Decoding Culinary Narratives: A Fusion of Transformer Embeddings and Topic Modelling for Recipe Analysis

Natalya Smith

Auckland University of Technology
Auckland, New Zealand
mnb8479@autuni.ac.nz

***Abstract*—In the evolving landscape of Natural Language Processing (NLP), understanding semantic similarity is crucial. This study delves into the potential of Bidirectional Encoder Representations from Transformers (BERT) and Latent Dirichlet Allocation (LDA) for semantically clustering recipes. The integration of BERT and LDA reveals not only ingredients and directions but also deeper culinary narratives. Through k-Means clustering and dimensionality reduction, an automated topic modelling system is developed. Its effectiveness is underscored by multiple evaluation metrics. The preliminary findings highlight a promising convergence of NLP and gastronomy, suggesting further exploration into hybrid methodologies for enhanced efficiency.**

***Index Terms*—Natural Language Processing, Semantic Similarity Analysis, Latent Dirichlet Allocation, Bidirectional Encoder Representations from Transformers, k-Means Clustering, UMAP, t-SNE, Topic Modelling, Dimensionality Reduction, Transformer Architectures, Recipe Analysis**

I. Introduction: Topic Modelling in the Age of Semantic Similarity

With the burgeoning growth of digital textual data, the significance of extracting meaningful insights has never been more paramount. Historically, from the rudimentary keyword-based approaches to sophisticated neural embeddings, the field of topic modelling (TM) has witnessed a transformative evolution. However, as with any evolving domain, it brings forth its set of challenges. Ambiguities in language, polysemy (multiple meanings for a word), and the vastness of the textual landscape pose formidable challenges in discerning accurate thematic structures.

In the vast field of Natural Language Processing (NLP), discerning and categorising thematic structures within textual data is vital. TM, a technique that identifies these hidden thematic structures, is critical in applications [1] from content recommendation to information retrieval. Traditional methods, foundational as they are, often struggle with the nuances and evolving semantics of contemporary languages [2]. This limitation becomes even more pronounced when considering the dynamic and evolving nature of language, where context and semantics play a pivotal role [3].

Beyond the traditional methods of TM, techniques that visualise and represent thematic clusters, like word clouds, have also emerged as valuable tools, especially when combined with advanced embeddings. They offer a window into the prevalent themes, revealing both broader categories and specific nuances, which is crucial in applications from content recommendation to information retrieval.

Emerging from this challenge is the field of semantic similarity (SS) studies. SS aims to bridge the lexical and contextual divide, enabling a deeper understanding of sentiment and nuanced meanings [4], [5]. The advent of transformer architectures, especially BERT (Bidirectional Encoder Representations from Transformers), heralds a new era in this domain. BERT, with its deep bidirectional transformers, provides context-aware embeddings teeming with semantic richness [6], [7].

This research marries the capabilities of BERT with clustering methodologies, notably the k-Means clustering, to pioneer a novel approach to TM. As we scale the intricacies of data, computational complexity often becomes a bottleneck. Therefore, we explore dimensionality reduction techniques to augment the performance of our clustering algorithms. By culminating these methods, an efficient integrated clustering framework, along with a combined BERT-LDA strategy, is conceptualised and brought to fruition for automatic TM. To provide a robust measure of the model's performance, the effectiveness of the proposed system is assessed using various metrics.

Given this context, we pose the question: *"How can BERT embeddings, representative of advancements in SS, be combined with clustering techniques to improve the precision and granularity of TM in vast textual datasets?"*. To address this, we utilise BERT to generate embeddings, followed by dimensionality reduction, and subsequently employ clustering to group semantically cohesive texts. This methodology underscores an integrative approach, positing enhanced TM over conventional paradigms. Preliminary results appear promising, paving the way for future work.

Specifically, the prospects of incorporating hybrid methodologies hold significant potential to enhance the efficiency and capabilities of our model. This paper will delve into the methods, experiments, and findings, beginning with our exploration on the RecipeNLG dataset, investigating the synergy between BERT and Latent Dirichlet Allocation (LDA) in TM.

The personal interest in this topic as well as why it is relevant and interesting to others, can be explained as follows. SS is a crucial aspect of NLP, especially when dealing with vast textual datasets. Understanding the semantic relationships between different pieces of text, therefore, can lead to more accurate and meaningful analyses. BERT represents one of the latest advancements in the field of NLP, especially in capturing context-aware embeddings. Its ability to understand the nuances and intricacies of language makes it, indeed, a valuable tool for any research aiming to delve deeper into textual data.

TM is a technique used to identify hidden thematic structures within textual data. Clustering, as a method to group similar data points, can be a valuable tool in TM. The question of how to effectively combine BERT embeddings with clustering techniques to improve TM, for example, is both relevant and timely. As textual data continues to grow exponentially, the need for precision (accuracy) and granularity (detail) in TM becomes paramount. This study aims to answer how combining BERT embeddings with clustering techniques can meet these objectives.

Moreover, the emphasis on "extensive textual datasets" not only highlights the challenges and complexities of dealing with vast amounts of data but also underscores the need for tools and methodologies that provide both a birds-eye view and granular insights into the thematic structures. Moreover, our initial findings suggest that the combined use of embeddings, clustering, and visualisation techniques provides a multi-faceted understanding of extensive textual datasets.

This combination not only highlights the challenges and complexities of handling large data sets but also underscores the need for methodologies offering both overarching and detailed insights. Given the above points, the research question is not only valid but also aligns with current trends and challenges in the field of NLP. It sets the stage for a study that aims to harness the latest advancements in the field to address longstanding challenges in TM, and offer detailed insights into the content.

The practical implications of our research are manifold. Enhanced TM can revolutionise sectors like content recommendation, sentiment analysis, and even targeted advertising, ensuring users receive more relevant and contextually apt content. In academic research, it can aid in literature reviews, helping researchers quickly identify central themes in vast bodies of text.

The remainder of this paper is organised as follows. Section II provides a comprehensive review of traditional TM techniques, the evolution of transformer architectures, and previous integrative approaches. Building upon this foundation, Section III outlines the methodology, and describes the dataset used, the pre-processing steps and the process. Section IV provides a critical analysis of the results. It provides comparison with traditional methods and insights derived from the integrated approach. Section V provides a reflection on the project's journey. Section VI summarises the findings, contributions, and suggests potential future work. The link to the Python code can be found in Section VII.

II. Literature Review

A. Semantic Similarity

SS is a measure of how alike two data points are in terms of their meaning or content [50]. It is often used in the context of NLP and textual data to quantify the similarity between words, phrases, or documents based on their semantic content [50], [51]. Several papers offer comprehensive overview of the different approaches to SS, as well as the challenges[1]of measuring SS.

For example, various approaches, the evaluation methods, the applications and challenges of SS are discussed in [8]. Similarly, a survey of the SS measures is presented in [9]. It also discuss the problem of polysemy (where a word can have multiple meanings) and synonymy (where two words have the same meaning).

Different approaches to measuring SS based on word embeddings (WEs) (a technique for representing words as vectors in a high-dimensional space) are discussed in [10]. This survey also highlights the challenges of measuring SS based on WEs: the problem of word sense disambiguation (where a word can have multiple meanings) and of synonymy. An overview of various similarity and relatedness metrics (categorised based on their underlying methodologies and applications) is presented in [11]. The highlighted gaps include the need for more fine-grained evaluation methodologies and the exploration of context-aware similarity measures.

Due to their interpretability and computational efficiency, the traditional metrics are argued to be suitable for tasks like text classification and clustering [8], [9], [10], [11], and [12]. However, such measures exhibit limitations in semantic understanding and word order sensitivity: they lack the capacity to discern intricate semantic relationships and struggle with capturing context, potentially limiting their performance in tasks where such nuances are vital [13].

Neural-network (NN)-based similarity scores have emerged as a potent alternative, showcasing profound semantic comprehension and contextual awareness [14]. These approaches, often leveraging deep learning architectures, can model complex relationships among words and documents. Indeed, by considering word order and contextual information, they excel in capturing semantics beyond lexical overlaps.

For example, [14] proposes a BiGRU-BERT model for SS learning. The model leverages contextual embeddings from BERT and bidirectional gated recurrent units (BiGRUs) to

[1]Of measuring SS in a heterogeneous world, where documents can be written in different languages, use different vocabularies, and have different structures.

enhance similarity measurement. The findings show that the proposed model achieves improved performance on sentence similarity tasks compared to traditional methods. This paper provides an opportunity for additional investigation into alternative neural architectures and how they influence the process of similarity calculation.

[15] developed Siamese recurrent networks for learning SS and applied them to tasks such as recognising textual entailment. These networks provided effective representations for measuring SS. [16] explored cross-lingual textual entailment and content synchronisation, which involves measuring similarity between news articles in different languages. The study applied machine learning techniques to measure textual entailment and align news articles across languages.

A supervised learning approach to train universal sentence representations using natural language inference data is introduced in [17]. These trained sentence representations can be used for various tasks, including sentence similarity measurement. While effective, the study did not extensively explore the challenges and limitations of the model's generalisation to various domains and languages.

The Universal Sentence Encoder, a neural model that generates semantically meaningful sentence embeddings for various NLP tasks is introduced in [18]. This model provided high-quality sentence embeddings, which can be used for sentence similarity measurement and other downstream tasks. Nevertheless, the research did not extensively examine potential concerns associated with selecting a particular architecture, addressing biases in the training data, or fine-tuning for specific domains.

Sentence-BERT, a model that employs siamese BERT networks to learn sentence embeddings optimised for SS in [19] outperformed existing methods for SS, providing more effective and interpretable sentence embeddings. This study, however, did not extensively address the challenges of fine-tuning Siamese architectures and potential difficulties in interpretability for more complex models, providing opportunities for further work.

Despite offering a compelling solution for tasks requiring deep semantic understanding, such as SS measurement [17], [18], [19], [27], and [28], the effectiveness of these models comes at the cost of increased complexity and resource demands [20]. Indeed, NN models require substantial labelled data and computational power for training, potentially limiting their applicability in resource-constrained environments [21]. Moreover, their intricate architectures may be challenging to interpret and fine-tune for specific domains [22], [23], [24], [25], and [26].

A notable body of research has combined different similarity metrics to enhance SS estimation in various NLP tasks. For example, [29] proposed a method that combines traditional and NN-based metrics for document clustering. By weighting and aggregating different metrics, they achieved improved accuracy in clustering, showcasing the potential of leveraging the strengths of diverse metrics.

Similarly, [30] explored the fusion of cosine similarity and NN-based embeddings to enhance sentiment classification. Their findings revealed that combining metrics led to better sentiment prediction accuracy, indicating that leveraging lexical and contextual information can yield superior results in tasks involving subjective content. Jaccard similarity and cosine similarity are used as features alongside NN-based approaches for SS classification in [31]. The study explores the contribution of these metrics to the overall performance of the model.

In [32], the authors conduct extensive experiments on benchmark datasets and provide insights into the performance of different metrics. They identify that no single similarity measure consistently outperforms others across all tasks, indicating the need for careful selection based on specific applications. The study also highlights the challenges of addressing the trade-offs between efficiency and effectiveness in similarity measurement.

For question answering, [33] show how the combined approach achieves better performance in single-relation question answering tasks. A multi-prototype vector-space model that combines traditional distributional similarity metrics with NN-based techniques is presented in [34]. This paper introduces the concept of using multiple prototypes per word to capture different aspects of meaning, and demonstrates improvements in word similarity tasks.

The authors of [35] propose a joint learning approach that combines traditional distributional similarity measures with latent semantic clustering, which improves word sense disambiguation accuracy. Study in [36] proposes a method that combines traditional distributional similarity metrics with sense embeddings for word sense disambiguation. The combination of these techniques is shown to enhance the accuracy of sense disambiguation.

However, while several studies have reported promising results, there are instances where combined metrics did not consistently outperform individual ones. For example, [37] investigated the combination of various similarity metrics for word sense disambiguation and found that performance gains were not consistently significant across all metrics. This suggests that the effectiveness of combined metrics might vary based on the specific task and dataset characteristics. [38] explored metric fusion for cross-lingual SS, revealing that while combining metrics could enhance performance, certain combinations were more effective than others, highlighting the importance of careful selection and evaluation of combined metrics.

Based on this literature, several similarities and dissimilarities between traditional and NN-based metrics can be identified in terms of:

- *Methodology*: if traditional metrics rely on simple mathematical formulas that operate on raw word counts or binary presence/absence vectors, NN-based approaches learn complex, distributed representations of words and sentences by considering their context and relationships within the training data;

- *Data Representation*: traditional metrics often use sparse representations of text, which may not capture semantic nuances effectively. NN-based approaches use dense, continuous representations that capture semantic meanings more accurately. For instance, the embeddings learned by BERT or similar models encode rich semantic information, allowing for more precise similarity measurements;
- *Semantic Understanding*: traditional metrics may struggle with semantic understanding, treating words as isolated units without considering their context. NN-based models excel at capturing semantic relationships and contextual meanings. They understand synonyms, antonyms, and context-specific word usage, improving similarity scores that align better with human perception;
- *Performance and Flexibility*: NN-based approaches generally offer higher performance and flexibility. They can be fine-tuned on specific tasks, making them adaptable to various NLP applications. Traditional metrics are limited by their simplistic nature and lack of context awareness, which may result in sub-optimal performance in complex NLP tasks.

B. Latent Dirichlet Allocation as a Traditional TM Technique

TM has been a cornerstone in the realm of text analysis. Traditional methods like LDA, introduced in [39], is a prominent technique in extracting thematic structures from vast textual datasets [39], [40]. For example, [1] delves into the use of TM, discussing the potential of LDA to uncover semantic relationships in text, making it a valuable tool for personalised content recommendations: indeed, designed to uncover hidden semantic structures in large volumes of text, LDA is invaluable for various NLP applications [40]. LDA is a generative probabilistic model that assumes each document is a mixture of topics and a topic is a distribution over words [39], [40]. It can model complex corpora with multiple topics per document and has a solid probabilistic foundation. However, it requires the number of topics to be specified a priori and can be computationally intensive.

C. Evolution of Word Embeddings

WEs, or high-dimensional vectors where words with similar meanings are located close to each other in the vector space, are a type of word representation that captures the semantic meaning of words based on their context in a text[2]. For example, [2] provides insights into their properties, especially in the context of human lexical knowledge and word associations. The potential of WEs to model semantic relationships between words and their ability to capture the nuances of human lexical representations are highlighted. WEs have revolutionised the way textual data is represented and understood. Early methods like Word2Vec, GloVe, and FastText transformed words into vectors, capturing semantic relationships in the process [7]. These embeddings, by considering the context of words, have been instrumental in tasks ranging from sentiment analysis to machine translation. Their significance lies in their ability to capture semantic nuances, bridging the gap between lexical content and contextual meaning, as highlighted in [5] and [6].

[2]The idea behind WEs is to transform words into dense vectors, where the position of each word in the vector space is determined by its context, i.e., the words that frequently appear nearby.

D. Transformer Architectures

Transformers

The advent of transformer architectures has resulted in a significant paradigm shift in the field of NLP [17], [54]. These architectures have had a profound impact on various NLP tasks since. Unlike their predecessors, transformers do not rely on recurrent or convolutional layers. Instead, they utilise a mechanism known as self-attention, which allows them to weigh the significance of different words in a sentence relative to a given word [52]. This ability to consider the entire context, or the full sequence of words, in a sentence is what sets transformers apart. One of the key breakthroughs in this domain is the introduction of BERT. BERT's ability to capture such rich context means that its embeddings (vector representations of words) are semantically dense. These embeddings can capture nuances, idiomatic expressions, and even sarcasm to some extent [46], [47], [48] and [54]. As a result, BERT has set new performance benchmarks for a plethora of NLP tasks, including but not limited to, sentiment analysis, question-answering, and named entity recognition.

BERT in TM

BERT has found application in TM, adding a new dimension to the field. Its embeddings, characterised by their depth and semantic richness, have been integrated into TM processes [46], [47], [48], and [55]. When coupled with conventional clustering techniques, BERT embeddings introduce a level of granularity in topic extraction that was previously challenging to achieve [46], [47], and [48]. The key advantage of leveraging BERT in TM stems from its profound contextual understanding. Unlike traditional methods that may overlook subtle contextual cues, BERT excels in capturing the intricate nuances of language. This contextual awareness enables it to discern and differentiate topics more effectively, contributing to more precise and informative TM outcomes.

E. Clustering Techniques in NLP

Clustering is a fundamental technique in data analysis, used to group similar data points together [56], [57], [58], and [59]. In the context of NLP, traditional clustering methods like K-means have found applications in various areas. For instance, K-means clustering can group similar documents together [56], making it useful for tasks like TM. By clustering text data based on content similarities, NLP systems can efficiently organise and retrieve information [57].

The connection between SS and cluster analysis lies in the idea that data points that are semantically similar should belong to the same cluster [58]. This connection can be established using various clustering techniques, such as K-Means, hierarchical clustering, or spectral clustering. By incorporating SS into the clustering process, one can discover patterns and

associations that may not be apparent when using traditional features alone [59].

In the realm of cluster analysis, SS can serve as a valuable feature for grouping data points: instead of relying solely on numerical or categorical attributes, clustering algorithms can leverage SS to identify clusters based on the underlying meaning or content of the data [60]. This approach can lead to more meaningful and interpretable clusters, especially in domains where the semantics of data play a crucial role.

F. Integrated Frameworks

The fusion of advanced NLP techniques, such as BERT embeddings with clustering, significantly elevates the performance of NLP tasks. While BERT excels in capturing contextual nuances, it can benefit from the structured organisation provided by clustering for tasks like TM [61]. This combination ensures richer textual representations, fostering coherent TM, precise document categorisation, and tailored content recommendations. Such integrations bolster semantic comprehension and adaptability across varied textual data. Notably, research exemplified by [52] and [54] underscores the advantages of merging distinct NLP techniques, highlighting the promise of enhanced SS estimation and paving the way for robust, multi-faceted NLP solutions.

TM and BERT Embeddings

Although traditional TM approaches, such as LDA, have been foundational in extracting thematic structures from textual data, the increasing complexity and dynamism of language necessitate models that can delve deeper into semantic nuances. Recognising this, the integration of BERT embeddings with traditional TM techniques like LDA has surfaced as a promising direction in NLP research.

A study by [47], for example, presented a fusion method combining LDA and BERT for text classification. They used LDA to assign topics to sentences and then integrated Word2Vec vectors, weighted by word probabilities, with BERT embeddings. This approach showcased enhanced sentence semantics and outperformed traditional fusion methods, especially for short text datasets.

In the context of financial news, [48] developed a BERT-LDA joint embeddings model for topic clustering. Combining BERT's contextual semantics with LDA's thematic insights, they used the HDBSCAN algorithm for clustering and a class-based TF-IDF for topic representation. Their results highlighted the model's edge over traditional methods in generating distinct and coherent topic words.

In a distinct study focusing on the Bangla news corpus, [49] adopted a hybrid clustering strategy that combines BERT and LDA. The primary objective of their experiments was to demonstrate the proficiency of clustering analogous topics from a specialised dataset comprising Bangla news articles. The results were telling: the BERT-LDA model excelled in deriving more cohesive topics, with its coherence value marginally surpassing that of standalone LDA.

In essence, these amalgamated models are designed to encapsulate both the contextual and thematic subtleties present in textual data, thereby furnishing a more enriched and holistic representation. The studies mentioned above shed light on the immense potential of such hybrid models across diverse domains, ranging from generic text classification to niche sectors like financial news and specific regional languages.

G. Dimensionality Reduction in NLP

High-dimensional data in NLP poses several challenges, including the curse of dimensionality, increased computational complexity, and difficulty in visualising and interpreting the data [62], [64]. These challenges can hinder the effectiveness of TM and clustering algorithms, as the inherent noise and sparsity in high-dimensional spaces can obscure meaningful patterns and relationships within the data. [63], [64]

Dimensionality reduction techniques like Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are valuable tools in addressing these issues: PCA reduces data dimensions by capturing the most significant variance, while t-SNE and UMAP focus on preserving local relationships and revealing data clusters in lower-dimensional spaces [65]. In TM and clustering, these techniques play a significant role by reducing the complexity of NLP data representations (e.g., WEs) while retaining relevant information [16] and [20]. This simplification aids in improving the efficiency and effectiveness of clustering algorithms, enabling them to discover more meaningful and coherent topics or clusters within text data [66].

H. Challenges and Limitations

The reviewed literature highlights TM's importance in NLP, discussing challenges like scalability, interpretability, and domain-specific issues. Current TM techniques, including advanced methods like BERT, grapple with challenges [3], [4], [18], [19], and [26] such as scalability, especially with large datasets, due to high computational demands. BERT, while powerful, have interpretability issues, making it challenging to understand the reasoning behind topic assignments. Domain-specific challenges are prevalent, as topic models may struggle to adapt to specialised or rapidly evolving terminology and contexts. Traditional n-gram language models, as highlighted in [4], for example, can fail to capture the nuances of modern languages effectively.

While NN models offer depth and improved performance, they require substantial computational resources and large amounts of labelled data, limiting their applicability in resource-constrained environments, as discussed in [18] and [19]. Moreover, their black-box nature can hinder the interpretability of TM results, making it challenging for users to trust and act upon the insights generated.

This literature also underscores the potential benefits of integrating diverse NLP techniques to enhance TM outcomes. The gap in the literature that our paper aims to address is the need for more effective and interpretable TM approaches.

Hence, an innovative approach is proposed that integrates dimensionality reduction techniques and advanced embeddings to tackle these gaps and elevate the quality and applicability of TM in NLP. By bridging these existing limitations, the paper aims to provide a more robust and interpretable framework for extracting meaningful topics from textual data, thereby contributing to the advancement of NLP research and applications.

I. Advantages of Semantic Clustering in Culinary Analysis

Semantic clustering, particularly within the culinary domain, offers a plethora of benefits. When recipes or any kind of culinary data are clustered based on SS, it denotes that items within each cluster share certain conceptual or thematic resemblances. Delving into these clusters offers valuable insights and practical applications:

1) Examining recipes within a cluster can reveal discernible patterns or themes. For instance, certain clusters might emphasise sweet undertones, while others may highlight savory or spicy nuances.
2) Individuals who enjoyed a recipe from one cluster are more likely to appreciate another recipe from the same cluster, paving the way for tailored recommendations.
3) Contrasting clusters unravels subtle differences and variations. This helps in answering questions related to the unique categorisation of certain recipes based on ingredients or cooking techniques.
4) For (culinary) platforms abundant with recipes, semantic clustering can assist in methodical content organisation, simplifying recipe discovery. For example, when users search with specific culinary adjectives, understanding cluster semantics can ensure more relevant recipe suggestions.
5) Culinary aficionados, whether chefs, food bloggers, or home cooks, can derive inspiration from these clusters to craft aligned recipes or intentionally create distinctive dishes.
6) Clustering can lead to the discovery of novel ingredient combinations, presenting opportunities for unique culinary creations.
7) Commercial kitchens and restaurants can optimise inventory based on ingredient patterns across popular clusters.
8) Clusters can be analysed to cater to specific diets or preferences, such as vegan, gluten-free, or low-carb diets.
9) Patterns within clusters might offer a window into regional or cultural culinary preferences, enabling platforms to offer region-specific recipe suggestions.
10) Monitoring cluster shifts over time can shed light on emerging culinary trends and evolving consumer preferences.

In essence, while the foundational principle of these clusters is rooted in SS, their real-world applicability spans from insights and recommendations to resource management and trend predictions, enriching the entire culinary ecosystem.

III. Methodology

A. Dataset

With the evolution of NN structures, as noted in [41], there has been a surge in pioneering work related to culinary content. For instance, the work in [42] focused on classifying a collection of around 100,000 food photographs into 101 unique classes. The introduction of the Recipe1M+ dataset in [43] and [44], has resulted in further of research opportunities. A standout project by [45] integrated this dataset with a vast library of 13 million food images, aiming to bridge the gap between recipes and their corresponding images. Studies like [43] and [44] leveraged the Recipe1M+ dataset to craft concise recipes, leaving out measurements, and assessed their outputs based on perplexity and alignment between text and images.

Cumulatively, these initiatives highlight the potential of deep NNs when combined with expansive culinary data sources. A fresh challenge emerged in [46] where the objective was to produce complete recipes, inclusive of measurements. In line with this, the authors introduced the RecipeNLG dataset, tailor-made for natural language generation (NLG) tasks, paving new pathways for culinary studies within the domain [43], [44], and [45]. RecipeNLG is a rich collection of recipes, gathered from multiple sources, with each offering a significant array of contributions (can be found here[3]).

For the sake of project manageability and given time constraints, we opted for a subset of this dataset. The primary objective was to ensure data quality was not sacrificed while maintaining a manageable volume. The final choice was the subset from www.foodnetwork.com, due to its representative size. The data from this site provides a solid base for the project, facilitating the discovery of culinary tendencies and insights without the challenges of processing the entire dataset. For this study, we utilised a set containing 49,443 entries.

B. Constructing BERT Embeddings for Recipes

To generate embeddings for the recipes, we utilised the BERT model. Specifically, the 'bert-base-uncased' variant was chosen for this purpose. BERT is a transformer-based model trained on vast amounts of text, and it's capable of understanding the context of words within sentences. This makes it particularly suitable for generating embeddings for our recipe dataset, where context is crucial for understanding the nuances of ingredients and preparation methods.

To compute the BERT embeddings, we proceeded follows:

1) Initialisation: A tokeniser and a BERT model were initialised using the 'bert-base-uncased' variant from the HuggingFace `transformers` library.
2) For each batch of recipes, the text was tokenised and encoded into input tensors. Padding was applied to ensure consistency in tensor dimensions across batches, and the sequences were truncated to a maximum length of 256 tokens.
3) The encoded input tensors were fed into the BERT model to obtain embeddings. The embeddings for each

[3]RecipeNLG Dataset Source

recipe were derived by averaging the embeddings of all tokens in the sequence, resulting in a single vector representation for each recipe.

4) Cosine similarity was computed between the embeddings of the recipes within each batch. This similarity matrix gives a measure of how semantically similar two recipes are, based on their BERT embeddings.
5) The dataset was divided into batches of 100 recipes each. For each batch, BERT embeddings were generated, and the similarity matrix was computed and saved. This batching approach was employed to manage memory consumption and computational efficiency.
6) The batch processing was equipped with error handling to capture and log any exceptions that might arise during the embedding generation or similarity computation. This ensures that the process can continue uninterrupted, even if issues arise with individual batches.

Utilising this procedure, we successfully generated BERT embeddings for the entire recipe dataset. These embeddings served as the foundation for the subsequent t-SNE visualisation and the following analysis, shedding light on the semantic landscape of the culinary world.

C. Analysis of Recipe Embeddings Visualisation

Visualisation Based on Reduced BERT Embeddings via t-SNE

To visualise the high-dimensional BERT embeddings (typically ranging from 768 dimensions upwards, depending on the variant of BERT used), we employed the t-SNE (t-distributed Stochastic Neighbour Embedding) technique [67], which is a machine learning algorithm for dimensionality reduction that is particularly well-suited for the visualisation of high-dimensional datasets [68].

It works by minimising the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding [67], [68]. The reduced space retains the relative distances and relationships between data points [67], [69], providing a visual representation of the semantic landscape of our recipe dataset.

The visualisation presented in Figure 1 allows us to observe clusters of recipes that share similar themes, ingredients, or preparation methods, as well as identify unique or outlier recipes.

Insights from the Visualisation

The visualisation derived from the reduced BERT embeddings of our recipe dataset offers intriguing insights into the semantic landscape of the culinary world. At a glance, distinct clusters and a pervasive spread of data points provide a rich tapestry of information.

1) The darker blue data points, associated with a value of 1, are distributed throughout the visualisation. This widespread distribution suggests a vast array of recipes with diverse characteristics. These recipes may not share strong commonalities in terms of ingredients, preparation methods, or other factors captured by the embeddings. Their ubiquitous presence may signify a versatile or generic set of recipes, adaptable to various tastes and requirements.
2) Three distinct clusters emerge from the visualisation: orange (value 6), darker orange (value 7), and light green (value 5). The absence of overlap between these clusters suggests recipes within each group share strong SS. The proximity of the lighter orange cluster to the light green one indicates potential overlaps in culinary themes or ingredients, albeit being distinct enough to form separate clusters.
3) While the darker blue data points represent a broad spectrum of recipes, the presence of defined clusters indicates pockets of specificity within the dataset. These clusters might represent popular or common types of recipes, variations of similar dishes, or specific culinary themes. The interspersion of the darker blue data points among these clusters hints at transitional recipes. These could be recipes that bridge the gap between distinct culinary categories or represent fusion cuisines.
4) The diverse yet structured nature of the visualisation underscores the richness of the recipe dataset. While there are clear groupings indicating popular or common recipe themes, the diversity captured by the darker blue points suggests a vast potential for culinary exploration and innovation. Future work might delve deeper into the darker blue group to discern potential sub-clusters or patterns. Refining the dimensionality reduction technique or clustering parameters might also provide more granular insights.

Incorporating this analysis into our research provided a foundational understanding of the dataset's structure and the potential avenues for further exploration. The interplay between the widespread darker blue data points and the distinct clusters paints a picture of both diversity and commonality in the world of recipes, setting the stage for more detailed investigations.

D. Process Overview

The primary objective of our research was to harness the capabilities of BERT embeddings in conjunction with clustering techniques to bolster TM. Our integrated Clustering and BERT framework revolves around the following steps:

1) *Data Pre-processing and Feature Extraction*: The RecipeNLG dataset underwent meticulous pre-processing to eliminate noise and standardise the text. Post-cleaning, each recipe was processed through the BERT model, leveraging its deep bidirectional transformers to generate dense embeddings. Owing to computational constraints, these embeddings were dimensionally reduced to a size of 100.
2) *TM Methods*: The embeddings were subjected to LDA to resonate with the inherent thematic structures within the recipes. Additionally, a combined BERT-LDA model was introduced, offering an enhanced TM approach.

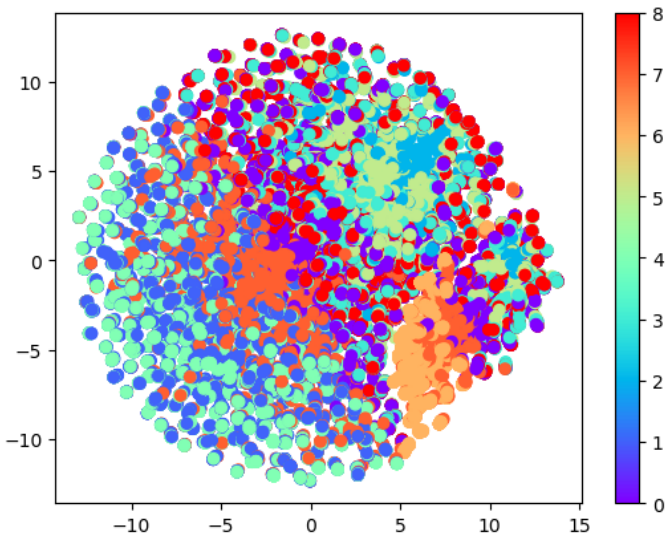

Fig. 1. t-SNE visualisation of reduced BERT embeddings for the recipe dataset.

3) *Dimensionality Reduction*: The high-dimensional nature of BERT embeddings necessitated the use of various dimensionality reduction techniques to ensure computational efficiency and the retention of pivotal information.
4) *Recipe Clustering*: The K-means clustering algorithm was employed post-dimensionality reduction, ensuring recipes with analogous themes clustered cohesively.
5) *Performance Evaluation*: The proposed integrated approach's robustness and efficacy were benchmarked against traditional TM techniques. Quantitative metrics, including the coherence score, silhouette score (SLS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI), illuminated the methodology's strengths and areas meriting refinement.

E. Data Processing Algorithm

A detailed breakdown of the TM process employed in this study is shown in Table I, where:

- R - Recipe
- RP - Pre-processed recipes
- RF - Vectorised recipe
- Tm - Topic model
- LDA – Topic model based on LDA algorithm
- $BERT$ – Topic model based on Bert algorithm
- $BERT\_LDA$ – Topic model based on combined BERT-LDA
- R_{DR_PCA}, R_{DR_t-SNE}, R_{DR_UMAP} – Topics obtained after Dimensionality reduction based on PCA, t-SNE and UMAP Algorithms
- R_{CT_PCA}, R_{CT_t-SNE}, R_{CT_UMAP} - Clustered topic models obtained after dimensionality reduction based on PCA, t-SNE and UMAP

We began by determining the optimal number of (a) topics for LDA TM and (b) components for PCA on the embeddings. The optimal number of topics for LDA was determined by coherence scores with a higher score indicates superior topics. The `compute_coherence_values` function calculated coherence scores for LDA models with an array of topics as shown in Figure 2, it commences from 2 topics and peaks at 40 topics, with an increment of 6 at each interval. Subsequently, the coherence scores are illustrated against the number of topics. The optimal number of topics, denoted as `optimal_topics`, is 20, as this number provided the highest coherence score.

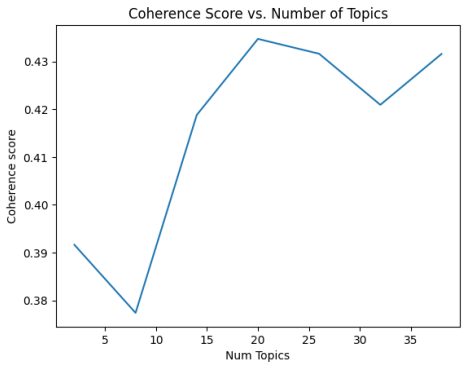

Fig. 2. Coherence Scores and Number of Topics

Choosing the right number of topics in LDA is pivotal. Having too few might result in the topics being exceedingly broad, while having too many might lead to overlap or excessive specificity. To deduce the optimal number, we assessed the coherence scores for various topic quantities. Coherence scores measure the quality of the topics generated by an LDA model, with heightened scores suggesting more meaningful topics.

The cumulative explained variance versus the number of components is illustrated in Figure 3. The subsequent step was to select the number of components where the cumulative

TABLE I
TM BASED ON INTEGRATED CLUSTERING AND COMBINED BERT-LDA

Step	Description
1	$\mathcal{R}$: the input recipes extracted from RecipeNLG dataset, where $i = 1, \ldots, n$ and n is the total number of recipes
2	Generate pre-processed recipes $\mathcal{R}_P$ from $\mathcal{R}$ by implementing stop word removal, data cleaning, stemming, lemmatisation, etc.
3	Extract the recipes' features $\mathcal{R}_F$ from $\mathcal{R}_P$ using word embedding (TF-IDF)
4	(i) Apply LDA as a TM algorithm on $\mathcal{R}_F$ to get $\mathcal{T}\Updownarrow_{LDA}$ (ii) Apply the transformer-based (Sentence Transformers) TM algorithm BERT on $\mathcal{R}_F$ to get $\mathcal{T}\Updownarrow_{BERT}$
5	Apply combined BERT-LDA based TM to get $\mathcal{T}\Updownarrow_{BERT\_LDA}$
6	Apply dimensionality reduction on topics $\mathcal{T}\Updownarrow_{BERT\_LDA}$ based on PCA, t-SNE and UMAP algorithms to generate $\mathcal{R}_{DR\_PCA}$, $\mathcal{R}_{DR\_t-SNE}$, $\mathcal{R}_{DR\_UMAP}$
7	Cluster the topics obtained from the last step after dimensionality reduction to get $\mathcal{R}_{CT\_PCA}$, $\mathcal{R}_{CT\_t-SNE}$, $\mathcal{R}_{CT\_UMAP}$
8	Compute Silhouette Score to evaluate the performance of the proposed model

explained variance exceeds 0.95. Based on the graph, the explained variance sees a noticeable increase around the 20-component mark, which suggests that this number captures a significant portion of the dataset's variance, i.e., about 80% of the data's variance.

Increasing components to 30 could capture more variance but at the cost of computational efficiency, especially given our large dataset. While more components enhance variance capture, they may complicate result interpretation, visualisation, and risk overfitting by capturing noise rather than genuine patterns. Thus, we opted for 20 components, balancing variance capture with efficiency and interpretability.

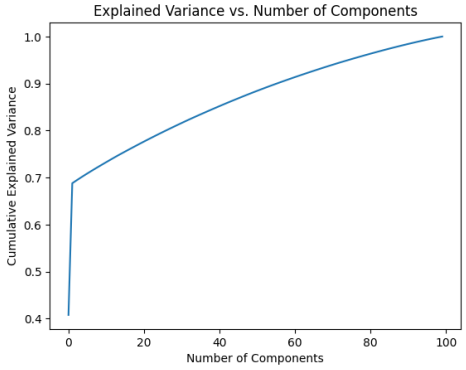

Fig. 3. Explained Variance and Number of Components

Given the challenges posed by high-dimensional data in clustering algorithms, we employed dimensionality reduction techniques, including PCA, t-SNE, and UMAP. After reducing the dimensions, we applied the k-means clustering algorithm to group the data and subsequently evaluated the results. To determine the optimal number of clusters (k), the Elbow Method [4]was used, with Figure 4 consistently suggesting 3 as the ideal choice across all methods. This uniformity reinforces our confidence in selecting three clusters.

[4]This is a standard technique for determining the best number of clusters in k-means clustering, and is represented through 'elbow plots' that show the rate of 'distortion' decline as clusters increase. An 'elbow' or significant change in this rate provides a good estimate for the actual number of clusters required.

IV. INTER-CLUSTER ANALYSIS 1

We next examine the 20 clusters/topics suggested by the coherence score, a metric used to judge the quality of the topics produced by a topic model.

Cluster 1: The words indicate *barbecue* and *grilling* recipes. Key ingredients suggest meats like 'pork', 'beef', and 'chicken', with an emphasis on 'ribs'. There is a marinade aspect indicated by 'vinegar', 'molasses', and 'brown sugar'. A smoke flavour is hinted at with 'wood chip' and 'smoke'.

Cluster 2: appears to represent *baking*, specifically cookies and pastries. Keywords like 'chocolate', 'cookie', 'bake', 'flour', and 'sugar' are indicative of dessert recipes.

Cluster 3: seems to have spaghetti or *Italian* pasta dishes. The presence of words such as 'spaghetti', 'pasta', 'marinara', 'tomato', 'garlic', and 'cheese' strongly supports this.

Cluster 4: appears to emphasise *Asian* cuisines, perhaps *Chinese or Japanese*. Terms like 'soy', 'sake', 'wasabi', 'ginger', and 'stir-fry' suggest dishes that are typical in East Asian cooking.

Cluster 5: points to classic *American or European* dishes. We can deduce this from words such as 'roast', 'potato', 'beef', 'carrot', and 'onion'. The mention of 'Yorkshire' also suggests *British* cuisine.

Cluster 6: indicates *sweet desserts and pastries*. 'Caramel', 'pecan', 'cream', and 'pie' are clear indicators. The presence of 'bourbon' suggests recipes that might incorporate alcoholic elements for flavour.

Cluster 7: focuses on *spicy or peppery dishes*. 'Chili', 'pepper', 'jalapeno', and 'hot' stand out, pointing towards spicy cuisines, perhaps *Mexican or Tex-Mex*.

Cluster 8: suggests *seafood* recipes, especially dishes involving shrimp. Words like 'shrimp', 'seafood', 'lobster', and 'shell' highlight this.

Cluster 9: revolves around *savory pie* recipes. Keywords such as 'pie', 'crust', 'mushroom', 'onion', and 'pastry' support this idea.

Cluster 10: hints at light *dishes or appetisers*. With words like 'spinach', 'salad', 'lemon', and 'olive', we can deduce recipes that are fresh and light.

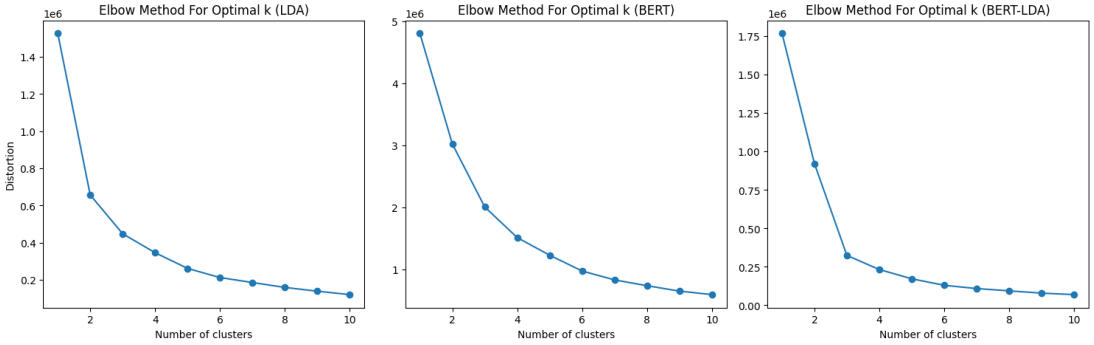

Fig. 4. Elbow method plot showcasing the cost function against different k values.

Cluster 11: the theme here is *soups or stews*. This can be inferred from words like 'soup', 'broth', 'bowl', 'stew', and 'chicken'.

Cluster 12: suggests sweet dishes or candies, potentially fudge or related confections. This is indicated by 'fudge', 'candy', 'sugar', 'chocolate', and 'butter'.

Cluster 13: the emphasis here is on recipes involving *fruits*, especially apples. Words like 'apple', 'cider', 'fruit', and 'spice' stand out.

Cluster 14: here we have *bread or baked goods*. 'Bread', 'flour', 'yeast', 'dough', and 'knead' are clear indicators.

Cluster 15: leans towards *Indian or spicy Asian* cuisines. Terms like 'curry', 'coconut', 'turmeric', and 'cumin' hint at this.

Cluster 16: seems diverse but might be pointing towards various *meat-based* dishes with marinades or sauces because of 'grilling', 'marinades', and various meat preparations.

Cluster 17: the focus is predominantly on *meat dishes*, perhaps involving beef or pork. The presence of words like 'breadcrumb', 'oven', 'pork', and 'beef' suggests this.

Cluster 18: the theme here appears to be *poultry* or chicken-based recipes. 'Chicken', 'butter', 'roast', and 'breast' confirm this.

Cluster 19: is quite diverse but seems to lean towards *gourmet or specialised dishes*, possibly with French influence. Words like 'truffle', 'foie', 'burrata', and 'pesto' are indicative of this.

Cluster 20: emphasis seems to be on drinks or beverages, especially *cocktails*. The presence of 'lime', 'rum', 'tequila', and 'margarita' supports this interpretation.

Therefore, the majority of clusters focus on specific cuisines or types of meals, ranging from grilled meats to baked goods and cocktails. Some clusters lean towards specific cultural cuisines, like Cluster 4 (Asian) or Cluster 15 (Indian). There is a good balance between savory dishes (e.g., Cluster 3's pasta dishes or Cluster 8's seafood) and sweet ones (Clusters 2 and 6). There is also an inclination towards specific meal types, with clusters focusing on main courses, while others look at desserts or drinks. Gourmet or upscale ingredients feature in some clusters, indicating a mix of everyday meals and more sophisticated recipes.

In summary, these clusters provide a comprehensive insight into a diverse range of recipes, ingredients, and culinary traditions. They span from everyday dishes to gourmet preparations and cover a wide spectrum of flavours, ingredients, and cooking methods.

V. Clustering Results

The clustering analysis that was conducted next focused on embeddings from different feature extraction methods, namely BERT, LDA, and a combination of BERT and LDA. BERT with PCA approach leveraged the power of BERT embeddings to understand the contextual significance of words within the recipes. Post which, PCA was employed to reduce dimensionality, making the data more suitable for clustering. BERT-LDA with UMAP is a combined modern NN approach (BERT) with a traditional TM technique (LDA). UMAP was then used for dimension reduction and potentially visualisation. LDA with UMAP method employed the traditional LDA approach for TM, followed by UMAP for dimension reduction.

These embeddings represented the semantic nuances of the recipes present in the dataset. Of particular note was the 'averaged_embedding' that was computed. This embedding was specifically tailored to the dataset, ensuring that it encapsulated the inherent characteristics and patterns within the recipes. By averaging out the embeddings, a holistic representation was achieved that effectively captured the essence of each recipe in the dataset.

Various dimensionality reduction techniques, such as PCA, t-SNE, and UMAP, were employed to reduce the complexity of these embeddings. Dimensionality reduction not only aided in visualisation but also often helped improve the efficiency of clustering algorithms. Each method had its unique strengths. For example, PCA is linear and worked well when the data exhibited a linear structure, while t-SNE and UMAP were capable of capturing nonlinear structures.

The analysis also utilised clustering evaluation metrics, including SLS, DBI, and CHI. Each of these metrics provided insights into the quality of the clusters that were formed. A higher SLS indicated better-defined clusters, while a lower DBI suggested that clusters were well-separated. Similarly, a higher CHI indicated the formation of dense, well-separated clusters.

TABLE II
SUMMARY OF CLUSTERING RESULTS

Embedding Type	Dimensionality Reduction	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
BERT	PCA	0.496667	0.692409	77945.750780
BERT	t-SNE	0.369102	0.874603	36679.435634
BERT	UMAP	0.365219	0.884391	36508.726023
LDA	PCA	0.195987	1.695571	6595.182827
LDA	t-SNE	0.383853	0.865595	39909.915703
LDA	UMAP	0.441901	0.838473	60059.934332
BERT-LDA	PCA	0.167555	1.914880	11824.500034
BERT-LDA	t-SNE	0.400824	0.862537	43980.888679
BERT-LDA	UMAP	0.579857	0.534542	120266.518231

As it can be seen in Table II, there were varying performances observed. These results are visualised in Figure 5. Based on these results, several observations can be made:

LDA with UMAP Clustering shows

- a clear distinction between three major clusters, which seem to cover distinct, non-overlapping regions in the UMAP space;
- clusters have well-defined boundaries, suggesting that the LDA model could segregate the data efficiently;
- the separation and density of the clusters might indicate distinct topics or categories within the data.

BERT with PCA Clustering confirms

- there are primarily two clusters: one dense elongated cluster and a few scattered points;
- the elongated structure of the main cluster might indicate the presence of a gradient or spectrum in the data representation, which could be a sequence of closely related topics or a transition from one topic to another;
- the few scattered points outside the main cluster might represent outliers or topics that BERT found difficult to group with the dominant patterns;
- PCA's linear dimension reduction might not capture the complexities in high-dimensional BERT embeddings as effectively as non-linear methods like UMAP.

BERT-LDA with UMAP Clustering suggests

- two distinct clusters, but their boundaries are not as well-defined as in the LDA with UMAP visualisation;
- the overlap and proximity between clusters might indicate that BERT-LDA provides a more nuanced separation, capturing some shared information between clusters;
- the presence of only two major clusters suggests that the combination of BERT and LDA might be generalising or grouping the data into broader categories compared to just LDA.

Based on these observations,

- UMAP seems to provide clearer cluster separation than PCA when visualising the embeddings, which is consistent with UMAP being a non-linear dimensionality reduction technique.
- LDA, being a TM algorithm, can generate distinct clusters, suggesting clear topics within the data.
- BERT, with its deep representation, combined with PCA, seems to represent data on a spectrum, which might be useful for understanding transitions between topics or gradients within a topic.
- Combining BERT and LDA introduces some overlap in the clusters, potentially capturing more nuanced relationships between topics.

Results for BERT embeddings with various dimensionality reduction methods

Figure 6 compares different methods and metrics. Results in Table II show the following. PCA has the highest SLS (0.496564) among the three methods. Higher values indicate better-defined clusters. This suggests that, based on this metric, PCA provides the most distinct clusters for the data. Lower values of DBI indicate better partitioning. Again, PCA performs the best with the lowest score (0.692596), suggesting better partitioning of the clusters compared to t-SNE and UMAP. Higher values of CHI indicate better-defined clusters. PCA also significantly outperforms the other two methods with a score of 78001.759817. This further reinforces the idea that PCA provides better clustering for the BERT embeddings in this dataset.

Therefore, based on these findings, PCA seems to be the most suitable dimensionality reduction method for clustering BERT embeddings in this particular dataset, as it consistently outperforms both t-SNE and UMAP across all three evaluation metrics. It provides better-defined and better-partitioned clusters. However, these are just quantitative metrics. Visualising the clusters, analysing the content within each cluster can further assist in making a conclusive decision.

Results for LDA embeddings with various dimensionality reduction methods

UMAP method shows the highest SLS (0.441901) among the three methods (see Table II), suggesting that UMAP provides the most distinct clusters for the LDA embeddings in this dataset. UMAP has the DBI lowest score (0.838473), indicating that it provides better partitioning of the clusters compared to PCA and t-SNE for the LDA embeddings. Higher values of CHI indicate better-defined clusters. UMAP significantly outperforms PCA with a score of 60059.934332. t-SNE is in between with a score of 39909.915703. This suggests that UMAP's clusters are better defined compared to PCA and t-SNE for this particular dataset.

Therefore, UMAP appears to be the most suitable dimensionality reduction method. It outperforms both PCA and t-SNE in all three evaluation metrics. UMAP provides better-defined and better-partitioned clusters for the LDA embeddings in this dataset. However, as with the BERT embeddings, the actual content of the clusters should be considered in addition to these quantitative metrics when making a final decision.

Results for BERT-LDA embeddings with various dimensionality reduction methods

UMAP stands out with a silhouette score of 0.579857, which is considerably higher than the scores of PCA and t-SNE (see Table II). This suggests that UMAP provides the most distinct clusters for the BERT-LDA embeddings in this dataset. Lower values of DBI indicate better partitioning. UMAP also excels here with the lowest DBI of 0.534542, which indicates superior partitioning compared to PCA and t-SNE for the BERT-LDA embeddings. Higher values of CHI indicate better-defined clusters. UMAP dramatically outperforms both PCA and t-SNE with a score of 120266.518231. This reinforces the observation that UMAP's clusters are particularly well-defined for the BERT-LDA embeddings.

Hence, UMAP appears to be the optimal choice for dimensionality reduction. It excels in all three evaluation metrics,

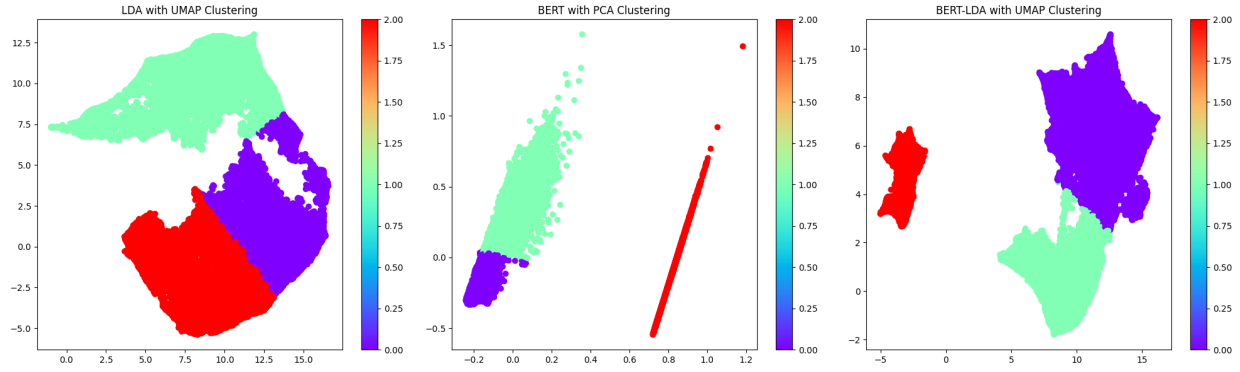

Fig. 5. t-SNE visualisation of reduced BERT embeddings for the recipe dataset.

suggesting superior cluster definition and partitioning. The results indicate that the combination of BERT-LDA embeddings with UMAP as the dimensionality reduction method might yield the most meaningful and well-segregated clusters. Again, alongside these quantitative metrics, visual inspection and domain expertise should be leveraged to validate the effectiveness and interpretability of the clusters.

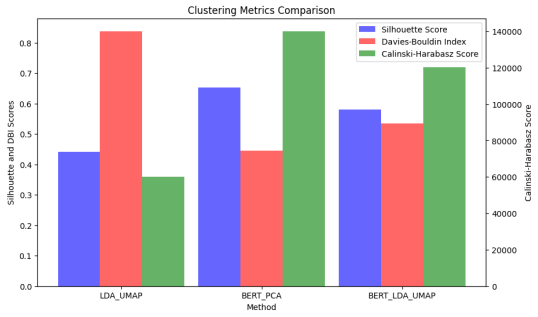

Fig. 6. Comparison of Clustering Metrics for the Methods.

A. Explanation of Clustering Results

Nature of Embeddings

BERT produces dense vector representations for text that capture deep contextual information: the embeddings are positioned in a high-dimensional space and are sensitive to the input's context. Hence, their distribution in this space may be challenging to reduce accurately using some techniques. LDA produces topic-based embeddings. These are more sparse and represent the distribution of topics in a document. Their nature is fundamentally different from BERT embeddings, which can affect how they respond to dimensionality reduction techniques. On the other hand, BERT-LDA Embeddings is a fusion, trying to combine the depth of BERT embeddings with the topic distributions from LDA. The hybrid nature can make them unique in how they behave with dimensionality reduction.

Dimensionality Reduction Techniques

As a linear technique, reducing dimensions by projecting data onto the principal components, PCA excels when the data has linear patterns and relationships. The success of PCA with BERT embeddings suggests that BERT embeddings in the dataset might exhibit such linear patterns. UMAP, on the other hand, is a non-linear technique designed to respect the local and global structure of data. Given its consistent performance across embeddings, it suggests that both LDA and BERT-LDA embeddings have complex structures, which UMAP can handle effectively. T-SNE method focuses on preserving local structures and can sometimes lose sight of global patterns. Its varying performance indicates that it might not always be the best choice for capturing the essence of complex embeddings.

Therefore, while PCA outperformed other methods for BERT embeddings, indicating potential linear patterns, UMAP's consistent top performance for LDA and BERT-LDA embeddings suggests its capability to handle more complex, possibly non-linear data structures inherent in these embeddings. The metrics further validate these observations. Saying that, it is crucial to recognise that while these quantitative metrics provide valuable insights, qualitative assessments (e.g., analysing cluster content) are equally critical to truly understand the clustering outcomes.

Inter-Cluster Analysis 2

To distill significant insights from our clusters, we investigated the composition and characteristics of the encapsulated recipes. Priority was given to embeddings with high SLS, suggesting well-defined clusters. This study aims to discern if the clusters correlate with culinary themes or principles. Table III provides a distribution overview of recipes across clusters for various embedding and dimensionality reduction combinations.

TABLE III
Cluster Distributions.

Method	Cluster 0	Cluster 1	Cluster 2
BERT with PCA	22,414	21,185	5,840
BERT-LDA with UMAP	30,436	13,163	5,840
LDA with UMAP	16,314	14,556	18,569

Key Observations from Clustering

- Both "BERT with PCA" and "BERT-LDA with UMAP" display a notably smaller Cluster 2, hinting at a specialised recipe subset.
- "LDA with UMAP" offers a balanced distribution, alluding to broader categorisations.
- "BERT with PCA" showcases prevalent overlapping ingredients, suggesting thematic overlaps.
- "BERT-LDA with UMAP" highlights distinct thematic differences, likely due to melding BERT's contextual strengths with LDA's topical focus.
- "LDA with UMAP", while capturing overarching themes, might miss the depth present in BERT-LDA.
- Ubiquitous ingredients like 'salt' and 'water' are frequently spotted across methods, indicating potential preprocessing enhancements.

Under the "BERT with PCA" clustering method:

Cluster 0: seems primarily centered on desserts or baked goods, with ingredients pointing towards oatmeal cookies, banana cream-based drinks, and pastries. There is a consistent presence of ingredients like 'butter', 'sugar', 'flour', 'oats', and 'baking powder'.

Cluster 1: appears to contain recipes with a mixture of desserts and savoury items, characterised by ingredients such as 'vanilla extract', 'almond', 'wine', 'cookies', 'marsala', and 'egg yolks' for desserts, and ingredients like 'pepper', 'chickpea', 'thyme', 'potato', and 'ginger' for savory dishes. There seems to be a focus on traditional European (especially Italian) and Jamaican cuisines.

Cluster 2: encompasses a more diverse range of savoury dishes. Ingredients such as 'fennel', 'saffron', 'anchovy', 'pine nuts', and 'currants' suggest Mediterranean and Italian influences. There is also a hint of seafood dishes, given the presence of crab, chervil, and parsley. Furthermore, the inclusion of ingredients like Irish cream and chocolate indicates that this cluster might also contain a few drink or dessert recipes.

In essence, while there is a predominant theme in each cluster, there is also some overlap, and a variety of recipes can be found within each group.

Under "LDA with UMAP" clustering method:

Cluster 0: This cluster highlights terms suggestive of intricate, savory recipes that incorporate a mix of meats and veggies. Frequent terms include 'pound', 'pepper', 'roast', 'salt', 'oil', 'beef', 'wine', and 'garlic'. The dishes from this cluster could be main course items such as roasts, stews, or braised meals. The myriad of meats, veggies, and techniques like 'roast' implies dishes suited for substantial dinners or midday meals.

Cluster 1: Centered on ingredients and procedures typical of sweet recipes or desserts, this cluster has predominant terms like 'sugar', 'cream', 'bake', 'chocolate', 'cocoa', 'rum', and 'flour'. It's probable that the recipes here are related to desserts, baked items, or sugary beverages. The mention of 'candi', 'mint', 'chocolate', and 'whip' suggests a spectrum of dessert varieties, from baked treats to sugary drinks.

Cluster 2: Dominated by ingredients and techniques tied with robust, savory, or brunch dishes, this cluster displays terms such as 'egg', 'butter', 'cheese', 'bacon', 'broccoli', 'olive', and 'salt'. They likely correspond to breakfast or brunch recipes, potentially inclusive of omelets, hashes, and cheesy concoctions. The strong presence of dairy, veggies, and eggs reinforces this notion.

To summarise, Cluster 2 seems to amass breakfast/brunch recipes. Cluster 1 has a dessert-centric vibe, and Cluster 0 likely gathers main course dishes. It's pivotal to understand that while this evaluation offers a general insight into each cluster's theme, there might be variations within individual recipes. A deeper dive into the complete recipes or a broader sample could provide further clarity.

Under "BERT-LDA with UMAP" clustering method:

Cluster 0: leans heavily towards savoury recipes encompassing diverse ingredients like chicken, bacon, onion, cider, and vinegar. The occurrence of terms such as 'simmer', 'saute', 'braise', and 'stir' indicates intricate cooking techniques. This cluster might encompass traditional or filling dishes, evident from ingredients like cabbage, apple, wine, and broth.

Cluster 1: exhibit a pronounced inclination towards baking and sweet treats. Terms like 'cookie', 'chocolate', 'buttermilk', 'powder', 'cupcake', and 'ganache' flag a baking theme. There is also a spotlight on specific baking methods and aspects like 'paddle', 'whisk', 'fold', and 'batter'. Ingredients such as strawberries, mascarpone, and Nutella suggest a gamut of sweet treats, ranging from cookies to cakes and pastries.

Cluster 2: is a melange of recipes. Terms like 'bean', 'lard', 'refry', and 'casserole' hint at comfort dishes or potential regional delicacies. Conversely, ingredients like shrimp, mayo, skewer, and roast signify dishes that might encompass seafood or grilled elements. A subtle healthy undertone is present with terms like 'spinach', 'orzo', 'pine', and 'salad'.

Overall, this method has effectively clustered the recipes into meaningful categories. Cluster 0 seems to be more about savoury, hearty dishes; Cluster 1 is clearly about desserts and baking; and Cluster 2 has a mix, with both comfort foods and lighter dishes. This clustering can be useful for someone looking to categorise a large set of recipes or to make recommendations based on a particular theme or preference.

Recommendations and Future Directions

- LDA with UMAP seems optimal for broad culinary themes.
- BERT-LDA with UMAP excels for detailed insights.
- BERT with PCA could shine when context matters but may require tweaks for our dataset.
- Subsequent efforts might explore results from diverse methods or delve into advanced embeddings like RoBERTa or XLNet.

In summary, both inter-cluster analyses consistently identified a dessert or baked goods-focused cluster. While the previous analysis had a generalised Cluster 1, post-dimensionality reduction refined clustering, segmenting savoury recipes into

specific categories. Dimensionality reduction with UMAP offers more precise clustering, delineating savory recipes into unique categories, while pre-reduction clusters gave a general dish type overview. The chosen method should match the desired granularity and insight specificity.

Both inter-cluster analyses identified a cluster focused on dessert or baked goods (Cluster 2 in the previous analysis and Cluster 1 in the current analysis). The previous Cluster 1 was more generalized in its grouping, suggesting foundational ingredients for many savory dishes. The post-dimensionality reduction clusters (Cluster 0 and Cluster 2) provided a more nuanced separation of savory dishes into main courses and breakfast/brunch recipes. Cluster 3 from the previous analysis (rich, creamy recipes) doesn't have a direct one-to-one match in the current analysis but seems to be distributed between Cluster 0 and Cluster 2. Overall, dimensionality reduction using UMAP seems to provide more refined clustering, separating savory recipes into distinct categories (main courses vs. breakfast/brunch). However, the clusters before the reduction provided a broader overview of the types of dishes (foundational savory, baked goods, creamy dishes). Each method has its advantages and can be chosen based on the granularity and specificity of the insights required.

B. Word Cloud and Frequency Analysis

As the final stage in this study, we turn to a word cloud (WC), which is a visual representation of text data where the size of each word indicates its frequency or importance[5]. It can be especially useful to quickly understand the most prominent terms in large volumes of text. In the context of our clustering analysis for recipes, using WCs can offer the several additional benefits. For example, by creating a WC for each cluster, we can immediately identify the most dominant terms or ingredients associated with each cluster. Visually comparing WCs from different clusters can help discern the uniqueness or similarity of the clusters. For instance, if 'salt' and 'pepper' dominate all clusters, it may indicate their ubiquitous use, while unique ingredients can show cluster specialties.

The word cloud visualisation represents three clusters possibly related to recipes or cooking instructions. Different models, BERT, LDA, and a combination of BERT-LDA, have generated interpretations of these clusters. This report discusses, compares, and concludes the analyses.

As depicted in Figure 7, the dominant words for Cluster 1 include "tablespoon", "salt", "onion", "add", and "garlic", indicating measurements and basic ingredients. In Cluster 2, the balanced emphasis is on "onion", "garlic", "salt", "tablespoon", and "water", suggesting balanced ingredients and measurements. In Cluster 3, words such as "onion", "salt", "tablespoon", "garlic", and "heat" stand out, implying a pattern similar to Cluster 1.

Figure **??** shows that Cluster 1 has words related to savory dishes like "mince", "add", "ground", "onion", "garlic", "salt", and "juice". Cluster 2 focuses on baking with terms such as "sugar", "mix", "egg", "tablespoon", "butter", "vanilla", and "chocolate". Cluster 3 overlaps with Cluster 1 and indicates general cooking terms.

In Figure 9, Cluster 0 represents basic cooking with words like "salt", "pepper", "cook", "oil", "onion", "water", "garlic", and "tablespoon". Cluster 1 pertains to baking with dominant words like "sugar", "egg", "butter", "bake", "tablespoon", "chocolate", and "vanilla". Cluster 2 seems to be associated with fresh ingredients, possibly salsas or tangy dishes, containing "salt", "onion", "juice", "lime", and "cilantro".

Upon comparing the analyses, several observations emerge:

- All three models identified the presence of key ingredients like "onion", "garlic", and "salt" across the clusters.
- Both BERT-LDA and LDA suggest a specific cluster related to baking or desserts, identifying words like "sugar", "butter", and "chocolate".
- BERT's analysis appears to be more generic, focusing primarily on the dominant words without diving deep into the theme of each cluster.
- LDA provides a more thematic interpretation, while BERT-LDA offers a mix of both thematic and generic interpretations.

Each model brings a unique perspective to the analysis of the WC clusters. While BERT gives a broader view, LDA offers a thematic breakdown, and BERT-LDA balances between the two. The choice of model might depend on the desired depth and specificity of the analysis required.

Concluding remarks

The identification of a dessert or baked goods cluster in both the clustering and WC analysis aligns, especially with the presence of words like "sugar", "butter", and "chocolate". The presence of key ingredients like "onion", "garlic", and "salt" in most clusters is consistent across both methods.

However, there are also some potential misalignments or additional insights. For example, the WC analysis highlighted certain specifics that the clustering did not, such as the possibility of a cluster related to fresh ingredients, salsas, or tangy dishes in the BERT-LDA WC analysis. While the clustering found a distinction between main courses and breakfast/brunch, this specific distinction is not as clear in the WC analysis. BERT's generic analysis aligns with the generalised grouping from the clustering before dimensionality reduction, while LDA's more thematic interpretation is in line with the post-reduction nuanced clusters.

Therefore, our findings are largely in line, especially when considering the broader themes such as the presence of baking-related clusters and the importance of foundational ingredients across clusters. However, some insights gained from one method are not directly mirrored in the other, which is expected given the different natures of the techniques. This

[5]In NLP, it can be a very useful tool as can quickly, for example, point out potential noise or irrelevant words that might have been included in the clustering. This can guide us in improving the data pre-processing steps. WCs are visually engaging and can be a great way to present the findings, especially to audiences that may not be familiar with technical details. While numerical or tabular results can provide detailed insights, visual aids like WCs can further support and strengthen those findings.

Fig. 7. t-SNE visualisation of reduced BERT embeddings for the recipe dataset.

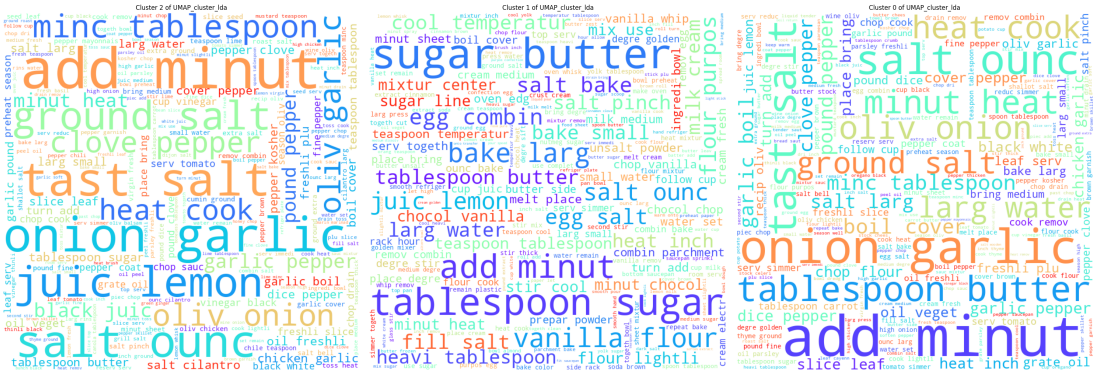

Fig. 8. t-SNE visualisation of reduced BERT embeddings for the recipe dataset.

Fig. 9. t-SNE visualisation of reduced BERT embeddings for the recipe dataset.

highlights the value of using multiple methods of analysis, as they can provide both reinforcement for key findings and offer additional or nuanced insights.

VI. Reflection on the Project

The journey of this project began with a seemingly simple aspiration: to delve into the realm of SS. This motivation, rooted in the desire to understand the intricate relationships within recipes, quickly unraveled into a labyrinth of challenges, discoveries, and learning.

The Complexity: Exploring SS, it became evident that the concept was far more multifaceted than initially perceived:

- The task was not merely about identifying similarities but understanding the depth and nuances of these relationships.
- Semantic relationships are often subjective. What appears semantically similar to one might differ for another, especially in a domain as diverse and culturally profound as cooking.
- Recipes are more than ingredients and directions lists. They encapsulate cultural nuances, diverse cooking methods, and unique personal touches, adding layers of complexity to discerning SS.

The Evolution and Challenges: From inception to realisation, the project's trajectory was marked by both hurdles and growth:

- The incorporation of advanced models like BERT emphasised the strength of deep learning and its complexities.
- A balance had to be struck between the depth of semantic insights and the volume of data.
- Every challenge, whether computational, methodological, or conceptual, morphed into a learning opportunity, refining the project's course and deepening our grasp of the domain.

One of the most formidable challenges was harnessing the power of BERT for embeddings. While BERT's potential in capturing semantic nuances is unparalleled, the computational demands were but substantial. Generating embeddings for a dataset as extensive as ours was both time-consuming and resource-intensive. Such challenges accentuated the importance of optimisation and the judicious selection of tools and methodologies. Indeed, this has been but a journey.

The project's evolution underscores the dynamism inherent in research. Simple ideas can lead to uncharted territories, pushing boundaries, challenging preconceived notions, and fostering growth and discovery. This venture, with its amalgamation of challenges and insights, exemplifies the profound depths and intricacies of semantic research in culinary arts.

VII. Conclusion, Limitations and Future Directions

In our unique approach, we crafted three distinct topic clustering models using BERT, LDA, and an innovative BERT-LDA combination, all tailored for a recipe dataset. For each model, we delved into three different dimensionality reduction techniques: PCA, t-SNE, and UMAP. These methods were meticulously compared, showcasing the novelty and effectiveness of our methodology in clustering the embeddings. These embeddings, rich in semantic content, underscore the importance of advanced NLP techniques in capturing intricate semantic relationships. Due to computational constraints, the employed dimensionality reduction techniques ensured the data remained manageable while retaining its core features.

LDA was subsequently applied, unveiling potential thematic structures within the recipes. This was complemented by the K-means clustering algorithm, which grouped recipes with similar themes, setting the stage for enhanced content organisation and recommendation systems. Quantitative evaluations, using metrics like coherence, silhouette scores, Davies-Bouldin Index, and Calinski-Harabasz Index, were compared against traditional TM method, revealing the efficacy of our approach.

Our study is distinguished by its innovative combination of BERT and LDA embeddings, aiming to capture both the syntactic and semantic intricacies of recipes. This dual approach ensures a holistic representation of the recipes, making our research a significant stride in the field of SS, which value, especially in practical applications like recipe recommendations, cannot be overstated. By understanding the nuanced relationships between different recipes, we can offer users more tailored and relevant suggestions.

Our study has also emphasised the importance of diverse methods of data exploration. The WC analyses provided thematic interpretations that complemented our clustering results. Specifically, while both methods confirmed the prevalence of foundational ingredients across clusters and the distinction between baking-related and other recipes, the WCs also hinted at nuanced interpretations, such as a potential focus on fresh ingredients or tangy dishes.

Moreover, our methodology's strength in identifying broader themes, as seen with BERT, and more specific, thematic clusters, as seen post-dimensionality reduction, reinforces the need for multiple methods of analysis. This multiplicity allows researchers and practitioners to choose a granularity level of insights based on the specific requirements of their applications.

Indeed, our research is not without limitations. Despite these advancements, there is still room for refinement. The overlap observed in some clusters, as pointed out by both the clustering and WC analyses, suggests that while we can capture broader themes effectively, there may still be sub-themes or nuances that our current approach may not fully distinguish. This is an exciting avenue for future research, where techniques can be developed to capture these finer nuances without compromising on the broader thematic insights.

While our quantitative evaluations have shown promising results, it is crucial to note that these findings are preliminary. While such metrics were invaluable, the absence of a comprehensive qualitative analysis might have left certain nuances unexplored. Additionally, the presence of overlapping clusters in some methods suggests potential shared themes or broader categorisations, which might not align with the goal of distinct clustering. Therefore, there is a significant scope for deeper exploration and validation.

Looking ahead, there is a plethora of opportunities for further research. Further investigation is warranted to fully understand the implications of our approach and to refine its potential applications. Indeed, more enhanced visualisation techniques can provide deeper insights into cluster dynamics. Exploring alternative dimensionality reduction and/or clustering methods might unveil novel insights. And, most importantly, understanding the intricacies of cluster distributions can significantly refine recipe categorisation and recommendation systems, making them more user-centric and efficient.

Our approach has the potential to redefine user experience on digital culinary platforms, offering recommendations that are not just based on superficial tags but delve deep into the semantic richness of each recipe. In the era of smart kitchens, integrating our recommendation system with IoT devices can lead to real-time recipe suggestions based on available ingredients. While our focus was content-based clustering, integrating collaborative filtering can factor in user preferences, leading to an even more personalised recommendation system.

Future iterations of the system can benefit immensely from a user feedback loop, refining recommendations based on

user interactions and preferences. While BERT has shown significant promise, transfer learning techniques leveraging other models could further enhance the quality of embeddings. As digital culinary platforms grow, the real challenge lies in real-time analysis and clustering of an ever-expanding recipe database. Our approach sets the foundational stone, but scalability remains an avenue to explore.

VIII. Link to Python code

The Python program used to carry out this study can be found here.

References

[1] Wang, Y., Huang, G., Li, J., Li, H., Zhou, Y., & Jiang, H. (2021). Refined global word embeddings based on sentiment concept for sentiment analysis. IEEE Access, 9, 37075-37085.

[2] Souza, F. D., & Filho, J. B. D. O. E. S. (2023). Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers. Neural Computing and Applications, 35(13), 9393-9406.

[3] Van Erp, M., & Groth, P. (2020). Towards entity spaces. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 2129-2137).

[4] Yuwen, L., Chen, S., & Yuan, X. (2021). G2Basy: A framework to improve the RNN language model and ease overfitting problem. Plos one, 16(4), e0249820.

[5] Wu, X., & Palmer, M. (2018). A survey on semantic similarity measures. Foundations and Trends® in Information Retrieval, 11(1), 1-135.

[6] Miraj, R., & Aono, M. (2021). Humour detection using a Bidirectional Encoder Representations from Transformers (BERT) based Neural Ensemble Model. In 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA) (pp. 1-6). IEEE.

[7] Sivakumar, S., & Rajalakshmi, R. (2022). Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. Social Network Analysis and Mining, 12(1), 104.

[8] Navigli, R., Punyakanok, V., & Lapata, M. (2009). Semantic similarity in a heterogeneous world. Computational Linguistics, 35(4), 615-639.

[9] Wu, X., & Palmer, M. (2018). A survey on semantic similarity measures. Foundations and Trends in Information Retrieval, 11(1), 1-135.

[10] Zhang, Y., & Lee, K. (2016). Semantic similarity based on word embedding: A survey. ACM Computing Surveys (CSUR), 49(2), 1-40.

[11] Bhatia, A., Kumar, M., & Mahata, D. (2020). A Comprehensive survey of similarity and relatedness metrics in NLP. In Proceedings of the Third Workshop on eLexicography, eLex2020 (pp. 27-37).

[12] Chen, X. and Kondrak, G. (2013). Joint learning of semantic and latent clusters for word sense disambiguation. Computational Linguistics, 39(4), 935-976.

[13] Nguyen, D. Q. and Nguyen, D. Q. (2017). Combining word and sense Eembeddings for word sense disambiguation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 333-338).

[14] Landauer, T. K., Foltz, P. W. and Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25(2-3), 259-284.

[15] Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), pp. 2786-2792.

[16] Cer, D., & Diab, M. (2013). Cross-lingual textual entailment for content synchronisation in the news domain. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pp. 1216-1223.

[17] Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 670-680.

[18] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity—Multilingual and Cross-lingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 1-14.

[19] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3973-3983.

[20] Hu, S., Ma, L., Liu, H. and Li, S. (2021). A Systematic Evaluation of Sentence Similarity Measures. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 6436-6446).

[21] Zhang, L., Liu, Z., & Liu, B. (2019). A survey of neural network architectures and learning approaches for sentence similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 13(5), 1-34.

[22] Caruana, R. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730.

[23] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144.

[24] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, Visualising and Interpreting Deep Learning Models. ITU Journal: ICT Discoveries, 1(1), 58-63.

[25] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[26] Lipton, Z. C. (2016). The Mythos of Model Interpretability. In Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning, pp. 112-121.

[27] Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning (ICML), Vol. 32, pp. 1188-1196.

[28] Devlin, J., Chang, M. W., Lee, K.., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Vol. 1, pp. 4171-4186.

[29] Zhang, L., Liu, Y., & Liu, B. (2019). A Multi-Metric Learning Approach for Document Clustering. IEEE Transactions on Knowledge and Data Engineering (TKDE), 31(10), 2291-2303.

[30] Chen, J., Liu, Y., & Liu, B. (2020). Fusing Cosine Similarity and Neural Network-Based Embeddings for Sentiment Classification. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(5), 1-26.

[31] Lee, S., & Kim, D. (2020). Sentence Similarity Modeling Using Recurrent Neural Networks for Sentence Pair Classification. Information, 11(10), 476.

[32] Hu, S., Ma, L., Liu, H., & Li, S. (2021). "A Systematic Evaluation of Sentence Similarity Measures." In Proceedings of the 29th International Conference on Computational Linguistics (pp. 6436-6446).

[33] Yih, W., He, X., & Meek, C. (2013). Semantic Parsing for Single-Relation Question Answering. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 643-653).

[34] Reisinger, J., & Mooney, R. J. (2010). Multi-prototype Vector-Space Models of Word Meaning. Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2010) (pp. 109-117).

[35] Chen, X., & Kondrak, G. (2013). Joint Learning of Semantic and Latent Clusters for Word Sense Disambiguation. Computational Linguistics, 39(4), 935-976.

[36] Nguyen, D. Q., & Nguyen, D. Q. (2017). Combining Word and Sense Embeddings for Word Sense Disambiguation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 333-338).

[37] Wang, J., Liu, H., Zhang, T., & Liu, Z. (2017). Investigating the effectiveness of combined similarity metrics for word sense disambiguation. Journal of Artificial Intelligence Research, 57, 407-436.

[38] Liu, Y., Liu, B., Wang, J., & Zhang, T. (2020). Metric fusion for cross-lingual semantic similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(5), 1-26.

[39] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[40] Goyal, A., & Kashyap, i. (2022). Latent Dirichlet Allocation - An approach for topic discovery, 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, pp. 97-102.

[41] Liang, M., & Hu, X. (2015). Recurrent Convolutional Neural Network for Object Recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[42] Bossard, L., Guillaumin, M., & Gool, L. V. (2014). Food-101–Mining Discriminative cComponents with Random Forests. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, Proceedings, Part VI 13. Springer International Publishing.

[43] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., & Torralba, A. (2017). Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11081-11090.

[44] Salvador, A., Drozdzal, M., Giró-i-Nieto, X., & Romero, A. (2019). Inverse Cooking: Recipe Generation from Food Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10453-10462.

[45] Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., & Torralba, A. (2021). Recipe1m+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food images. IEEE transactions on Pattern Analysis and Machine Intelligence, 43(1), pp. 187-203.

[46] Bień, M., Gilski, M., Maciejewska, M., Taisner, W., Wisniewski, D., & Lawrynowicz, A. (2020). RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation. In Proceedings of the 13th International Conference on Natural Language Generation, pp. 22-28.

[47] Zhang, P., Zhao, H., Wang, F., Zeng, Q., & AMOS, S. (2022). Fusing LDA Topic Features for BERT-based Text Classification.

[48] Zhou, M., Kong, Y., & Lin, J. (2022, July). Financial Topic Modelling Based on the BERT-LDA Embedding. In 2022 IEEE 20th International Conference on Industrial Informatics (INDIN) (pp. 495-500). IEEE.

[49] Paul, P. C., Uddin, M. S., Ahmed, M. T., Hoque, M. M., & Rahman, M. (2022). Semantic Topic Extraction from Bangla News Corpus Using LDA and BERT-LDA. In 2022 25th International Conference on Computer and Information Technology (ICCIT) (pp. 512-516). IEEE.

[50] Milajevs, D., Sadrzadeh, M., & Roelleke, T. (2015). IR meets NLP: On the semantic similarity between subject-verb-object phrases. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval (pp. 231-240).

[51] Joseph, K., & Carley, K. M. (2016). Relating semantic similarity and semantic association to how humans label other people. In Proceedings of the First Workshop on NLP and Computational Social Science (pp. 1-10).

[52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[53] Nugroho, K. S., Sukmadewa, A. Y., & Yudistira, N. (2021). Large-scale news classification using bert language model: Spark nlp approach. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (pp. 240-246).

[54] Gupta, H., & Patel, M. (2021). Method of text summarization using LSA and sentence based topic modelling with Bert. In 2021 international conference on artificial intelligence and smart systems (ICAIS) (pp. 511-517). IEEE.

[55] Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.

[56] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

[57] Zhong, S. (2005). Efficient streaming text clustering. Neural Networks, 18(5-6), 790-798.

[58] Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. Mining text data, 77-128.

[59] Mohammed, S. M., Jacksi, K., & Zeebaree, S. (2021). A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 22(1), 552-562.

[60] Deshpande, R., Vaze, K., Rathod, S., & Jarhad, T. (2014). Comparative study of document similarity algorithms and clustering algorithms for sentiment analysis. International Journal of Emerging Trends in Technology and Computer Science, 3(5), 196-199.

[61] Dhina, M. M., & Sumathi, S. (2021). An innovative approach to classify hierarchical remarks with multi-class using BERT and customised naïve bayes classifier. International Journal of Engineering, Science and Technology, 13(4), 32-45.

[62] Pavithra, M., & Parvathi, R. M. S. (2017). A survey on clustering high dimensional data techniques. International Journal of Applied Engineering Research, 12(11), 2893-2899.

[63] Doan, M. T. (2019). Scalable clustering of high dimensional data in non-disjoint axis-parallel subspaces (Doctoral dissertation, University of Melbourne, Parkville, Victoria, Australia).

[64] Asyaky, M. S., & Mandala, R. (2021). Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP. In 2021 8th international conference on advanced informatics: Concepts, theory and applications (ICAICTA) (pp. 1-6). IEEE.

[65] Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

[66] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

[67] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[68] Cai, T. T., & Ma, R. (2022). Theoretical foundations of t-sne for visualising high-dimensional clustered data. The Journal of Machine Learning Research, 23(1), 13581-13634.

[69] Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-SNE effectively. Distill, 1(10), e2.