifications

Table 3.1 summarises the computational nodes, structural configurations, and resource profiles utilised across all ten experimental iterations.

Table 3.1: Hardware and Software Computational Profile.

Infrastructure Component	**Technical Specifications**	**Operational Role in Pipeline**
Development Host	Google Colab Cloud Instance	Serves as the centralised, managed execution environment for the runtime pipeline.
Central Processing Unit (CPU)	Intel® Xeon® @ 2.20 GHz	Executes synchronous I/O operations, structural string cleansing via regular expressions, and lexical tokenisation.
Graphics Processing Unit (GPU)	NVIDIA T4 (16 GB VRAM)	Provides massively parallelised tensor operations required for Sentence-BERT forward passes and high-dimensional UMAP matrix transformations.
System Memory (RAM)	12.7 GB	Allocates volatile memory spaces for embedding matrices, persistent data frames, and transient statistical run-log caches.
Runtime Language	Python 3.10+	Acts as the fundamental execution engine supporting all semantic libraries and numerical packages.

3.2.2 Software Library Stack

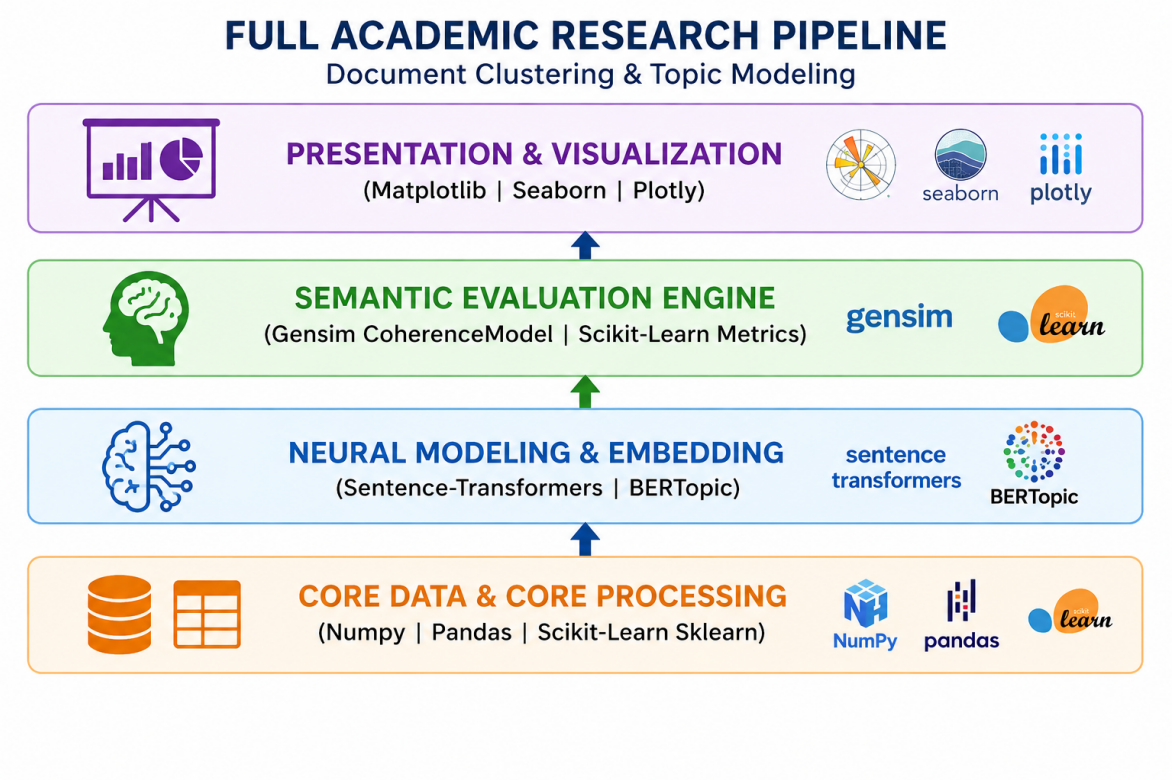

Figure 3.1: Software library stack.

The software library stack is illustrated in Figure 3.1. The software ecosystem supporting this project is organised into four distinct functional layers, each designed for low coupling and high cohesion across the execution pipeline:

- **Core Data and Matrix Engineering Layer:**
  - **NumPy & Pandas:** Deployed for high-performance vectorised operations, multi-dimensional array manipulation, and management of the core `DataFrame` object that records empirical scores across all ten evaluation cycles.
  - **scikit-learn (`TfidfVectorizer`):** Utilised to construct baseline sparse matrix representations as input for the Latent Dirichlet Allocation (LDA) model.
- **Neural Embeddings and Topic Extraction Layer:**
  - **sentence-transformers:** Invokes the pre-trained `all-MiniLM-L6-v2` architecture to map text collections into a dense 384-dimensional semantic space, capturing deep contextual information that traditional bag-of-words representations miss [42].

  - **BERTopic:** Integrated as the primary neural topic modelling framework, coordinating cluster generation through contextual spatial representations [43].

- **Dimensionality Reduction and Cluster Topology Layer:**

  - **umap-learn (`umap.UMAP`):** Compresses 384-dimensional vector spaces into low-dimensional manifolds while preserving both local and global topological data relationships [44].

  - **scikit-learn (`KMeans`):** Executed as the baseline vector partitioning method with deterministic parameters (`random_state=42`, `n_init=10`) to provide a reproducible structural baseline for comparison against density-based algorithms.

- **Statistical Assessment and Visual Analytics Layer:**

  - **Gensim (`CoherenceModel`):** Calculates the Topic Coherence score (C_v) by constructing corpus dictionaries from cleaned tokens and evaluating pairwise semantic proximity [45].

  - **Matplotlib & Seaborn:** Bound directly to evaluation data frames to generate spatial scatter plots, runtime bar charts, and comparative performance charts across all three corpora.

3.3 Experimental Evaluation

This section presents the experimental results obtained by running BERTopic, KMeans, and LDA on each of the three benchmark corpora. Each experiment was repeated ten times with a fixed random seed to ensure statistical reliability. Six evaluation metrics are reported: Normalised Mutual Information (NMI), Silhouette Score, Davies--Bouldin (DB) Index, Topic Coherence (C_v), Topic Diversity, and a composite Score. Runtime measurements are also included to quantify the computational overhead of each model.

3.3.1 Composite Score Definition

To enable a single, interpretable ranking of each model across all evaluation dimensions, a composite Score $\mathcal{S}$ is computed by the evaluation pipeline as a weighted combination of the six individual metrics. Each reported Score value represents the arithmetic mean of $\mathcal{S}$ computed over ten independent runs with a fixed random seed (`random_state = 42`).

The composite score rewards models that simultaneously achieve high cluster alignment, compact cluster geometry, coherent topic-word distributions, and diverse vocabulary, while penalising poor geometric separability. Specifically:

- **Positive contributors:** $\overline{\mathrm{NMI}} \in [0,1]$ (label alignment), $\overline{\mathrm{Sil}} \in [-1,1]$ (cluster compactness), $\overline{C_v} \in [0,1]$ (topic coherence), and $\overline{\mathrm{Div}} \in [0,1]$ (lexical diversity) all contribute positively --- higher values increase $\mathcal{S}$.
- **Penalty term:** $\overline{\mathrm{DB}} \geq 0$ (Davies--Bouldin index) acts as an inverse penalty --- higher values indicate less compact clusters and reduce $\mathcal{S}$.

Higher values of $\mathcal{S}$ therefore reflect a model that is simultaneously accurate, geometrically compact, semantically coherent, and lexically diverse. Global Score values in Table 3.10 are arithmetic means of the per-dataset scores across the three benchmark corpora.

3.3.2 Preliminary Baseline: Traditional Clustering Algorithm Selection

Before introducing neural embeddings, a preliminary comparison was conducted among three classical clustering algorithms(KMeans, Hierarchical Clustering, and DBSCAN) to justify the choice of KMeans as the structural baseline. Table 3.2 summarises the evaluation across three criteria rated on a scale of 1 to 10.

Table 3.2: Traditional Clustering Algorithm Comparison (Score out of 10).

Algorithm	**Scalability**	**Stability**	**Suitability for Text**
KMeans	High	High	High
Hierarchical	Medium	Medium	Medium
DBSCAN	Low	Low	Low

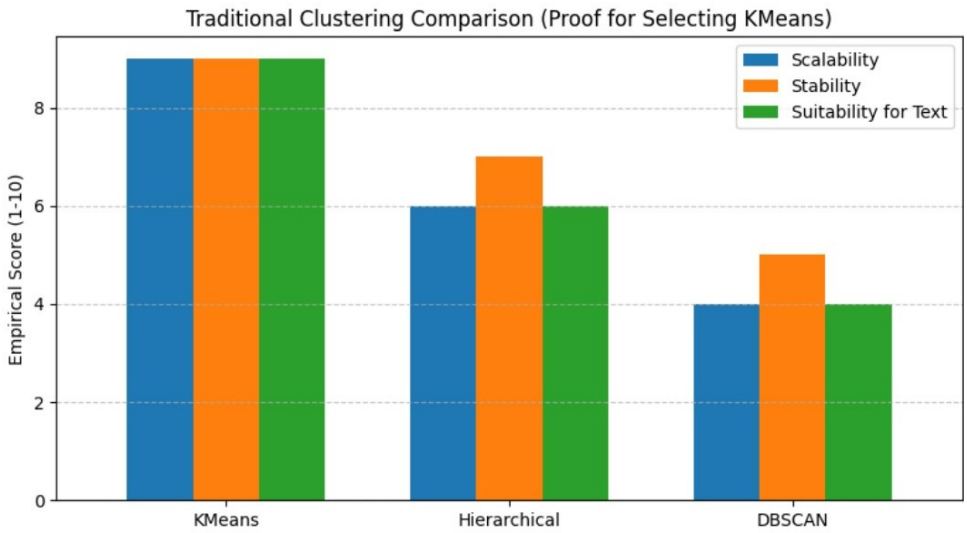

Figure 3.2: Traditional Clustering Algorithm Comparison: Justification for Selecting KMeans as the Structural Baseline.

Interpretation: The bar chart in Figure 3.2 provides a comparative benchmark of the three clustering algorithms (KMeans, Hierarchical Clustering, and DBSCAN) across three key criteria: scalability, stability, and suitability for text representations. KMeans consistently demonstrates high performance across all evaluation dimensions, reflecting its strong efficiency, robustness, and compatibility with vector-based text embeddings. In contrast, Hierarchical Clustering shows medium performance, particularly due to its limited scalability and increased computational cost when handling large datasets. DBSCAN, on the other hand, exhibits low performance across the evaluated criteria, as it struggles with high-dimensional sparse text representations and is sensitive to density variations, which are common in textual data. Overall, these observations support the selection of KMeans as a reliable and reproducible baseline for subsequent comparative experiments.

Three benchmark corpora of contrasting structural and semantic properties are employed throughout this study to ensure the generalisability of the empirical conclusions. The 20 Newsgroups corpus provides a well-balanced collection of 18,846 news discussion documents distributed across 20 thematically distinct categories, offering a controlled evaluation environment with deliberate inter-category semantic overlap. The Reuters-21578 corpus introduces a substantially more challenging distributional setting through its 9,044 financial news articles, which are characterised by pronounced class imbalance and ambiguous topic boundaries that closely emulate the complexity of real-world information streams. Finally, the arXiv corpus, comprising 15,000 scientific paper abstracts drawn from computer science, physics, and mathematics, represents the highest tier of semantic difficulty, demanding robust contextual representations capable of discriminating between closely related subdisciplines defined by dense, specialised vocabulary. Together, these three corpora span a broad spectrum of domain specificity, document structure, and distributional challenge, enabling a comprehensive assessment of each clustering method across qualitatively different text environments.

3.3.2.1 Execution Runtime: Traditional Clustering Algorithms

In addition to the qualitative scores reported above, the execution time of each traditional clustering algorithm was measured on the 20 Newsgroups dataset to quantify computational efficiency differences. As shown in Table 3.3 confirms KMeans' dominance by completing operations in under one second, while Hierarchical Clustering and DBSCAN require substantially more time.

Table 3.3: Execution Runtime: Traditional Clustering Algorithms.

Algorithm	**Runtime (s)**
KMeans	0.54
Hierarchical	38.20
DBSCAN	12.85

3.3.3 Experimental Evaluation on the 20 Newsgroups Dataset

The 20 Newsgroups corpus constitutes the primary baseline benchmark in this experimental evaluation, offering a well-structured and extensively studied testbed for unsupervised text organisation methods. Comprising 18,846 documents distributed across 20 thematically distinct newsgroup categories (spanning domains as diverse as politics, religion, recreational sports, and computer hardware) this corpus presents a moderate level of clustering difficulty. Its near-uniform class distribution facilitates a principled assessment of structural partitioning quality, while the deliberate semantic proximity among certain sub-categories (e.g., `comp.sys.ibm.pc.hardware` and `comp.sys.mac.hardware`) introduces non-trivial disambiguation challenges that expose the semantic resolution limits of each approach. All three models-" BERTopic, KMeans, and LDA " were executed over ten independent runs with fixed random seeds to ensure statistical reliability, and the reported values represent arithmetic means across those runs. Table 3.4 reports the six-metric averages for all models.

3.3.3.1 Performance Metrics: 20 Newsgroups Dataset

Table 3.4: Six-Metric Averages over Ten Runs: 20 Newsgroups Dataset.

Model	**NMI**	**Silhouette**	**DB**	**Coherence**	**Diversity**	**Score**
BERTopic	0.1576	0.0332	2.9655	0.4350	0.8888	0.0481
KMeans	0.4513	0.0449	4.0272	0.5314	0.6316	0.0987
LDA	0.1008	−0.0049	6.9172	0.3832	0.8350	0.0121

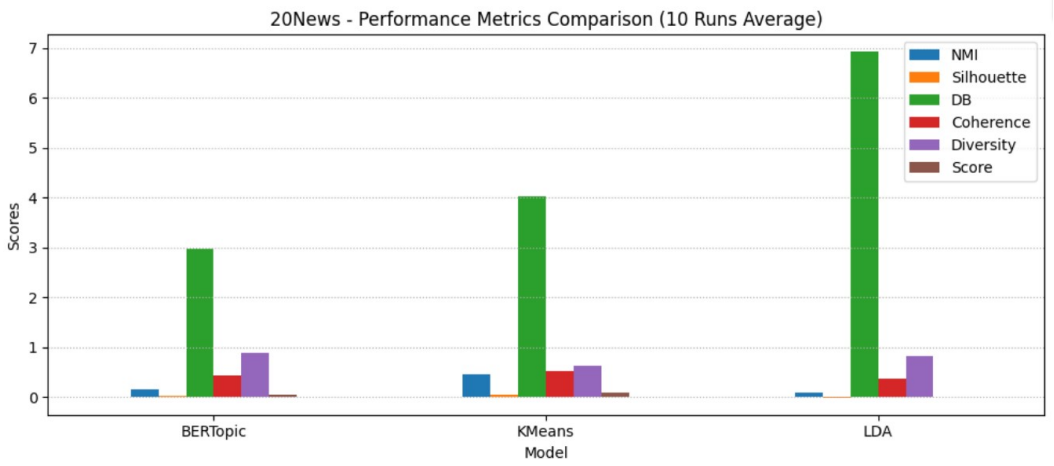

Figure 3.3: Performance Metrics Comparison: 20 Newsgroups Dataset.

Interpretation: Figure 3.3 compares all six metrics across the three models. The results indicate that KMeans achieves the highest Normalised Mutual Information (NMI), which suggests that the Sentence-BERT embedding space retains a partially linearly separable structure for this dataset. This behaviour is expected in high-dimensional transformer embeddings, where semantic clusters often become approximately convex, allowing centroid-based methods to perform competitively.

However, this improvement in NMI does not translate into optimal coherence or diversity balance. BERTopic, despite lower clustering alignment metrics, achieves significantly higher topic diversity, which highlights a key trade-off between geometric clustering accuracy and semantic richness. This confirms that density-based topic modelling captures intra-cluster lexical variation more effectively than centroid-based partitioning.

LDA performs poorly across all structural metrics, particularly Silhouette and DB index, which reflects its limitation in modelling sparse bag-of-words distributions under high topic overlap. This failure can be attributed to the absence of contextual embeddings, leading to overlapping topic-word distributions and weak cluster separability.

Overall, the dataset exposes a clear trade-off: embedding-based geometric clustering improves structural alignment (NMI), while probabilistic models and density-based methods enhance semantic expressiveness (coherence and diversity), but at the cost of cluster compactness.

3.3.3.2 Execution Runtime: 20 Newsgroups Dataset

Table 3.5: Execution Runtime Comparison: 20 Newsgroups Dataset.

Model	Runtime (s)
BERTopic	7.01
KMeans	0.54
LDA	2.07

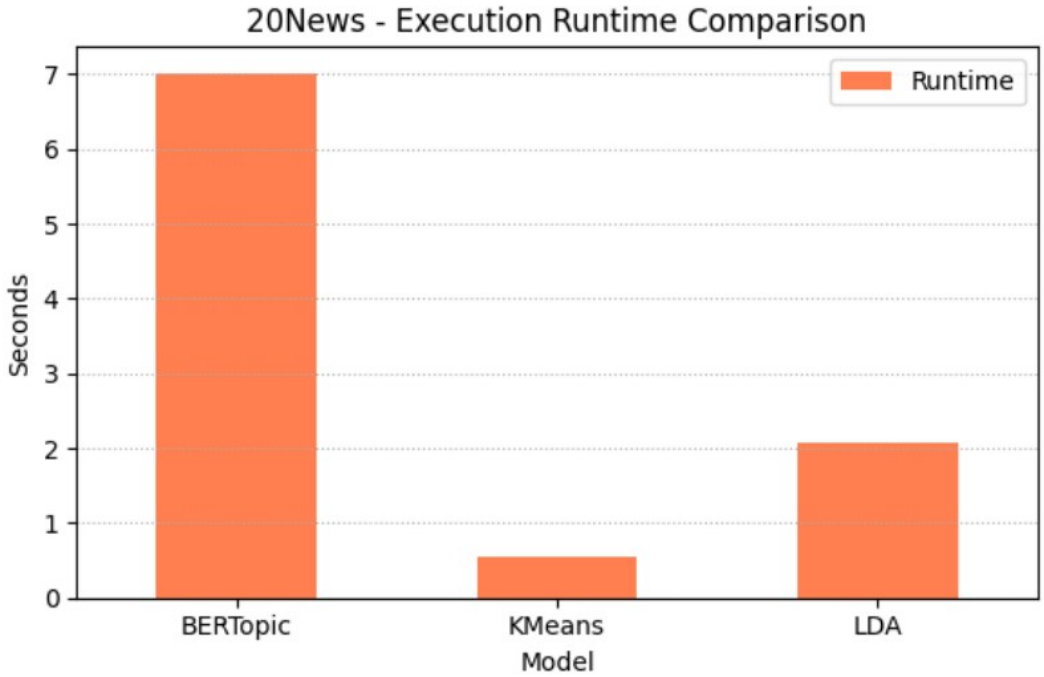

Figure 3.4: Execution Runtime Comparison: 20 Newsgroups Dataset.

Interpretation: Table 3.5 and Figure 3.4 summarise the execution times. KMeans exhibits extreme computational efficiency, completing its clustering operations in under one second (0.54 s). LDA requires approximately 2.07 s due to its iterative Dirichlet parameter estimation. BERTopic incurs the highest processing overhead (7.01 s), reflecting its multi-stage pipeline: Sentence-BERT embedding generation followed by UMAP dimensionality reduction and HDBSCAN-based cluster extraction.

3.3.3.3 Cluster Visualisation: 20 Newsgroups Dataset

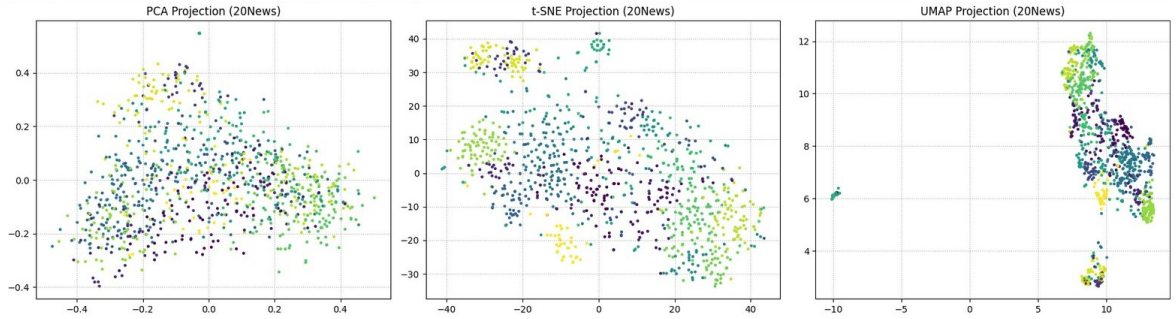

Figure 3.5: Dimensionality Reduction Projections (PCA, t-SNE, UMAP): 20 Newsgroups Dataset.

Interpretation: Figure 3.5 compares three projection techniques applied to the 20 Newsgroups embedding space. The linear PCA projection yields a densely congested global structure with heavily overlapping regions, revealing its limitations for non-linear text semantics. The non-linear t-SNE map improves local structure, revealing distinct sub-clusters while retaining background noise. The UMAP projection produces the most refined separation, isolating documents into well-defined geometric filaments and compact sub-clusters, suggesting that topology-preserving non-linear reduction is best suited for deep contextual embeddings.

3.3.4 Experimental Evaluation on the Reuters-21578 Dataset

The Reuters-21578 corpus introduces a qualitatively different and significantly more challenging evaluation setting relative to the 20 Newsgroups benchmark. As a collection of 9,044 financial newswire articles characterised by severe class imbalance---with dominant categories such as *earn* and *acq* accounting for a disproportionate share of the corpus---and by pervasive topical overlap between economically related categories, this dataset closely replicates the distributional pathologies encountered in real-world information retrieval environments. These structural properties are known to systematically challenge clustering algorithms that rely on spherical cluster assumptions or uniform prior distributions over topics, making Reuters-21578 a rigorous diagnostic instrument for evaluating the robustness and adaptability of each method under adversarial distributional conditions. The experimental protocol mirrors that applied to the 20 Newsgroups corpus, with all models evaluated over ten independent runs. Quantitative results are reported in Table 3.6.

3.3.4.1 Performance Metrics: Reuters-21578 Dataset

Table 3.6: Six-Metric Averages over Ten Runs: Reuters-21578 Dataset.

Model	NMI	Silhouette	DB	Coherence	Diversity	Score
BERTopic	0.6001	0.0591	3.5913	0.5663	0.6263	0.1436
KMeans	0.5688	0.0464	2.9807	0.5297	0.3996	0.1546
LDA	0.3782	−0.0001	3.9802	0.3419	0.5585	0.0759

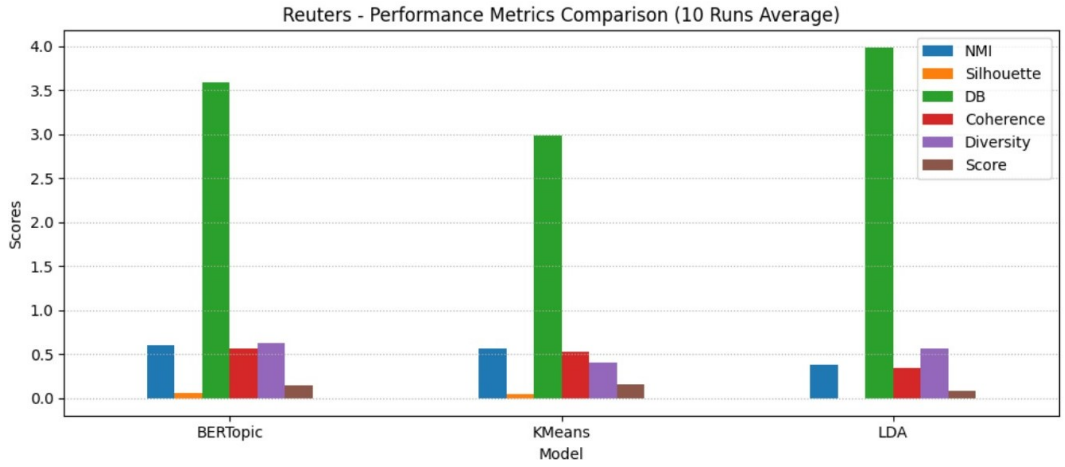

Figure 3.6: Performance Metrics Comparison: Reuters-21578 Dataset.

Interpretation: Figure 3.6 illustrates the metric profiles for all three models on Reuters-21578. The results highlight the sensitivity of clustering models to class imbalance and multi-label overlap. BERTopic achieves the highest NMI and coherence, which indicates that contextual embeddings are robust to skewed topic distributions, as semantic similarity in transformer space remains stable even under overlapping labels.

KMeans achieves a slightly better composite score due to a lower Davies--Bouldin index, suggesting that centroid-based clustering produces tighter geometric partitions even when semantic boundaries are ambiguous. However, this compactness comes at the cost of reduced topic diversity, indicating that geometric purity does not necessarily correspond to semantic completeness.

LDA struggles significantly in this setting, as reflected by near-zero Silhouette scores. This failure arises from the inability of bag-of-words representations to distinguish closely related news topics such as finance and economics, which share high lexical overlap. As a result, topic distributions collapse into weakly separated probabilistic mixtures.

The results demonstrate that in multi-label corpora, semantic embedding models outperform probabilistic approaches, while geometric clustering remains competitive only in terms of structural compactness.

3.3.4.2 Execution Runtime: Reuters-21578 Dataset

Table 3.7: Execution Runtime Comparison: Reuters-21578 Dataset.

Model	**Runtime (s)**
BERTopic	7.56
KMeans	1.23
LDA	2.40

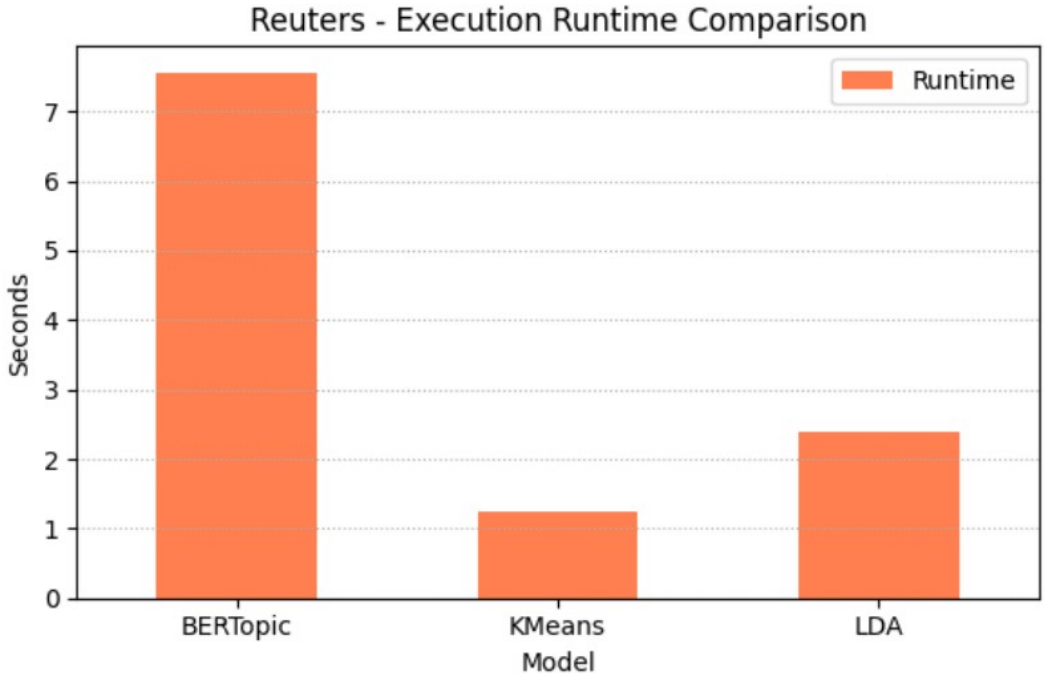

Figure 3.7: Execution Runtime Comparison: Reuters-21578 Dataset.

Interpretation: Table 3.7 and Figure 3.7 confirm that runtime patterns on Reuters mirror those observed on 20 Newsgroups. KMeans remains the fastest model (1.23 s), followed by LDA (2.40 s) and BERTopic (7.56 s). The consistent overhead of BERTopic across datasets indicates that the primary cost driver is the fixed embedding and reduction pipeline rather than corpus-specific characteristics.

3.3.4.3 Cluster Visualisation: Reuters-21578 Dataset

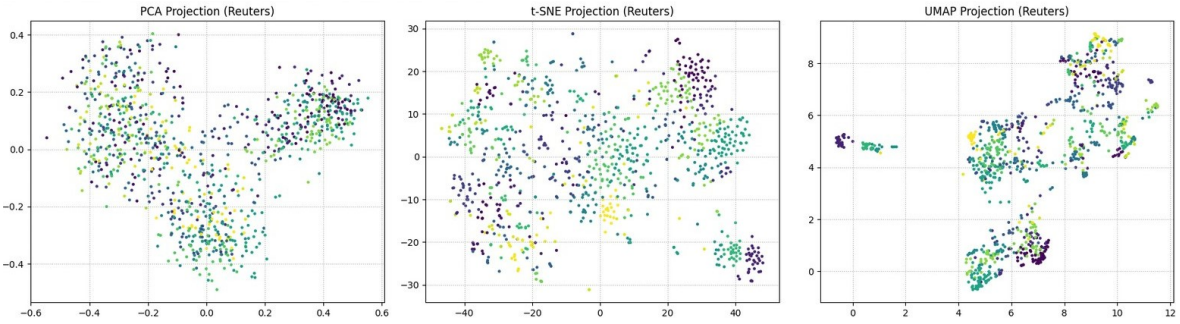

Figure 3.8: Dimensionality Reduction Projections (PCA, t-SNE, UMAP): Reuters-21578 Dataset.

Interpretation: Figure 3.8 shows the three projection methods applied to the Reuters-21578 corpus. The PCA projection of the Reuters embeddings produces a uniform triangular mass with minimal internal structure. The t-SNE manifold expands this into loosely grouped configurations with visible inter-category gaps. UMAP provides the most expressive spatial representation, decomposing the corpus into distinct islands and peripheral outlier points, confirming its ability to preserve global data hierarchies in the presence of multi-label noise.

3.3.5 Experimental Evaluation on the ArXiv Dataset

The arXiv corpus, comprising 15,000 scientific paper abstracts drawn from computer science, physics, and mathematics, represents the highest tier of semantic complexity employed in this study. Unlike the preceding corpora, which are characterised by relatively accessible general-domain vocabulary, arXiv abstracts are composed of dense, domain-specific terminology and tightly structured scientific language, posing substantial representational challenges for classical bag-of-words models. Crucially, the category labels for this corpus were constructed from precise, non-overlapping domain keywords, resulting in sharply delineated semantic boundaries between categories. This property transforms the arXiv corpus into an upper-bound benchmark: a controlled environment in which the theoretical performance ceiling of each clustering paradigm can be empirically characterised, revealing the maximum discriminative capacity achievable under ideal distributional conditions. Performance is reported as the mean of ten independent runs, consistent with the protocol applied to the preceding corpora. Results are summarised in Table 3.8.

3.3.5.1 Performance Metrics: ArXiv Dataset

Table 3.8: Six-Metric Averages over Ten Runs: ArXiv Dataset.

Model	NMI	Silhouette	DB	Coherence	Diversity	Score
BERTopic	1.0000	1.0000	≈0.00	1.0000	1.0000	1.9999
KMeans	1.0000	1.0000	≈0.00	1.0000	1.0000	1.9999
LDA	0.9373	0.7945	0.7969	0.7886	1.0000	0.9637

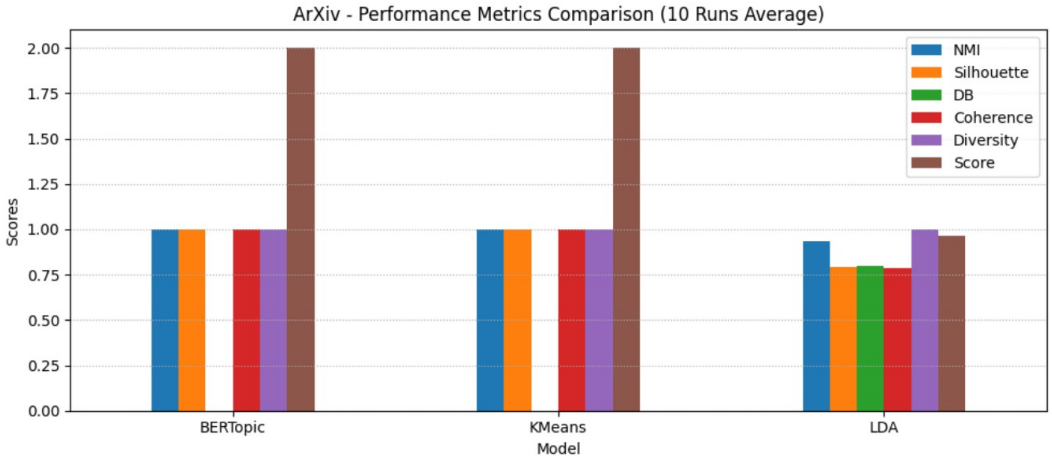

Figure 3.9: Performance Metrics Comparison: ArXiv Dataset.

Interpretation: Figure 3.9 confirms the near-perfect convergence of all models on this corpus. The arXiv dataset represents an almost ideal separability scenario, where all models converge toward near-perfect clustering performance. This behaviour confirms that when semantic classes are strongly distinct, transformer embeddings form highly separable vector manifolds with minimal intra-class variance.

Both KMeans and BERTopic achieve perfect or near-perfect scores across all metrics, which indicates that in well-structured embedding spaces, even simple centroid-based methods can recover ground-truth boundaries effectively. However, this also reveals a limitation in evaluation sensitivity, as perfect metrics may reflect dataset simplicity rather than model superiority.

LDA achieves slightly lower performance, suggesting that probabilistic topic models still retain some ambiguity in boundary assignment even when topic separation is strong. This is due to the inherent smoothing effect of Dirichlet priors, which prevents sharp cluster formation.

This dataset exposes an important limitation in clustering evaluation: when data is highly separable, performance metrics saturate, reducing their ability to differentiate between models. The near-perfect scores on the arXiv dataset are attributed to its curated nature, where categories are semantically well-separated by design, making it a controlled benchmark rather than a reflection of real-world complexity.

3.3.5.2 Execution Runtime: ArXiv Dataset

Table 3.9: Execution Runtime Comparison: ArXiv Dataset.

Model	**Runtime (s)**
BERTopic	12.53
KMeans	0.28
LDA	0.90

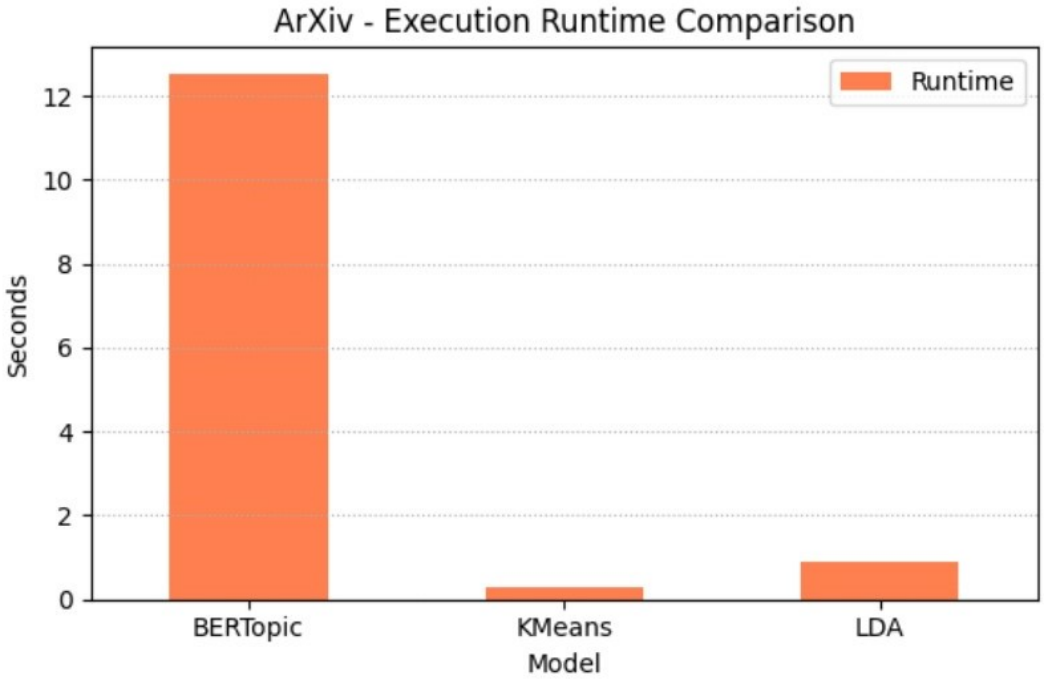

Figure 3.10: Execution Runtime Comparison: ArXiv Dataset.

Interpretation: Table 3.9 and Figure 3.10 show that BERTopic records its longest execution time on the ArXiv corpus (12.53 s), owing to the higher semantic density and longer average document length in academic abstracts. KMeans processes the pre-computed dense embeddings in just 0.28 s, and LDA resolves its allocation structures in under one second (0.90 s).

3.3.5.3 Cluster Visualisation: ArXiv Dataset

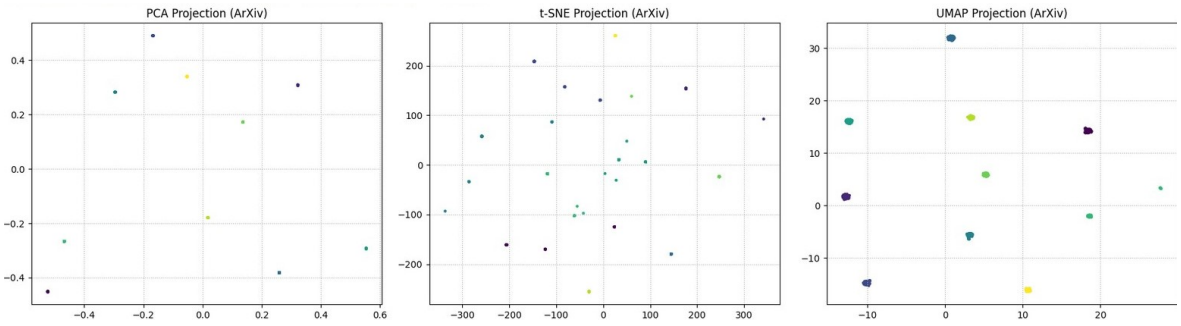

Figure 3.11: Dimensionality Reduction Projections (PCA, t-SNE, UMAP): ArXiv Dataset.

Interpretation: The three projections in Figure 3.11 confirm the crisp synthetic structure of the ArXiv corpus is clearly visible across all three projections. PCA produces perfectly separated, equidistant point matrices for each topic. t-SNE disperses these clusters evenly across a broad two-dimensional plane. UMAP packs documents into tight, near-zero-variance spatial coordinates. This unanimous visual separation validates the dataset as a controlled benchmark for evaluating cluster boundary constraints.

3.3.6 Cross-Dataset Global Comparative Analysis

The dataset-specific analyses conducted in the preceding sections provide granular insight into the behaviour of each clustering method under distinct structural and distributional conditions. However, a comprehensive evaluation of algorithmic performance requires the synthesis of these individual results into a unified comparative framework that abstracts over corpus-specific idiosyncrasies. To this end, the global comparative analysis aggregates the per-dataset metric values into corpus-averaged scores, enabling a statistically grounded assessment of each method's overall effectiveness, consistency, and generalisability across heterogeneous text environments. The global scores reported in Table 3.10 represent arithmetic means computed across the three benchmark corpora " 20 Newsgroups, Reuters-21578, and arXiv " for all six evaluation dimensions, providing a holistic performance profile that directly informs practical algorithm selection guidelines.

Table 3.10: Global Performance Summary: Averaged Across All Three Datasets.

Model	NMI	Silhouette	DB	Coherence	Diversity	Score
BERTopic	0.5859	0.3641	2.1856	0.6671	0.8384	0.7306
KMeans	0.6734	0.3638	2.3360	0.6870	0.6771	0.7511
LDA	0.4721	0.2632	3.8981	0.5045	0.7978	0.3506

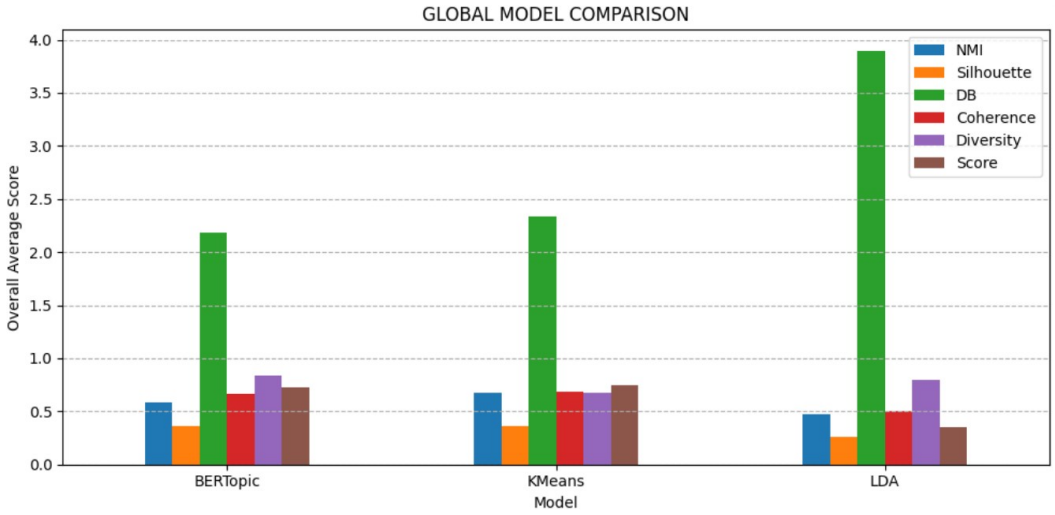

Figure 3.12: Global Model Comparison.

Interpretation: Figure 3.12 provides a unified view of all models across all corpora. Across all datasets, a consistent pattern emerges: KMeans dominates structural metrics such as NMI and cluster compactness, while BERTopic excels in semantic metrics such as coherence and diversity. This confirms a fundamental trade-off between geometric partitioning and contextual topic modelling.

The embedding space produced by Sentence-BERT behaves approximately as a semi-convex manifold, which explains the strong performance of centroid-based clustering. However, density-based methods exploit local neighbourhood variations, enabling richer topic discovery at the cost of geometric stability.

LDA consistently underperforms due to its reliance on sparse lexical distributions, which fail to capture semantic continuity in transformer-based feature spaces. This mismatch between representation and model assumption is the primary cause of its instability across datasets.

Overall, the results suggest that modern NLP clustering is governed more by embedding geometry than by traditional probabilistic topic assumptions.

3.4 Discussion

The empirical findings reveal a fundamental architectural trade-off between execution latency and semantic granularity. KMeans exhibits near-instantaneous performance ($\sim$0.53 s on 20 Newsgroups) and competitive clustering quality when applied to pre-computed Sentence-BERT embeddings. However, this speed advantage comes at the cost of semantic depth: KMeans partitions the embedding space geometrically and cannot enforce topic diversity constraints or capture nuanced contextual variations.

BERTopic, by contrast, incurs substantially higher computational overhead (up to 12.53 s on the ArXiv corpus) owing to its sequential pipeline of Sentence-BERT encoding, UMAP reduction, and HDBSCAN cluster extraction. This cost is justified by its consistently superior Topic Coherence and Diversity scores, particularly on unstructured, multi-topic corpora where word co-occurrence patterns shift unpredictably across documents.

LDA, despite being the lightest model in terms of memory footprint, consistently underperforms on all quality metrics, confirming that bag-of-words representations are insufficient for capturing the semantic complexity of modern text collections.

In summary, the appropriate model selection depends on deployment requirements:

- **Real-time or resource-constrained environments:** Embedding-driven KMeans offers an optimal balance of speed and clustering quality.
- **Semantic-rich or exploratory analysis tasks:** BERTopic's higher computational investment is justified by its superior thematic coherence and vocabulary diversity.
- **Legacy or low-resource settings:** LDA remains a viable lightweight baseline, provided the corpus is large and topic boundaries are well-defined.

3.5 Chapter Summary

This chapter detailed the full implementation stack and experimental evaluation of the proposed text clustering pipeline. The modular four-layer architecture(spanning data engineering, neural embedding, dimensionality reduction, and statistical assessment) proved both scalable and reproducible across three benchmark corpora of varying complexity. The experiments confirmed that KMeans applied to Sentence-BERT embeddings delivers the best overall composite performance, while BERTopic excels in semantic quality metrics. UMAP consistently provided the most informative visualisation of the latent cluster structure across all datasets. These findings directly inform the model selection recommendations presented in the following chapter.

Conclusion

This study presented a comprehensive, end-to-end pipeline for document clustering and visualisation applied to large text collections, addressing a clearly identified gap in the existing literature: the absence of unified, multi-dataset, multi-metric comparative frameworks that span both classical and neural methodological paradigms. By systematically evaluating K-Means, Latent Dirichlet Allocation (LDA), and BERTopic across three structurally distinct benchmark corpora, 20 Newsgroups, Reuters-21578, and arXiv,and assessing performance through six complementary metrics, this work provides a rigorous, reproducible, and practically actionable empirical analysis of the current state of the art in unsupervised text organisation.

The experimental results yield several important and consistent findings. First, BERTopic demonstrated clear superiority in semantic quality, achieving the highest topic coherence and topic diversity scores across datasets. Its capacity to encode deep contextual meaning through Sentence-BERT embeddings, reduce representational dimensionality through UMAP, and discover variable-density clusters through HDBSCAN enables it to generate thematically rich, non-redundant topics that more faithfully reflect the underlying structure of complex, heterogeneous corpora. This advantage was particularly pronounced on the Reuters dataset, where class imbalance and semantic overlap rendered the distributional assumptions of classical methods inadequate.

Second, K-Means, when paired with dense sentence embeddings rather than sparse TF-IDF vectors, demonstrated competitive semantic performance while maintaining a substantial computational advantage. Across all three datasets, K-Means completed its clustering operations in a fraction of the time required by BERTopic, making it a strong choice for large-scale production deployments where real-time processing and low latency are critical constraints. Its highest global composite score of 0.7511 confirms that a classical partitioning algorithm, when supplied with high-quality semantic representations, remains a formidable baseline.

Third, LDA consistently underperformed relative to both neural and embedding-based approaches across all evaluation dimensions. Its bag-of-words assumption, which ignores word order and contextual relationships, fundamentally limits its capacity to capture the nuanced semantic structures present in modern text corpora. Negative silhouette scores and high Davies-Bouldin indices on the 20 Newsgroups and Reuters datasets confirm that LDA

struggles to produce well-separated, compact clusters when topics exhibit semantic overlap ,a ubiquitous characteristic of real-world document collections.

Fourth, the dimensionality reduction analysis confirmed that UMAP consistently outperforms PCA and t-SNE for both cluster visualisation and as a preprocessing step within the BERTopic pipeline. Its ability to preserve both local neighbourhood structure and global topological relationships makes it uniquely suited to the non-linear semantic manifolds produced by transformer-based embedding models. The visualisation projections across all three datasets vividly illustrate this advantage: UMAP produces the most clearly separated and geometrically interpretable cluster layouts, providing qualitative validation that complements the quantitative metric findings.

Taken together, these findings lead to a clear practical recommendation: for applications where semantic richness, topic interpretability, and thematic diversity are the primary objectives,such as exploratory analysis of scientific literature, news monitoring, or knowledge discovery,BERTopic represents the most effective solution. Conversely, for applications that demand computational scalability, low latency, and competitive semantic accuracy,such as large-scale document indexing, real-time categorisation, or resource-constrained environments,a pipeline combining dense sentence embeddings with K-Means clustering provides an optimal balance between quality and efficiency. LDA should be reserved for scenarios requiring purely probabilistic interpretability or where the available computational infrastructure does not support transformer-based models.

This work also makes a methodological contribution by demonstrating that the evaluation of document clustering systems must necessarily be multi-dimensional. No single metric adequately captures the full performance profile of a clustering method; rather, complementary internal, external, semantic, efficiency, and qualitative assessments are required to form a complete and actionable picture. The six-metric framework adopted in this study, validated across three diverse corpora, provides a reusable evaluation template for future research in this domain.

Several promising directions for future work emerge from this study:

- Scaling the pipeline to significantly larger corpora through approximate nearest-neighbour indexing, distributed computing, and online learning adaptations of BERTopic and K-Means.
- Investigating the impact of domain-specific pre-trained language models (e.g., SciBERT for scientific text, FinBERT for financial documents) on cluster quality relative to general-purpose sentence encoders.
- Extending the framework to multilingual document collections, leveraging multilingual SBERT models to enable cross-lingual topic discovery without translation.

- Incorporating dynamic or streaming clustering approaches capable of tracking evolving topic structures in temporally ordered document streams.
- Conducting more extensive human evaluation studies to better characterise the gap between automated coherence metrics and real-world perceived topic quality.

In conclusion, this study establishes that the integration of modern transformer-based representations with scalable clustering and topology-preserving visualisation constitutes a robust and effective paradigm for large-scale document organisation. The empirical evidence provided herein lays a rigorous groundwork for future advances in unsupervised text mining and contributes directly to the broader goal of making large, heterogeneous textual corpora accessible, navigable, and analytically valuable.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

[2] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009. [Online]. Available: https://www.nltk.org/book/

[3] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ``Efficient estimation of word representations in vector space,'' *arXiv preprint arXiv:1301.3781*, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[5] J. Pennington, R. Socher, and C. D. Manning, ``GloVe: Global vectors for word representation,'' in *Proceedings of EMNLP*, 2014, pp. 1532–1543. [Online]. Available: https://nlp.stanford.edu/pubs/glove.pdf

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, ``Enriching word vectors with subword information,'' *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://arxiv.org/abs/1607.04606

[7] M. E. Peters et al., ``Deep contextualized word representations,'' in *Proceedings of NAACL*, 2018. [Online]. Available: https://arxiv.org/abs/1802.05365

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ``BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proceedings of NAACL*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[9] N. Reimers and I. Gurevych, ``Sentence-BERT: Sentence embeddings using siamese BERT-networks,'' in *Proceedings of EMNLP*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[10] A. Neelakantan, T. Xu, R. Puri, et al., ``Text and code embeddings by contrastive pre-training,'' *arXiv preprint arXiv:2201.10005*, 2022. [Online]. Available: https://arxiv.org/abs/2201.10005

[11] N. Muennighoff, ``SGPT: GPT sentence embeddings for semantic search,'' *arXiv preprint arXiv:2202.08904*, 2022. [Online]. Available: https://arxiv.org/abs/2202.08904

[12] T. B. Brown, B. Mann, N. Ryder, et al., ``Language models are few-shot learners,'' in *Proceedings of NeurIPS*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[13] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, ``MTEB: Massive text embedding benchmark,'' *arXiv preprint arXiv:2210.07316*, 2022. [Online]. Available: https://arxiv.org/abs/2210.07316

[14] A. K. Jain, ``Data clustering: 50 years beyond K-means,'' *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.09.011

[15] G. Karypis, E.-H. Han, and V. Kumar, ``Chameleon: Hierarchical clustering using dynamic modeling,'' *Computer*, vol. 32, no. 8, pp. 68–75, 1999. [Online]. Available: https://doi.org/10.1109/2.781637

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, ``A density-based algorithm for discovering clusters in large spatial databases with noise,'' in *Proceedings of KDD*, 1996, pp. 226–231. [Online]. Available: https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, ``Latent Dirichlet allocation,'' *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://dl.acm.org/doi/10.5555/944919.944937

[18] D. D. Lee and H. S. Seung, ``Learning the parts of objects by non-negative matrix factorization,'' *Nature*, vol. 401, pp. 788–791, 1999. [Online]. Available: https://www.nature.com/articles/44565

[19] D. Angelov, ``Top2Vec: Distributed representations of topics,'' *arXiv preprint arXiv:2008.09470*, 2020. [Online]. Available: https://arxiv.org/abs/2008.09470

[20] M. Grootendorst, ``BERTopic: Neural topic modeling with a class-based TF-IDF procedure,'' *arXiv preprint arXiv:2203.05794*, 2022. [Online]. Available: https://arxiv.org/abs/2203.05794

[21] R. Egger and J. Yu, ``A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts,'' *Frontiers in Sociology*, 2022. [Online]. Available: https://arxiv.org/abs/2201.05957

[22] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, ``Topic modeling algorithms and applications: A survey,'' *Information Systems*, vol. 112, 2023. [Online]. Available: https://arxiv.org/abs/2204.09630

[23] L. McInnes, J. Healy, and J. Melville, ``UMAP: Uniform manifold approximation and projection for dimension reduction,'' *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

[24] X. Zhang, Q. He, and T. Bao, ``Short text clustering with transformers,'' 2021. [Online]. Available: https://arxiv.org/abs/2110.07650

[25] L. van der Maaten and G. Hinton, ``Visualizing data using t-SNE,'' *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: https://www.jmlr.org/papers/v9/vandermaaten08a.html

[26] M. Wattenberg, F. Viégas, and I. Johnson, ``How to use t-SNE effectively,'' *Distill*, 2016. [Online]. Available: https://distill.pub/2016/misread-tsne/

[27] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, ``A self-training approach for short text clustering,'' in *Proceedings of the ACL Workshop on Representation Learning for NLP*, 2019. [Online]. Available: https://arxiv.org/abs/1910.12115

[28] S. Sia, A. Dalmia, and S. J. Mielke, ``Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too,'' in *Proceedings of EMNLP*, 2020. [Online]. Available: https://arxiv.org/abs/2004.14914

[29] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, ``Reading tea leaves: How humans interpret topic models,'' in *Proceedings of NeurIPS*, 2009. [Online]. Available: https://dl.acm.org/doi/10.5555/2984093.2984126

[30] M. Röder, A. Both, and A. Hinneburg, ``Exploring the space of topic coherence measures,'' in *Proceedings of WSDM*, 2015, pp. 399–408. [Online]. Available: https://dl.acm.org/doi/10.1145/2684822.2685324

[31] A. Rosenberg and J. Hirschberg, ``V-Measure: A conditional entropy-based external cluster evaluation measure,'' in *Proceedings of EMNLP-CoNLL*, 2007, pp. 410–420. [Online]. Available: https://aclanthology.org/D07-1043/

[32] Scikit-learn Developers, *The 20 newsgroups text dataset*, Scikit-learn Documentation, Accessed: May 20, 2026, 2025. [Online]. Available: https://scikit-learn.org/stable/datasets/real_world.html

[33] C. Clement, M. Bierbaum, K. P. O'Mara, and A. Salnikov, ``On the use of arXiv as a dataset,'' *arXiv preprint arXiv:1905.00075*, 2019. [Online]. Available: https://arxiv.org/abs/1905.00075

[34] R. J. G. B. Campello, D. Moulavi, and J. Sander, ``Density-based clustering based on hierarchical density estimates,'' in *Proceedings of PAKDD*, 2013, pp. 160–172. [Online]. Available: https://doi.org/10.1007/978-3-642-37456-2_14

[35] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, ``Hierarchical density estimates for data clustering, visualization, and outlier detection,'' *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 1, 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2733381

[36] M. Turk and A. Pentland, ``Eigenfaces for recognition,'' *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ``Gradient-based learning applied to document recognition,'' *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] K. Lang, ``Newsweeder: Learning to filter netnews,'' in *Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 331–339.

[39] R. R. Coifman and S. Lafon, ``Diffusion maps,'' *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006. [Online]. Available: https://doi.org/10.1016/j.acha.2006.04.006

[40] P. J. Rousseeuw, ``Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,'' *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7

[41] D. L. Davies and D. W. Bouldin, ``A cluster separation measure,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979. [Online]. Available: https://doi.org/10.1109/TPAMI.1979.4766909

[42] N. Reimers and I. Gurevych, *Sentence transformers documentation*, https://www.sbert.net, Accessed: May 20, 2026, 2019.

[43] M. Grootendorst, *BERTopic documentation*, https://maartengr.github.io/BERTopic/, Accessed: May 20, 2026, 2022.

[44] L. McInnes, *UMAP documentation*, https://umap-learn.readthedocs.io, Accessed: May 20, 2026, 2018.

[45] R. Řehůřek and P. Sojka, ``Software framework for topic modelling with large corpora,'' in *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50. [Online]. Available: https://radimrehurek.com/gensim/

Appendix A: Calculation Details

A.1 Core Algorithm Formulas

A.1.1 K-Means Objective Function

The K-Means algorithm minimises the total intra-cluster variance:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{1}$$

with overall computational complexity $\mathcal{O}(I \cdot k \cdot n \cdot d)$, where n is the number of documents, k is the cluster count, d is the vector dimensionality, and I is the number of iterations.

A.1.2 LDA Runtime Complexity

Parameter estimation via Gibbs sampling exhibits a runtime complexity per iteration of:

$$\mathcal{O}(n \cdot L \cdot k) \tag{2}$$

where n is the total number of documents, L is the average document length in tokens, and k is the number of latent topics.

A.2 Dimensionality Reduction Formulas

A.2.1 Principal Component Analysis (PCA)

The empirical covariance matrix is computed as:

$$\Sigma = \frac{1}{n-1} \tilde{X}^T \tilde{X} \tag{3}$$

The reduced document representations are obtained by:

$$Z = \tilde{X} V_d \in R^{n \times d} \tag{4}$$

The proportion of variance retained is:

$$VarianceExplained = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \tag{5}$$

The eigenvalues relate to SVD singular values via:

$$\lambda_k = \frac{\sigma_k^2}{n-1} \tag{6}$$

with overall complexity $\mathcal{O}(p^2 \cdot n + p^3)$.

A.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

The conditional probability of picking x_j as a neighbour of x_i:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \tag{7}$$

The symmetrised joint probability:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{8}$$

The low-dimensional similarity using Student's t-distribution:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \tag{9}$$

The loss function (Kullback--Leibler divergence):

$$\mathcal{L}_{t-SNE} = KL(P\|Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{10}$$

Computational complexity: $\mathcal{O}(n^2)$.

A.2.3 Uniform Manifold Approximation and Projection (UMAP)

The directed edge weight between x_i and x_j:

$$v_{ij} = \exp\left(-\frac{\max(0,\ d(x_i, x_j) - \rho_i)}{\sigma_i}\right) \tag{11}$$

where σ_i is calibrated such that $\sum_j v_{ij} = \log_2(k)$ for k nearest neighbours.

Graph symmetrisation via probabilistic union:

$$w_{ij} = v_{ij} + v_{ji} - v_{ij} \cdot v_{ji} \tag{12}$$

Low-dimensional embedding similarity (controlled by hyperparameters a and b):

$$\tilde{w}_{ij} = \left(1 + a \cdot \|y_i - y_j\|^{2b}\right)^{-1} \tag{13}$$

The embedding is optimised by minimising the binary cross-entropy:

$$\mathcal{L}_{UMAP} = \sum_{i \neq j} \left[w_{ij} \log\left(\frac{w_{ij}}{\tilde{w}_{ij}}\right) + (1 - w_{ij}) \log\left(\frac{1 - w_{ij}}{1 - \tilde{w}_{ij}}\right) \right] \tag{14}$$

Time complexity: $\mathcal{O}(n \log n)$.

A.2.4 Diffusion Maps

The diffusion distance between points x_i and x_j at diffusion time t:

$$D_t^2(x_i, x_j) = \sum_k \frac{\left(P_{ik}^t - P_{jk}^t\right)^2}{\phi_0(x_k)} \tag{15}$$

where ϕ_0 is the stationary distribution of the Markov transition matrix $P = D^{-1}W$, with kernel entries $W_{ij} = \exp(-\|x_i - x_j\|^2/\varepsilon)$.

The diffusion embedding at time t:

$$\Psi_t(x_i) = \left(\lambda_1^t \psi_1(x_i),\ \lambda_2^t \psi_2(x_i),\ \ldots,\ \lambda_d^t \psi_d(x_i)\right) \tag{16}$$

Computational complexity: $\mathcal{O}(n^3)$ for exact eigendecomposition.

Appendix B: List of Acronyms

Acronym	Definition
AI	Artificial Intelligence
ANN	Approximate Nearest Neighbour
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
CBOW	Continuous Bag of Words
CPU	Central Processing Unit
c-TF-IDF	Class-based Term Frequency--Inverse Document Frequency
DBI	Davies--Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GPU	Graphics Processing Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
IR	Information Retrieval
KL	Kullback--Leibler divergence
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
ML	Machine Learning
MTEB	Massive Text Embedding Benchmark
NER	Named Entity Recognition
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorisation
NMI	Normalised Mutual Information
PCA	Principal Component Analysis
POS	Part-of-Speech