FAITHFUL CONCEPT-BASED EXPLANATIONS FOR PARTITION-BASED DOCUMENT CLUSTERING

by

Syed Mohammad Baqir Husain

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2024

Table of Contents

List of Tables

List of Figures

List of Abbreviations Used

ELMO Embeddings from Language Models

BBM black-box Model

CDCEM Conceptual Document Clustering Explanation Model

N.B. Naive Bayes

L.R. Logistic Regression

D.T. Decision Tree

Abstract

This research introduces the Conceptual Document Clustering Explanation Model (CDCEM), a novel explanation model for explaining unsupervised textual clustering. CDCEM explains the discovered clusters and document assignments. Furthermore, it ensures faithfulness—meaning it accurately reflects the decision-making process—using the core elements of black-box textual clustering, such as document embedding and centroids from k-means. This faithfulness and comprehensiveness boost user trust and understanding and help debug clustering. Using Wikipedia, CDCEM first performs wikification, which extracts real-world concepts from the text. It then evaluates these concept's significance for cluster assignment to produce concept-based explanations. CDCEM determines the importance of each concept within a cluster by measuring the cosine similarity between the concept's embedding (representing its contextual meaning) with the cluster centres (representing the cluster's theme), both of which it derives from a black-box model (using ELMO for embeddings and K-means for clustering). This concept's importance for each cluster facilitates generating concept-based explanations at two levels: cluster-level explanations, which describe the concepts that best represent the clusters, and document-level explanations, which clarify why the black-box model assigns a document to a particular cluster. We quantitatively evaluate the faithfulness of CDCEM using AG News, DBpedia, and Reuters-21578 datasets, comparing it with explainable classification methods (Decision Tree, Logistic Regression, and Naive Bayes) by treating clusters as classes and computing the agreement between the black-box model's predictions and explanations. Additionally, a user study was conducted to compare CDCEM with the best baseline in terms of comprehensiveness, accuracy, usefulness, user satisfaction, and usability of the explanation visualization tool on the AG News dataset. CDCEM showed higher faithfulness than the baseline model in quantitative evaluations, indicating accurate explanations of unsupervised clustering decisions. Qualitative evaluations revealed that users preferred CDCEM's cluster-level and document-level explanations for accuracy, clarity, logic, and comprehensibility.

Acknowledgements

I want to express my deepest gratitude to everyone who supported me while writing my thesis, titled "Faithful Conceptual Explanations for Partition-Based Document Clustering."

First and foremost, I owe a huge thank you to my supervisors, Professor Evangelos E. Milios and Professor Enayat Rajabi. Your guidance, insightful feedback, and constant support have been invaluable. Your patience and encouragement kept me going, and I am incredibly grateful for everything you have done.

A big shout-out to my mentors, colleagues, and friends who have been there for me. Lakshita, Rakshit, Devi Ayyagari,Anmol,Shubh, Mariano Maiso, Aman, Juan, and Akhil—thank you all for your support, advice, and friendship. Your help and our stimulating discussions played a big part in getting this thesis done.

I also want to thank my father, S. M. Anwar Husain, for always believing in me. Your encouragement and unwavering support have been my greatest motivation throughout this journey.

Lastly, I want to acknowledge the inspiration and support from my supervisors. Your dedication to helping me grow as a researcher has been a significant driving force behind this work.

Thank you all for making this possible.

Chapter 1

Introduction

Our research addresses the challenge of enhancing explainability in unsupervised document clustering. Explainability in clustering means providing reasons that humans can understand why a complex clustering model groups documents in a certain way. These clustering models are so complex and large-scale that it's hard to understand how they assign documents to clusters. Therefore, there is an apparent necessity for an explanation model that can translate these complex decisions into understandable terms. The need for explainability is critical in unsupervised clustering, which, unlike classification tasks, lacks predefined categories or ground truth. The absence of ground truths often makes clustering subjective, as different algorithms can organize datasets into various patterns and themes tailored to specific audiences or perspectives.

Consequently, it becomes crucial to determine whether the patterns identified by a black-box clustering model align with the user's expectations. In this context, explanations play a vital role, helping end-users understand the model's decisions and enabling developers and domain experts to identify and correct any incorrect cluster formations or misassignments. Achieving a consensus on the generated clusters and gaining insights into the underlying clustering patterns and themes makes the results more practical and influential.

In the context of Explainable AI (XAI), "faithfulness" refers to the accuracy with which an explanation reflects the accurate reasoning process of the AI model. Faithfulness measures how well the explanation captures the complex AI system's actual workings and decision-making process [29]. A faithful explanation accurately represents the model's behaviour, ensuring that the insights provided directly relate to how the model processes and interprets data. Using a simpler explanation model to generate faithful and user-friendly explanations for a complex black-box model is a challenging task [18]. This complexity becomes even more pronounced in the domain of document clustering. Maintaining faithfulness is more challenging in document clustering because of the complicated interaction between clustering algorithms, such as k-means and document representations, like BERT (Bidirectional Encoder Representations from Transformers) embeddings. Trivial explanation approaches are merely exploratory analyses, exemplified by methods like FREQ [11], which attempt to characterize clusters based on the most frequent words. A closer examination reveals that these approaches spotlight standard terms such as 'news,' 'report,' or 'analysis' as indicative of the clusters but do not necessarily contribute substantially to the decision-making process within the **black-box Model (BBM)**. Considering such terms as explanations would be equivalent to misleading interpretations. These approaches fail to capture the intricate logic underpinning these opaque BBM operations. In addition to faithfulness, another challenge in explainability involves choosing the granularity of explanations, which can be global or local. These types of explanations aid in comprehending the decision-making process of black-box models at varying levels of detail.

Researchers use these two kinds of explanations—global and local—to interpret AI models and understand their decision-making processes at different levels of detail [28]. We explain each as follows:

Global Explanations: These provide an overall understanding of the model's behaviour. They aim to explain the general decision-making process of the model across all instances. Global explanations assist in understanding the model's logic on a macro scale, illustrating how different features generally affect predictions [7].

Local Explanations: In contrast, local explanations focus on individual predictions. They aim to explain why a model made a specific decision for a single instance. This explanation is beneficial for understanding the model's behaviour in a specific context or for a particular data point [51].

Many explanation approaches in document clustering primarily offer either local or global explanations. For instance, methods like TREQ [11] focus on providing global explanations by identifying representative words for each cluster. Conversely, approaches like LIME [36] provide local explanations, shedding light on why the BBM assigns a document to a specific cluster. We contend that global and local explanations are indispensable components of a comprehensive understanding. The absence of global explanations can lead to confusion in cluster assignments, as exemplified by the following scenario:

- Doc 1: Progress in medical technology revolutionizes healthcare. (Cluster 1)
- Doc 2: Groundbreaking cancer treatment offers hope for patients. (Cluster 2)

Without global explanations, it becomes difficult to understand why these documents belong to different clusters. This lack of clarity also makes it difficult to see how well the system separates technology (cluster 1) from healthcare documents (cluster 2).

Whereas, in a news clustering system devoid of local explanations, the following global explanation is provided as follows:

- Cluster 1: “Science News”
- Cluster 2: “Sports News.”

Now, consider the cluster assignment for this document:

- Document: “Biomechanics and aerodynamics are crucial for sports performance.”

The absence of local explanations makes it unclear why the document resides in Cluster 2 (Sports News). In case of misclassification, users cannot identify which parts of the document contribute too much or too little. This lack of clarity hampers explainability and the ability to rectify misclassifications effectively. Explanation approaches capable of providing both local and global explanations encompass post-hoc interpretable models like Naive Bayes (N.B.) and Logistic Regression (L.R.). These models are called post hoc because they are applied after performing clustering. These models are also called interpretable because their logic is simple enough for humans to understand their predictions. Interpretability refers to the extent to which a human can understand the cause of a decision made by such a model. These models can generate global explanations in the form of representative words and local explanations that clarify which words influence the document’s proximity to specific clusters. However, these post hoc models exhibit two notable limitations. Firstly, they may not offer explanations in terms of meaningful concepts extracted from the text. Secondly, being model-agnostic, they may disregard the inherent logic of the BBM and provide explanations based solely on their understanding of the data. Model-agnostic methods can be applied to any machine learning model, regardless of its specific structure or learning algorithm. To address the limitations of faithfulness and comprehensiveness in explainability, we introduce a novel approach to explaining document clustering in our research. Our proposed model, CDCEM, generates feature-based local and global explanations. Local explanations highlight the contribution of textual features to justify the placement of documents in a cluster. In contrast, global explanations identify the essential features of each cluster, highlighting the overall theme of each cluster. CDCEM is specifically tailored to explain the clustering performed by a particular BBM framework. In this framework, document embeddings are clustered using a partition-based method inspired by state-of-the-art techniques [11]. To ensure our explanations align with the clustering logic of the BBM, we utilize two fundamental elements: document embeddings generated by a pre-trained language model and cluster centre embeddings derived from the BBM's partition-based clustering algorithm. These embeddings are crucial as they help identify significant features within documents and facilitate feature contribution-based explanations. We gain insights into the intricate clustering process by analyzing both the document's feature representation and the cluster centres, which serve as pseudo-ground truths. This method allows us to move beyond merely examining the inputs and outputs of the BBM, incorporating its core components to produce explanations that accurately reflect the model's decision-making process.

Furthermore, CDCEM employs Wikification, where document features or phrases are linked to relevant Wikipedia concepts (each Wikipedia page represents a concept) rather than using raw text snippets or isolated words. This strategy avoids including punctuation and stop-words, unlike a few contemporary explanation models [43]. Instead, our explanations use meaningful and widely recognized real-world concepts. Therefore, our explanations offer a clear understanding of the model's workings in terms of relevant and significant concepts without the clutter of less informative elements, making them easier to interpret and more informative.

These design strategies enable explanations that faithfully reflect the 'black-box' nature of the clustering model and are also more comprehensive and user-friendly. However, it is essential to note that CDCEM has a significant limitation due to its design choice. It is tailored to a particular clustering framework and, therefore, is not model-agnostic, unlike methods like LIME [36] or SHAP [25].

There are two **claims** that we make.

- First, we assert that our developed post-hoc explanation model, which utilizes both the intermediate and final outcomes of the clustering model, provides more faithful explanations.

- Secondly, we contend that integrating local explanations at the document level with global explanations at the cluster level improves the overall quality and comprehensiveness of the explanations for users in unsupervised document clustering. The improvement is significant for perceived accuracy, clarity, logic, and comprehensibility of the clustering explanations.

This research introduces the following contributions:

1. A novel cluster explanation model: we designed our model, CDCEM, to provide faithful and comprehensive explanations for clusters and individual document assignments, all grounded in real-world concepts.

2. An extensive evaluation of CDCEM: we conducted a user study to evaluate our method qualitatively against the best-performing baseline. In parallel, we performed a quantitative evaluation with baselines to assess our approach's faithfulness and alignment (fidelity) with the BBM.

3. Tool to visualize clustering explanation: we constructed a user-friendly visualization tool demonstrating feature-based explanations for unsupervised document clustering.

The thesis is organized as follows: Chapter 2 provides a comprehensive literature review, covering document clustering frameworks, explainable clustering, the need for comprehensive explanations, and the challenges and open problems in the field. Chapter 3 details the methodology, including problem formulation, the steps for generating explainable models, and visualization techniques. Chapter 4 presents the experimental setup, including datasets, baselines, evaluation metrics, and both quantitative and qualitative(user study) results. Chapter 5 discusses the results, focusing on quantitative analysis and insights from the user study. Finally, Chapter 6 concludes the thesis with a summary of findings and directions for future work.

Chapter 2

Literature review

This chapter explores the key concepts and methodologies in document clustering and explainable artificial intelligence (XAI). We start by providing an overview of the document clustering frameworks relevant to our objective of generating clear explanations. Next, we highlight the crucial importance of explainability in clustering models. We then examine significant gaps in the current literature, such as the demand for more thorough explanations and the advantages of incorporating concept-based explanations. Finally, we discuss baseline methods and evaluation criteria, shedding light on the challenges and open issues our proposed method aims to tackle.

2.1 Document Clustering Frameworks

Clustering document data is a fundamental task in various applications, typically involving feature extraction followed by the application of clustering techniques [24, 46, 27, 34]. A noteworthy state-of-the-art framework, as exemplified by [11], incorporates a pre-trained text encoder (like Embeddings from Language Models (ELMO)) for feature extraction and employs K-means to generate clusters. ELMO is a deep learning model that generates contextualized word representations by employing a bidirectional LSTM (Long Short-Term Memory) architecture trained to predict the next word in a sequence [31]. K-means then uses these embeddings to identify clusters within the data.

2.2 Explainable Clustering

While these clustering frameworks excel in grouping documents, they often fall short regarding explainability. They are frequently referred to as “black-box” models, chosen primarily for their superior performance over inherently transparent models [28] like Decision Tree (D.T.) [4], N.B. [20], or L.R. [17]. Nevertheless, there is a growing demand for models combining performance and explainability, facilitating trust and understanding of the logic behind document clustering.

2.3 The Need for Comprehensive Explanations

Addressing the need for comprehensive explainability, we argue that clustering models should explain both cluster formation and document assignments. Existing approaches tend to focus on representative keywords for each cluster but often neglect individual document-to-cluster assignments [11, 49] or vice versa [36, 25] [21]. Some noteworthy methods from the literature that can provide both include: L.R., which assigns cluster numbers as pseudo-labels to words, elucidating their contributions to the decision-making process. (e.g., [11, 33, 8, 44]); D.T., which yields if-else rules as explanations. (e.g., [30, 1, 45]) and N.B., which provides explanations for features through posterior probability calculations. (e.g., [32, 14, 15, 40])

2.4 Concept-Based Explanations

There is a pressing need for a new approach to concept-based explanations. Existing attempts, both for classification and clustering, have fallen short in providing semantically meaningful insights. This gap in our understanding underscores the importance of our research and the potential impact of a successful solution [43].

Researchers have hinted at a potential solution by leveraging knowledge graphs (like Rožanec et al. [39], and Lecue et al. [22]), such as Wikidata, which hosts millions of interconnected items representing real-world concepts [47]. By extensively drawing from Wikipedia to populate its database, Wikidata helps us extract meaningful concepts from the text, which can then be used to generate feature/concept-based explanations, providing richer and more meaningful cluster explanations.

2.5 Baselines

In the existing literature for non-textual data, researchers generate explanations for clustering using a two-step procedure post-clustering [2]. Firstly, we train a classifier after clustering using cluster indexes as labels. Then, we explain this classification task using Explainable Artificial Intelligence (XAI) methods or employing an inherently transparent model. One existing work we found that deals with textual data instead of tabular following the mentioned explanation generation method is by Guan et al. [11], where they trained an L.R. classifier to identify the representative concepts for each cluster to explain them. We chose D.T., L.R., and N.B. as our baselines due to their established use([11, 33, 8, 44, 30, 1, 45, 32, 14, 15, 40]), interpretability, and ability to act as surrogate models providing both global and local explanations. Each model brings distinct explanatory strengths to our study. Below is a brief description of each method and the nature of the explanations they offer:

1. **Decision Trees (D.T.)**

    - **Nature of Explanation**: D.T.s provide explanations through their transparent, hierarchical structure. Each node and branch in the tree represents a decision rule based on the features of the document, leading to a clear path from input to cluster assignment.

    - **Explanation Type**: They are particularly effective in providing local explanations, as one can trace each document's clustering path. On the other hand, the tree's overall structure can be seen as a global explanation, showcasing the general decision-making logic.

2. **Logistic Regression (L.R.)**

    - **Nature of Explanation**: This method offers explanations in terms of the significance and weight of each feature in determining the cluster assignments. The coefficients in L.R. models indicate the direction and strength of the relationship between features and the predicted outcome.

    - **Explanation Type**: L.R. primarily provides global explanations, as it elucidates the overall influence of features on the clustering process. However, the impact of each feature on specific instances can also be interpreted, thus offering a form of local explanation.

3. Naive Bayes (N.B.)

- **Nature of Explanation**: N.B. explains decisions through a probabilistic approach. It calculates the probability of a document belonging to a particular cluster based on the features it exhibits, assuming independence among these features.

- **Explanation Type**: The model's strength lies in offering local explanations by detailing the probability distribution that led to a document's clustering. In terms of global explanation, it provides insights into the general probabilistic trends influencing the clustering across the entire dataset.

Each of these methods brings a unique perspective to the table. D.T.s, with their rule-based approach; L.R., with its feature-weight significance; and N.B., with its probabilistic reasoning, together, form a comprehensive set of baselines. This diversity ensures a thorough evaluation of our approach to generating explanations for document clustering, highlighting the strengths and potential areas for improvement in our XAI methodology.

2.6 Evaluation of Explanations

The evaluation of traditional clustering algorithms often relies on metrics such as the Silhouette Coefficient [38], the Adjusted Rand Index (ARI) [48], Normalized Mutual Information (NMI) [23], and Accuracy [23] to test clustering quality. However, assessing the quality of explanations provided for clustering algorithms requires different metrics. Various studies use different terminologies for several aspects of explanation evaluation. Three criteria for evaluating explanations have been clearly articulated and differentiated [19].

- Readability: The ease with which users can comprehend the explanation. Literature evaluates readability through quantitative metrics that check the number of features used in the explanation, the length of the explanation [37], or through interactive user studies.

- Plausibility: The degree to which an explanation convinces users, especially when dealing with BBMs. Plausibility is evaluated with user feedback or without user through proxies for explanation quality, such as the quality of feature importance or the performance of the explanation model [9].

- Faithfulness: How accurately an explanation characterizes the underlying BBM. Faithfulness, however, cannot be judged by humans because the workings of the BBM are unknown. If humans understood the rationale of the BBM, we would not need an explanation [18]. Several methods to evaluate faithfulness are summarized in a survey by Lyu et al. [26]. One method recommended to evaluate faithfulness is predictive power evaluation [26]. This method evaluates whether the surrogate explanation model's predictions on new data align with the BBM's, thereby indicating how accurately it captures the BBM's rationale [26].

2.7 Challenges and Open Problems

Explaining a BBM, as required for clustering, involves the challenge that we expect a simpler explanation model to encapsulate the intricate logic of the underlying complex black-box system, which inherently leads to a lossy nature of the interpretation process [19]. The clustering model explanation introduces its unique set of challenges, as it requires explanations for clusters and assigning a document to the clusters. This complexity is made even more challenging by the need to balance the performance of advanced algorithms with the clarity and understandability of the explanations provided. Current methods often focus on the logic behind cluster formation or the specifics of document-to-cluster assignments, but rarely both, limiting their practical utility. Moreover, there is a significant challenge in enhancing the semantic richness of explanations to include more nuanced and contextually relevant concepts, pushing us to explore further methods that can deliver technically accurate and meaningful explanations to users.

Chapter 3

Methodology

We propose an innovative explainable model that builds upon black-box clustering techniques, addressing the explainability challenges detailed in Section 2.7. This post-hoc explanation model generates feature-based explanations to justify the placement of documents in particular clusters and elucidates the theme of each cluster by presenting the most significant Wikipedia concepts within the documents. Our explanation model takes these clusters, generated by K-means, and identifies key concepts prevalent within each group by examining their proximity to the cluster centroids in the embedding space. We use these concepts to generate intuitive explanations for the clustering outcomes, helping users understand the basis for document groupings.

Our approach uses intermediate clustering results to enhance explainability and ensure that the explanations are faithful to the BBM's decision-making process.

For enhanced accessibility and user engagement, we have developed a visually intuitive web tool. This tool, designed with the user in mind, presents the explanations generated by our clustering algorithms in a clear and understandable manner. It assists users in grasping the intricate relationships and thematic structures within clusters, making the process of understanding the clustering outcomes more straightforward and user-friendly.

3.1 Problem Formulation

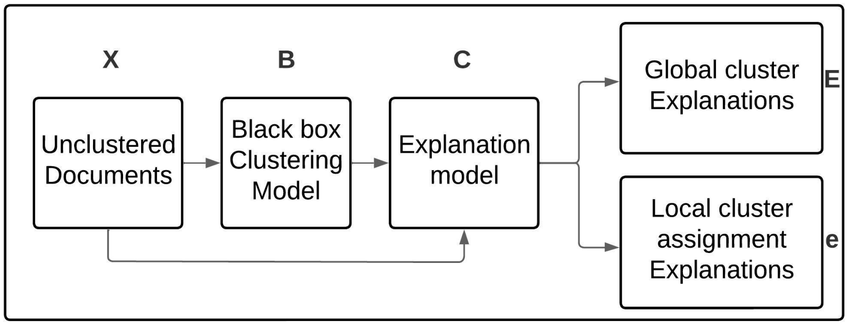

X
B
C
Unclustered Documents
Black box Clustering Model
Explanation model
Global cluster Explanations
E
Local cluster assignment Explanations
e

Figure 3.1: Defining the Problem for an Explanation Model in Unsupervised Document Clustering. This figure presents the challenge of developing an explanation model c for interpreting a black-box clustering model B. Model c is expected to elucidate the clustering of a document dataset X, offering Global explanations E for cluster-wide themes and Local explanations e for individual document placement. These explanations, $E = \epsilon_g(c, X_{\text{cluster}})$ and $e = \epsilon_l(c, x)$, are derived from the explanation functions ϵ_g, ϵ_l, providing insights into the clusters and assignment of a document to the clusters, respectively.

We formally define the problem of explaining black-box models using an explanation model based on the survey paper [12]. While their formulation addresses generic black-box predictors, our definition specifically applies to unsupervised document clustering. We combine elements from the Model Explanation Problem, which focuses on providing global explanations for the entire model, and the Outcome Explanation Problem, which focuses on local explanations for specific predictions. Our combined formulation allows us to incorporate cluster-level (global) explanations that describe the clusters and document-level (local) explanations that clarify why the black-box model assigns a document to a particular cluster. Our combined problem formulation is as follows:

Given a dataset X and a black-box model B, we seek to develop an explanation model c such that:

$$c = f(B, X)$$

Wherein the objectives are to determine the following:

- Cluster Explanation (E): A comprehensive insight into the cluster formulated using the subset of X that pertains to the identified cluster, expressed as:

$$E = \varepsilon_g(c, X_{\text{cluster}})$$

Here, $X_{\text{cluster}} \subseteq X$ represents the data points within X that belong to a specific cluster.

- Assignment Explanation (e): A detailed rationale behind the assignment of an individual instance x to a specific cluster, expressed as:

$$e = \varepsilon_l(c, x)$$

Where:

- B is a complicated model like ELMO [31], which is uninterpretable by humans.
- The explanation model c is designed to provide insights at different levels of granularity:
  - Cluster Explanation, also referred to as Global Explanations (E), are derived through the process ε_g, providing an overview of the data points within a specific cluster.
  - Assignment Explanation, also referred to as Local Explanations (e), are produced by the process ε_l, focusing on the individual assignment of data points to clusters.

3.2 Overview

Before detailing the steps of our explainable model, we present an overview of the explanation generation methodology for document clustering, as outlined in Fig. 3.2. This overview highlights the inputs, outputs, and processes of each step, emphasizing their contributions to the final explanation.

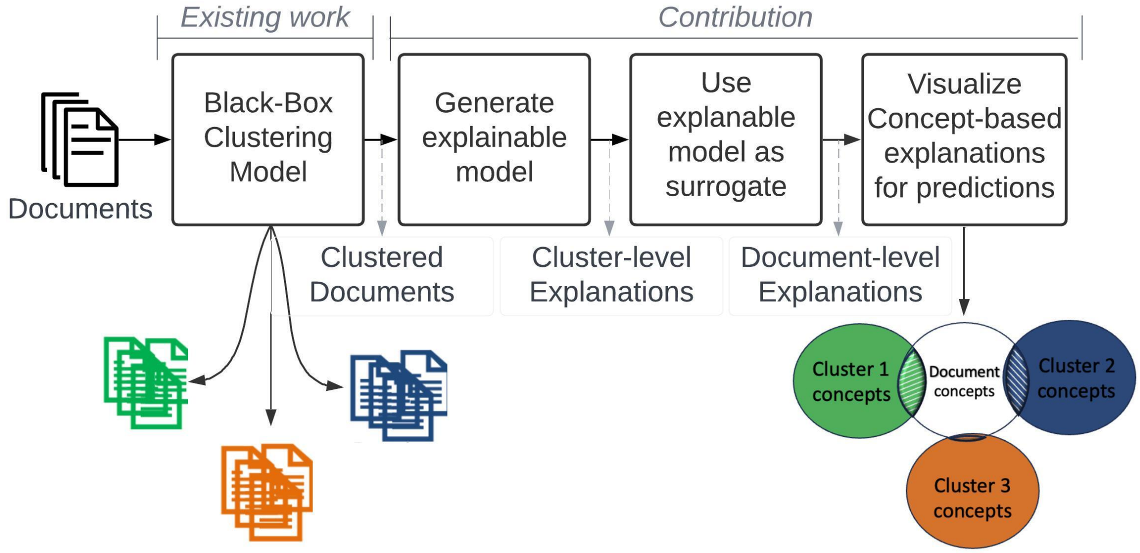

Existing work
Contribution
Documents
Black-Box Clustering Model
Generate explainable model
Use explainable model as surrogate
Visualize Concept-based explanations for predictions
Clustered Documents
Cluster-level Explanations
Document-level Explanations
Cluster 1 concepts
Document concepts
Cluster 2 concepts
Cluster 3 concepts

Figure 3.2: Methodology Overview. This figure outlines the process of interpreting document clusters, from the initial clustering by a black-box model to the articulation of cluster themes via global explanations and individual document assignments via local explanations. The process concludes with the visualization of these explanations. Each square block corresponds to a step of our methodology.

- **Step 1: Black-Box Clustering**

  In the initial step (Block 1 in Fig. 3.2), the process starts with the input of a dataset of documents and the desired number of clusters (k). These documents are embedded using ELMO embedding and clustered via the k-means algorithm. The output of this step consists of the labels for the clustered documents along with their respective k cluster centres, which embody the core traits of each cluster. We also store document embeddings to utilize them in later steps. This foundational step organizes the documents into coherent groups based on their content similarities, setting the stage for generating insightful explanations. In later steps, we utilize the cluster centres and contextual embedding generated for the words of each document.

- **Step 2: Generate Explainable Model**

  The second step (Block 2 in Fig. 3.2) involves generating an explainable model by extracting concepts from the dataset and quantifying concept-cluster relationships.

  - **Step 2a: Extracting Concepts from Clustered Documents**
    Following the clustering, this sub-step takes the dataset as input. The output for this stage includes concepts extracted from the dataset and linked to corresponding Wikipedia pages. We use a tool, Wikifier [3], to annotate the documents with real-world entities. This sub-step enhances cluster explainability by using these real-world concepts, ensuring that the explanations are grounded in recognizable and meaningful ideas.

  - **Step 2b: Quantify Concept-Cluster Relationships**
    In this sub-step, the inputs are the concepts annotated in documents (from the previous sub-step) and the previously identified cluster centres and document embeddings (from Step 1). The output is a concept-cluster similarity matrix that contains cosine similarity scores, quantifying the relationship between each concept's embedding and the cluster centres. This quantification is critical as it measures how closely concepts align with the thematic essence of each cluster, thus providing a concrete basis for evaluating the contribution of concepts to clusters. Additionally, the similarity score allows us to identify concepts with strong relationships or associations with each cluster. Such quantification is essential for building explanations that accurately reflect the clustering logic.

- **Step 3: Use Explainable Model as Surrogate**
  This step (Block 3 in Fig. 3.2) focuses on explaining individual document assignments to clusters. The input includes a document and the concept-cluster similarity matrix (developed in the previous step). By mapping the concepts identified within a document to the concept vocabulary and extracting importance scores from the similarity matrix for each cluster, we find concept contributions and compute aggregate scores to determine the document's cluster assignment. Acting as a surrogate, the explanation model provides both predictions and explanations for these predictions. Consequently, this step's output comprehensively explains the document's cluster assignment, detailing the model's reasoning behind each categorization.

- **Step 4: Visualize Concept-Based Explanations for Predictions**

The final step (Block 4 in Fig. 3.2) involves visualizing the explanations for document clustering. The input encompasses the role of each document's concepts in determining their cluster assignments (from the previous step) and the most significant concepts for each cluster (from Step 2a). The outputs are visual representations of local explanations for cluster assignments and global explanations for each cluster. These visualizations display the scaled contributions of concepts to clusters, illustrating each concept's relative importance and influence. Additionally, global explanations offer a broader overview of each cluster's prevalent themes or concepts. This step enhances the transparency and interpretability of the clustering process, allowing for a more intuitive understanding of both individual document assignments and overall cluster characteristics.

3.3 Step 1: Utilize Black-Box Clustering

This Section describes the architecture of black-box clustering models (in Fig. 3.3) for which our model generates explanations. We identify and extract core elements from the clustering architecture used in later steps to generate faithful explanations.

The clustering approach we explain uses ELMO embedding combined with the K-means algorithm. Here is a brief description of the process:

- Input text is processed using the pre-trained language model ELMO to generate embeddings that capture the essence of the text.

- Due to variations in text length, the size of the generated representations from the pre-trained language model differs. The document's feature representations are combined to form fixed-size representations by Guan et al. [11].

- A feature normalization module is applied in the third step, utilizing methods like layer normalization. This step ensures the feature's numerical stability and adherence to specific qualities, such as a normal distribution.

- In the final step, the normalized features are fed into a chosen clustering algorithm, K-means.

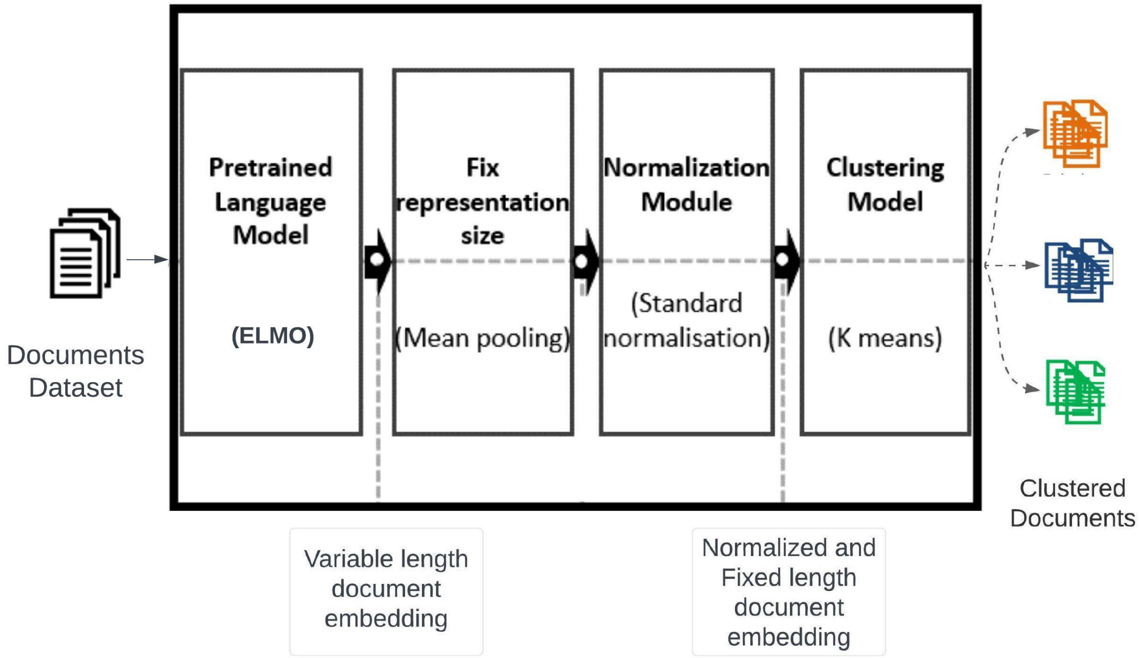


Figure 3.3: Workflow of the black-box clustering model proposed by Guan et al. [11]. This diagram showcases the specific process used for text data clustering, starting with transforming document data into variable-length embeddings via the ELMO pre-trained language model. These embeddings are then standardized through mean pooling and normalization to ensure consistent size and scale, making them suitable for clustering. The K-means algorithm then clusters these normalized and fixed-length embeddings, creating document clusters around centroids within the embedding space. Our explanation methodology utilizes these centroids and embeddings to elucidate the clustering rationale, as shown in Fig. 3.2.

- The outcome of this process is a set of k clustered documents formed around cluster centres within the embedding space.

The black-box model (BBM) organizes the documents into clusters based on their content, with cluster centres acting as focal points representing the commonalities within each cluster in the embedding space. As we advance to the subsequent steps of our methodology, we will continue to leverage the method of generating contextual embeddings with ELMO and the distance metric employed by K-means. This strategic alignment ensures that our explanation model functions in harmony with the BBM, preserving the integrity of our explanations. Without such coherence, we compromise the reliability of our explanations, particularly if the underlying methods of the BBM and the explanation model diverge significantly. This alignment is crucial for maintaining the efficacy and trustworthiness of the explanations we aim to provide.

In the explanation model we develop in the following steps, we will utilize uncompressed variable-length embeddings to identify the semantic contribution of each token, unlike BBM, which uses compressed fixed-size representations for clustering. This approach allows us to pinpoint the specific elements within the text that contribute to the clustering results, providing a detailed and transparent understanding of how the BBM operates.

Algorithm 1 Pseudocode for Black-box Document Clustering

```
Input: Dataset $D = \{d_1, d_2, \ldots, d_n\}$, Number of clusters $K$
Initialize vocabulary set $V = \emptyset$
Initialize set of full document embeddings $F = \emptyset$
Initialize set of preprocessed embeddings $P = \emptyset$
for each document $d_i \in D$ do
    Compute document embedding $f_i = ELMO(d_i)$
    $F \leftarrow F \cup \{f_i\}$    ▷ Add $f_i$ to the set of full document embeddings
    Compute fixed-size embedding $e_i = \text{MeanPooling}(f_i)$
    Normalize embedding $n_i = \text{L2\_Normalization}(e_i)$
    $P \leftarrow P \cup \{n_i\}$    ▷ Add $n_i$ to the set of preprocessed embeddings
    Update vocabulary $V$ with tokens in $d_i$    ▷ $V \leftarrow V \cup \text{tokens}(d_i)$
end for
Apply k-means clustering: $\{L, C\} \leftarrow \text{KMeans}(P, K)$ ▷ $L$ is the set of labels, $C$ is the set of cluster centers
Output: Labels $L$, Cluster centers $C$, Full document embeddings $F$
```

3.4 Step 2: Generate Explainable Model

Building on the black-box clustering model output, we receive a dataset where each document is assigned to specific clusters through indices or labels defined by the BBM. In this step, we model the logic of these clusters, aiming to enhance the transparency of the clustering process.

This step models the relationships between documents and their assigned clusters. At this point, our explanation model starts by extracting relevant concepts from the documents. It then examines the importance of these concepts across different clusters, providing a clear view of the themes that influenced the clustering decisions. This systematic process of extracting and assessing concept importance forms the base for creating detailed and easy-to-understand explanations of the clustering results. This approach ensures that our explanations are both informative and solidly based on the real-world entities found in the documents, effectively connecting abstract cluster labels to specific, observable themes.

The two steps of generating an explanation model are presented below: First, concepts are extracted from the documents. Then, the importance of these concepts across different clusters is assessed, providing a clear view of the themes that influenced the clustering decisions.

3.4.1 Step 2a: Extracting concepts from clustered documents

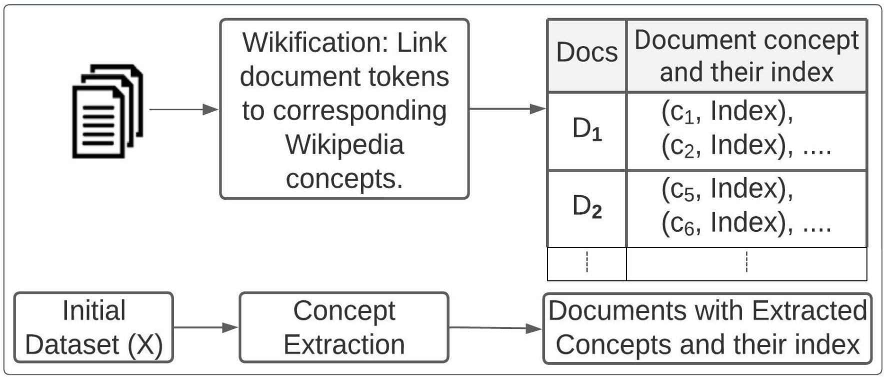
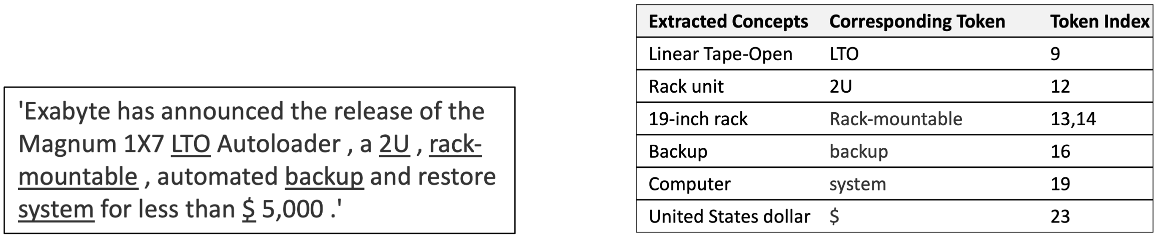

Wikification: Link document tokens to corresponding Wikipedia concepts.
Docs
Document concept and their index
D1
(c1, Index), (c2, Index), ...
D2
(c5, Index), (c6, Index), ...
Initial Dataset (X)
Concept Extraction
Documents with Extracted Concepts and their index

Figure 3.4: Concept Extraction for Explainability. This diagram illustrates the initial stage of our methodology, where the Wikifier tool annotates a document dataset X with Wikipedia-linked concepts. The token index records the mapping between concepts and their associated phrases, which is essential for tracking and referencing throughout the explanation model. The annotation allows our explanation model to generate meaningful insights.



'Exabyte has announced the release of the Magnum 1X7 LTO Autoloader , a 2U , rack-mountable , automated backup and restore system for less than $ 5,000 .'

(a) Input: Sample text from the AG-NEWS dataset

Extracted Concepts	Corresponding Token	Token Index
Linear Tape-Open	LTO	9
Rack unit	2U	12
19-inch rack	Rack-mountable	13,14
Backup	backup	16
Computer	system	19
United States dollar	$	23

(b) Output: Extracted concepts and their support.

Figure 3.5: Illustration of the Concept Extraction from the clustered document 3.4.1. The first image (Fig. 3.5a) shows the input document for concept extraction via the Wikifier tool. The second image (Fig. 3.5b) shows the output containing extracted concepts, their corresponding tokens, and token indices within the document. These serve as the groundwork for our model's explanation generation.

We intend to generate an explanation model based on the importance of features. Features within a document could be phrases or words present within it. Instead of using such phrases or words directly, we chose to find meaningful concepts corresponding to excerpts within the text. By meaningful, we mean that the concept is semantically meaningful on its own, as defined by Ghorbani et al. [10]. We annotate the documents with Wikipedia concepts. We generate explanations grounded in annotated real-world tangible concepts free from punctuation or stopwords.

To extract concepts from previously clustered documents, we extract real-world entities from each cluster's documents with a tool called Wikifier [3]. Wikifier identifies relevant Wikipedia concepts linked to specific phrases in a document. It determines the correct concept for each phrase by constructing a bipartite graph. The phrases are connected to potential concepts with weight based on their frequent association in Wikipedia. Furthermore, concepts are linked among themselves with the weight of their similarity. The page rank algorithm runs on this graph, iteratively distributing PageRank scores among concepts based on their association with phrases and their relatedness to other concepts. Finally, the concepts with the highest page rank value are chosen for each document phrase. It further employs several heuristics to filter the final concepts annotated to phrases within the document to reduce noise and ambiguity. This phrase-concept graph method finds appropriate concepts of all the document phrases in their context, ensuring that the concepts selected are coherent with the document's overall topic. Therefore, the phrase "Tesla" will be linked to a car manufacturing company if other concepts are about cars instead of linking to Nicola Tesla.

To illustrate our concept extraction process clearly, we present an example in Fig. 3.5. This figure depicts a document from a cluster, as shown in Fig. 3.5a. The annotation of concepts within this document using the 'Wikifier' tool is detailed in 3.5b. This tool tags specific phrases in the text with corresponding Wikipedia concepts and indicates the position of each phrase. For instance, the phrase at the ninth token, "LTO", is linked to the Wikipedia concept "Linear Tape-Open." This example demonstrates the tool's ability to extract pertinent concepts from the document.

This extraction step is applied to all documents in the dataset. The details of the extracted concepts are stored and later used to determine their role in assigning documents to their respective clusters.

Algorithm 2 Pseudocode for Annotating Documents Depicted in Fig. 3.4

```
Input: Dataset $D = \{d_1, d_2, \ldots, d_n\}$
Initialize set of document concepts $C_d = \emptyset$
Initialize vocabulary set $V = \emptyset$
for each document $d_i \in D$ do
    Compute document concepts $C_i = \text{Wikifier}(d_i)$    ▷ Wikifier returns a set of tuples $(\text{concept}, \text{index})$
    $C_d \leftarrow C_d \cup \{(d_i, C_i)\}$    ▷ Add $(d_i, C_i)$ to the set of document concepts
    for each $(\text{concept}, \text{index}) \in C_i$ do
        $V \leftarrow V \cup \{\text{concept}\}$    ▷ Add concept to the vocabulary set
    end for
end for
Output: Set of document concepts $C_d$, Vocabulary $V$
```

3.4.2 Step 2b: Quantify Concept-Cluster Relationships

In this phase of constructing the explanation model, we model how various concepts within the documents influence their placement within specific clusters. Concepts more similar to a cluster's characteristics will more likely draw documents toward that cluster than others. We prepare to explain document clustering decisions in subsequent phases by quantifying and documenting these concept-cluster similarities. We first obtain their respective embeddings to measure the similarity between a cluster and a concept. For the cluster's embedding, we use the centroid embedding produced by the k-means algorithm, as the centroid represents the core semantic essence of the cluster. We derive the embedding for a concept from the mean embedding of the phrase linked to the concept generated by the ELMO model before clustering. If a concept occurs in multiple documents within a cluster, we have multiple embeddings of the same concept. We average them to get a single embedding for a concept within a cluster. Once both embeddings are obtained, we calculate the similarity between each concept and the clusters using cosine similarity, which quantifies the alignment between two vectors by measuring the cosine of the angle between them. Specifically, we used the implementation from scikit-learn, to perform cosine similarity. The similarity values ranges between 0 and 1 with 1 being high similarity and 0 indicating no alignment.

We organize the results of these similarity calculations into a table where each entry shows the degree of similarity between a concept and the centre of the clusters. This process is visually depicted in Fig. 3.6. Building on the extracted concepts from step 3.4.1, we produce a concept-cluster similarity matrix using embeddings generated for documents and cluster centres during clustering 3.3.

To understand this step better, consider the example in Fig. 3.7. This step takes the concepts extracted in the previous step as input (shown in Fig. 3.7a). It computes the similarity between the concepts and the clusters. The output of this step is depicted as a heatmap in Fig. 3.7b, which visualizes the similarity between the concepts and each cluster. The step is applied across the dataset to model and gauge concept importance with respect to each cluster.

Crucially, our method uses the same ELMO embedding for these concepts as was used by the BBM during the clustering step. Moreover, like BBM, we use the distance metric to measure the similarity between a concept and a cluster centre. This alignment ensures that the explanation process remains consistent with the workings of our black-box deep learning approach. In other words, we ensure our explanations are grounded in the same logic as the BBM, maintaining that essential connection.

In the next step, we will use this concept-cluster similarity matrix to explain the clusters and the cluster assignment.

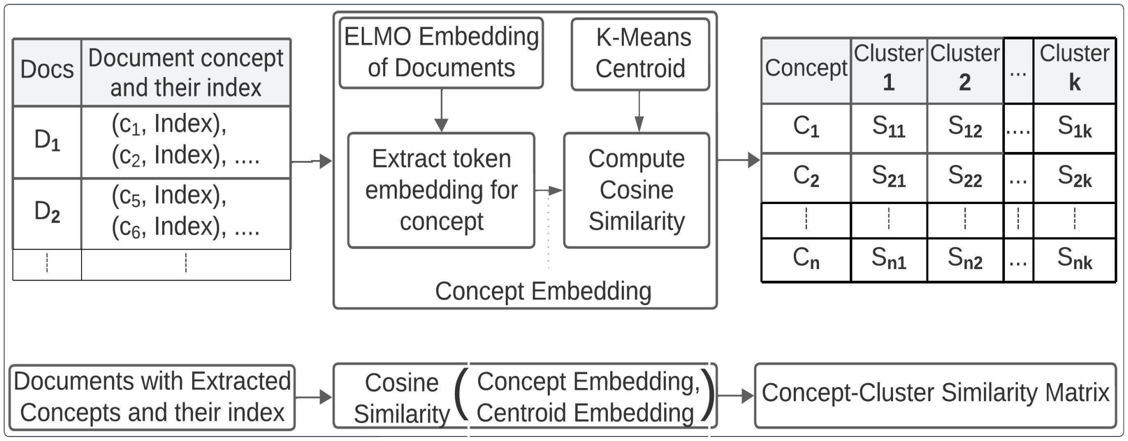
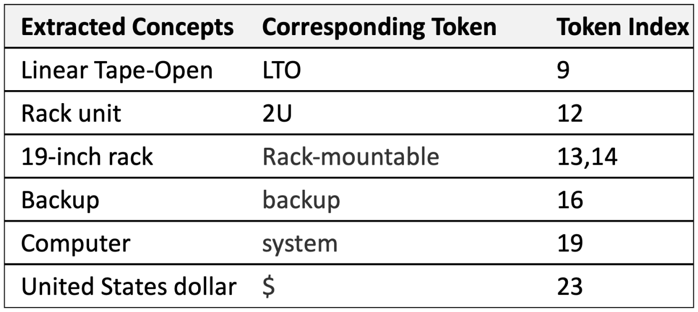

Docs
Document concept and their index
D1
(c1, Index), (c2, Index), ...
D2
(c5, Index), (c6, Index), ...
ELMO Embedding of Documents
K-Means Centroid
Extract token embedding for concept
Compute Cosine Similarity
Concept Embedding
Concept
Cluster 1
Cluster 2
...
Cluster k
C1
S11
S12
....
S1k
C2
S21
S22
...
S2k
Cn
Sn1
Sn2
...
Snk
Documents with Extracted Concepts and their index
Cosine Similarity
Concept Embedding, Centroid Embedding
Concept-Cluster Similarity Matrix

Figure 3.6: Determining Concept-Cluster Importance. This step involves measuring the semantic similarity between document concepts and cluster centroids using cosine similarity. The resulting concept-cluster similarity matrix highlights the relative importance of each concept in the clustering process.

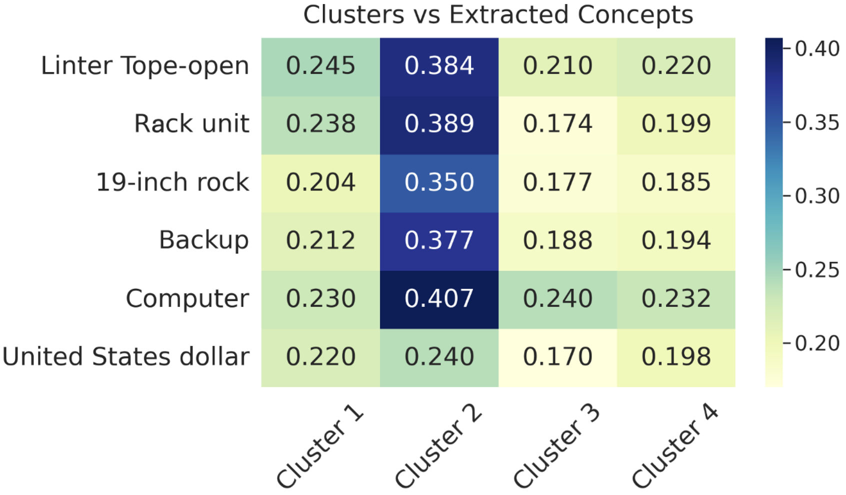
Algorithm 3 Pseudocode for Extracting Concept Importance Depicted in Fig. 3.6

```
Input: Number of clusters $K$, Set of cluster centers $C = \{c_1, c_2, \ldots, c_K\}$, Document concepts list $C_d$, Full document embeddings $F$, Vocabulary $V$
Initialize similarity matrix $S$ = empty list ▷ A list where each entry corresponds to a concept and contains a list of similarity values for $K$ clusters
for each concept $v \in V$ do
    Initialize concept-cluster similarities list $C_{cs}$ = empty list of size $K$ ▷ Holds similarity values between the concept and each cluster centroid
    for each cluster $i \in \{1, 2, \ldots, K\}$ do
        Extract concept embedding ${e_v}^i$ = Extract_Embedding$(v, C_d, F, i)$
        Compute semantic similarity ${s_v}^i$ = cosine_similarity$({e_v}^i, c_i)$
        $C_{cs}[i] \leftarrow {s_v}^i$ ▷ Store the similarity value for cluster $i$
    end for
    $S \leftarrow S \cup \{C_{cs}\}$ ▷ Add the list of concept-cluster similarities to the similarity matrix
end for
Output: Similarity matrix $S$
```



Extracted Concepts	Corresponding Token	Token Index
Linear Tape-Open	LTO	9
Rack unit	2U	12
19-inch rack	Rack-mountable	13,14
Backup	backup	16
Computer	system	19
United States dollar	$	23

(a) Input: Showcasing the concepts extracted from a document with their positions, serving as input for assessing concept-cluster importance.



(b) Output: Visualizing the importance of extracted concepts to each cluster for a sample document, as determined in the concept-cluster importance phase.

Figure 3.7: Illustration of the input and output of the Concept-Cluster Importance phase. Fig. 3.7a displays the concepts extracted from a document and their positions, which serve as input for evaluating concept-cluster importance. Fig. 3.7b visualizes this importance in a heatmap, offering a comparative view of concept significance across clusters, a technique applied dataset-wide to gauge concept importance relative to clusters.

3.5 Step 3: Explain Document Assignments with Concepts

From the previous step, we now have a model encapsulating the relationship between document concepts and the clusters in the form of a concept-cluster similarity matrix. This explanation model is used to elucidate the assignment of documents to their respective clusters.

Given a document, we extract its concepts using Wikifier, as described in Section

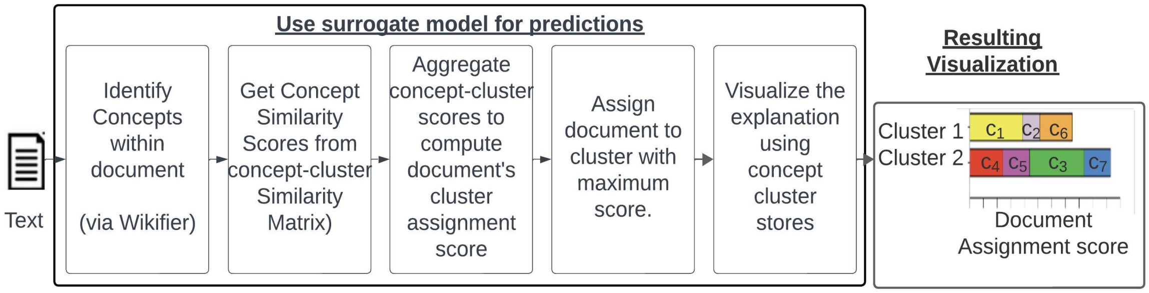

Text
Use surrogate model for predictions
Identify Concepts within document (via Wikifier)
Get Concept Similarity Scores from concept-cluster Similarity Matrix)
Aggregate concept-cluster scores to compute document's cluster assignment score
Assign document to cluster with maximum score.
Visualize the explanation using concept cluster stores
Resulting Visualization
Cluster 1
Cluster 2
C_1
C_2
C_6
C_4
C_5
C_3
C_7
Document Assignment score

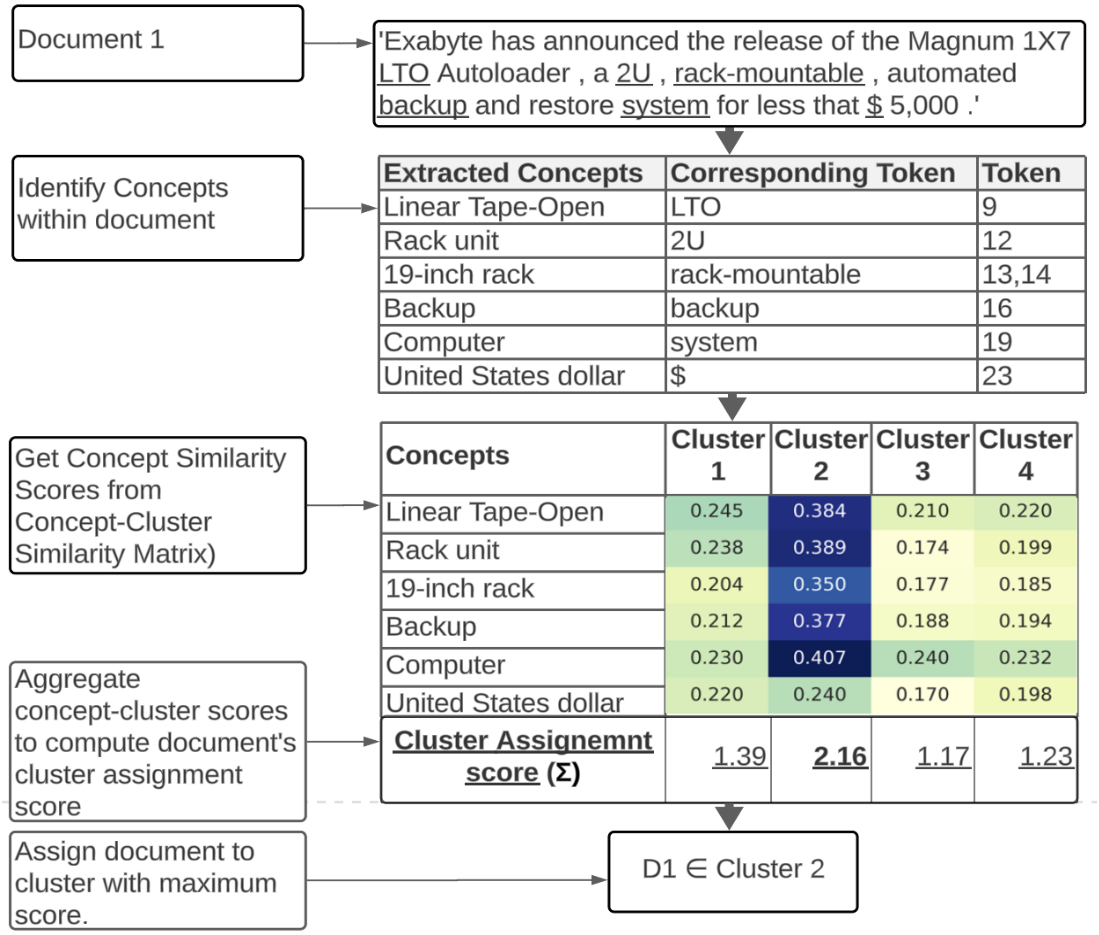
Figure 3.8: Generation of explanations using surrogate explanation model. Starting with concept identification within a document (as introduced in Fig. 3.4), the model concept-cluster similarity matrix (computed in Fig. 3.6) is used to retrieve contribution scores. Our model aggregates these scores to assign the document to a cluster, culminating in a visualization that explains the clustering decision through concept contributions, as highlighted in 3.7b.

3.4.1. We then determine the importance score for each document concept across the clusters from the concept-cluster similarity matrix. These importance scores illustrate how the concepts contribute to the document's inclination toward each cluster.

By aggregating the similarity scores of the concepts, we compute a cluster assignment score for each document. The cluster with the highest assignment score is deemed the document's assigned cluster. Thus, our explanation model provides insights and acts as a surrogate model, making cluster assignment predictions based on the scores. We use this functionality to verify the alignment of our explanation model with the black-box model (BBM).

Consider an example document shown in Fig. 3.9 to demonstrate this step. From the document text, we extract pertinent concepts, such as "Linear Tape-Open." Focusing on these identified concepts, we gather the importance scores for each cluster from the similarity score table. This results in explanations in the form of similarity scores, indicating the document's affinity for Cluster 2. By aggregating these scores, we compute the assignment score for each cluster and conclude that the document belongs to Cluster 2. Thus, our model generates explanations and uses these explanations to predict this document's cluster assignment, serving effectively as a surrogate.

In the next step, we improve the visualization of the explanation to make it more readable and helpful for the user by extracting and presenting global and local explanations from our explanation model.


Document 1
Identify Concepts within document
Get Concept Similarity Scores from Concept-Cluster Similarity Matrix)
Aggregate concept-cluster scores to compute document's cluster assignment score
Assign document to cluster with maximum score.
D1 ∈ Cluster 2

'Exabyte has announced the release of the Magnum 1X7 LTO Autoloader , a 2U , rack-mountable , automated backup and restore system for less that $ 5,000 .'

Extracted Concepts	Corresponding Token	Token
Linear Tape-Open	LTO	9
Rack unit	2U	12
19-inch rack	rack-mountable	13,14
Backup	backup	16
Computer	system	19
United States dollar	$	23

Concepts	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Linear Tape-Open	0.245	0.384	0.210	0.220
Rack unit	0.238	0.389	0.174	0.199
19-inch rack	0.204	0.350	0.177	0.185
Backup	0.212	0.377	0.188	0.194
Computer	0.230	0.407	0.240	0.232
United States dollar	0.220	0.240	0.170	0.198
Cluster Assignemnt score (Σ)	1.39	**2.16**	1.17	1.23

Figure 3.9: Surrogate Explanation Model Application. This figure demonstrates how the proposed explanation model predicts document clustering. It begins by identifying key concepts within a document (as explained in the concept extraction phase 3.4). Next, it assesses the contribution of these concepts to predefined clusters by calculating similarity scores(as shown in 3.6), followed by synthesizing the scores to gauge the document's propensity towards each cluster. The resulting visualization highlights the concept's alignment, particularly with Cluster 2—indicated by the intensity of colour, with darker hues signifying greater similarity. While this serves as an explanation, our model as a surrogate model can also categorize the document into the cluster with the highest cumulative score, exemplified here by its assignment to Cluster 2.

Algorithm 4 Pseudocode for Explainable Prediction Demonstrated in Fig 3.8

```
**Input**: Document $d$, Document concepts list $C_d$, Similarity matrix $S$, Vocabulary $V$, Number of clusters $K$
Extract document concepts $C_i = \text{Wikifier}(d)$
Initialize assignment score vector $A = \{0, 0, \ldots, 0\}$ of length $K$
Initialize contribution matrix $M = \text{zeros}(|C_i| \times K)$   $\triangleright$ Matrix with $|C_i|$ rows and $K$ columns
**for** each concept $c_j \in C_i$ with index $j$ **do**
    Find the index of $c_j$ in $V$: $v_j = \text{index}(c_j, V)$
    Retrieve concept similarities: $S_j = S[v_j]$
    **for** each cluster $i \in \{1, 2, \ldots, K\}$ **do**
        Update assignment score: $A[i] = A[i] + S_j[i]$
        Update contribution matrix: $M[j][i] = S_j[i]$
    **end for**
**end for**
**Output**: $\text{argmax}(A)$, Contribution matrix $M$
```

3.6 Step 4: Visualization

Visualization is a key tool in understanding the clustering model's behavior, providing detailed explanations. It showcases the reasons behind document-cluster assignments, highlighting cluster patterns and the topics covered in the text. This Section outlines methods for both local and global explanations, utilizing various visualization techniques to comprehensively explain the model's internal mechanisms.

3.6.1 Local Explanation

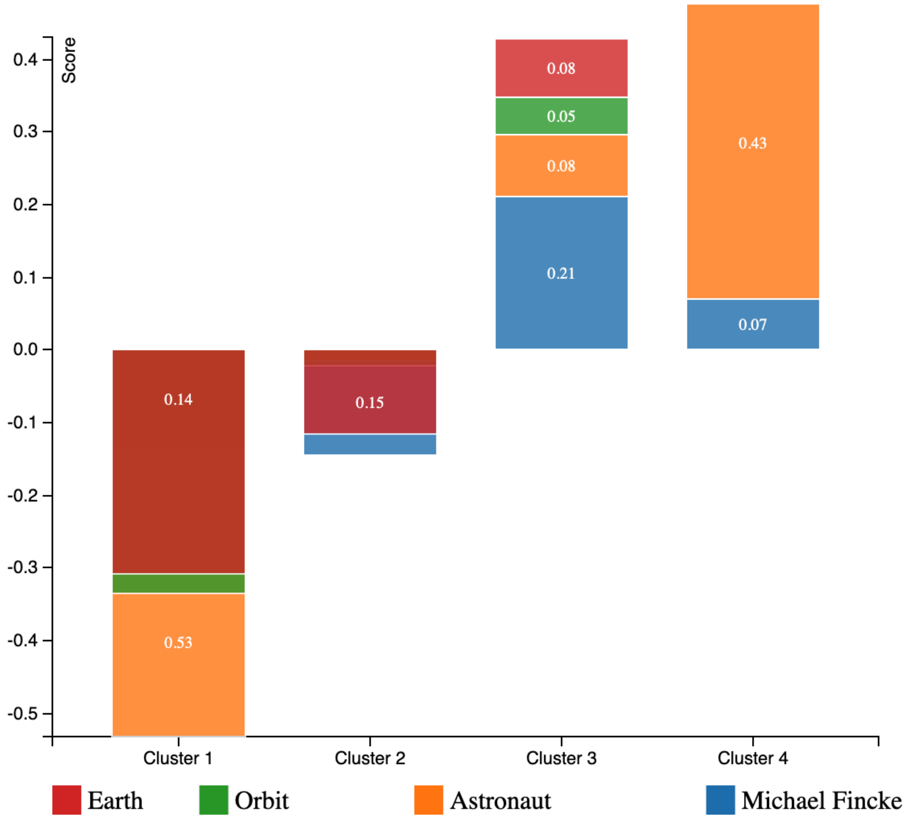


Figure 3.10: Stacked bar chart shows the contribution of document concept in cluster assignment. Each bar, segmented by concept, illustrates its relative contribution. Segments above the baseline denote stronger concept associations with the cluster; those below signal weaker ones. This visualization assists in understanding the clustering model's document assignments from concept embeddings (The visualization is for text documents: *Mike Fincke, whose daughter became the first child born to an astronaut in orbit June 18, is preparing to return to Earth and his family in Texas .)*

Local explanation in the context of clustering explains why a document is assigned to a specific cluster. We visualize local explanation using a bar chart depicting the contribution of a concept in placing the document in a cluster as shown in Fig. 3.10. A vertical bar in the chart across the x-axis corresponds to a cluster. Each bar is made up of segments that represent the individual contributions of document concepts. The total height of the stacked bar represents the document's belongingness to a cluster.

Pre-processing Steps for Highlighting concept Contributions:

We get concept contribution scores from the concept-cluster similarity matrix. We measure these scores by applying cosine similarity between the cluster centre and the concept in the embedding space and considering cosine similarities range from 0 to 1, the preprocessing steps aim to amplify the differences across clusters for better comprehension of the visualization.

1. **Calculate concept mean contribution across Clusters:** For each concept, we calculate the mean contribution score across all clusters. This step establishes a baseline for determining whether a concept's score within a cluster is above or below the overall mean.

   The mean (contribution) score for a concept across all clusters is given by:

   $$\text{meanScore}_c = \frac{1}{N}\sum_{i=1}^{N}\text{score}_{c,i} \tag{3.1}$$

   - N is the number of clusters.
   - $\text{score}_{c,i}$ is the cosine similarity score for concept c in cluster i.

2. **Compute Scaled Relative Differences:** For each concept within each cluster, we calculate the relative difference from its mean across all clusters. This relative difference is then scaled to highlight the variance.

   The applied formula is:

   $$\text{scaledDifference}_{c,i} = (\text{score}_{c,i} - \text{meanScore}_c) \times 10 \tag{3.2}$$

   Here, 10 is a scaling factor selected to enhance the visibility of the differences in the chart.

Interpretation of the Chart

In the local explanation, each bar's overall height represents the sum of scaled relative differences for a document's concepts within a specific cluster (in Fig. 3.10). Positive values indicate where a concept's contribution to a cluster is above the mean, suggesting a stronger association with that cluster. Conversely, negative values indicate a below-mean contribution.

By comparing the size and direction of each segment within the bars, stakeholders can discern the magnitude of a concept's influence and its comparative significance. This visualization method enables a nuanced understanding of the clustering algorithm's behaviour, which is crucial for interpreting complex models in machine learning and data science.

The visualization facilitates a detailed examination of the document's relationship to each cluster, offering insights into the clustering process and the reasons behind the document's assignment to a specific cluster.

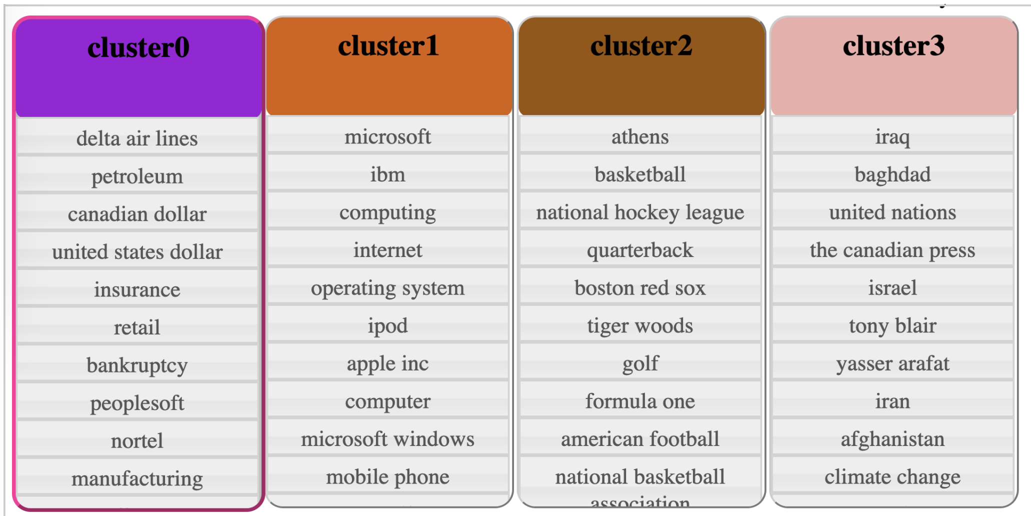
However, this local explanation has limited value since it does not show the theme of each cluster. If it were a classification task, this explanation would suffice. Therefore, we shed light on the clusters using representative concepts, highlighting the overarching theme of a cluster.

3.6.2 Global Explanations

Global explanations in document clustering reveal each cluster's characteristics, themes, or most influential concepts. For our proposed model and baselines, global explanations consist of a sorted list of the most influential real-world concepts extracted from the documents within each cluster. This list provides users with insights into what each cluster represents.

CDCEM : Global explanations

We employ a centroid-based method commonly used in multi-document summarization. Initially, we extract the contextual embeddings of all concepts within the documents of a cluster. Subsequently, we compare these embeddings with the cluster's centre in the embedding space, determined by the k-means algorithm. Utilizing cosine



cluster0	cluster1	cluster2	cluster3
delta air lines	microsoft	athens	iraq
petroleum	ibm	basketball	baghdad
canadian dollar	computing	national hockey league	united nations
united states dollar	internet	quarterback	the canadian press
insurance	operating system	boston red sox	israel
retail	ipod	tiger woods	tony blair
bankruptcy	apple inc	golf	yasser arafat
peoplesoft	computer	formula one	iran
nortel	microsoft windows	american football	afghanistan
manufacturing	mobile phone	national basketball association	climate change

Figure 3.11: The figure illustrates the representative concepts for each cluster, providing a global explanation of the themes and concepts covered by each cluster. This explanation, generated by the proposed approach for the AG News dataset, categorizes the clusters as Finance, Technology, Sports, and World Affairs from left to right.

similarity, we identify the concepts closest to the centre, noting their similarity. The underlying idea is that the concepts nearest to the centre significantly influence the positioning of the cluster in the embedding space. The figure illustrates the global explanations generated by our method for the AG news dataset.

Baseline Methods - Global explanations

The baseline methods do not consider the intermediate results of the BBM, nor do they account for the cluster's centroid. Instead, they rely solely on the input documents and cluster assignments to determine concept importance using coefficients and conditional probabilities. We will examine each baseline method for generating global explanations:

- Decision Tree: We employ Permutation Feature Importance from sklearn for D.T.s to determine the importance of global representative features. This method randomly shuffles each feature's values across the dataset to see how the model's accuracy changes. The more the accuracy decreases with the shuffling of a feature, the more important that feature is considered. This technique, initially introduced in the context of Random Forests by Breiman (2001), helps identify the features most critical for the clustering decisions made by the D.T. model.

- Logistic Regression: For L.R., we extract global explanations by analyzing the coefficients associated with each concept, representing the influence of these concepts on the classification decisions for each class. This process involves using an L.R. model alongside a count vectorizer. The steps include extracting the vocabulary from the vectorizer and the corresponding coefficients from the L.R. We then preprocess the vocabulary to ensure clarity and consistency in naming and sorting the vocabulary and coefficients. This sorted list allows us to see which concepts (features) have the highest coefficients and are, therefore, the most influential for each class. These concepts are then written to a file, forming a clear, class-by-class breakdown of influential concepts, which serves as the basis for our global explanations. This method provides a direct view into how each concept contributes to defining each class, offering valuable insights into the model's reasoning process.

- Naive Bayes: We extract global explanations for the N.B. model by utilizing the inbuilt log probabilities of features for each class. This approach involves examining the log probabilities that the N.B. algorithm assigns to each feature given a class, which indicates the influence or weight of each feature in the decision-making process for that class. Features or concepts with higher log probabilities are considered more representative and influential for their respective classes. This method effectively highlights the most significant concepts that define each cluster, providing clear insights into what characterizes different groupings in the data.

Chapter 4

Experimentation

This Section evaluates the explanations generated for unsupervised clustering as depicted in Fig. 4.1. The primary objective is to identify the best post hoc explanation model that effectively and accurately communicates the reasoning of a BBM to users. Initially, we describe the dataset utilized in the experiments and provide a brief overview of the baseline models, which we have explained in detail in the related work section.

We divide the evaluation into two main components: quantitative and qualitative. The quantitative component assesses the accuracy of the explanation model in aligning with the BBM. The qualitative component, on the other hand, evaluates the user satisfaction with the explanations. Together, these components address our truthtions and test the efficacy of our explanation model.

The outcomes of this experimentation will support or challenge our initial hypothesis, contributing insights to the development of explainable AI for document clustering. Each part of the quantitative and qualitative evaluation includes discussions on experimental design, hyper-parameters, evaluation metrics, observed results, and their implications for the research questions.

4.1 Datasets

In this study, we replicate the experimental setup of the clustering model introduced by Guan et al. [11], utilizing datasets they used to validate our findings. The following text provides an overview of these datasets:

We tested our approach using AG news[1], DBpedia[2], Reuters-21578[3] datasets. The AG news and DBpedia datasets were initially curated by Zhang et al. (2015) [50]. Due to the extensive size of these datasets, conducting experiments on them in their

[1] https://huggingface.co/datasets/fancyzhx/ag_news

[2] https://huggingface.co/datasets/fancyzhx/dbpedia_14

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578

entirety would be inefficient. Hence, we used condensed versions, randomly selecting 1,000 examples from each category in every dataset, in line with Guan et al. [11]. In their initial tests, the performance on these abridged datasets was comparable to the full versions. We derived the R2 and R5 datasets from the Reuters-21578 collection. Although these are labelled datasets, we withheld their labels from the clustering and explanation model. Let us explore deeper into each dataset:

AG news is a dataset meant for news categorization. It comprises top news categories from a vast collection of web news articles from over 2,000 news portals, compiled by Zhang et al. [50]. Each entry in AG news comprises the original headline and article body. It encompasses four sectors: World, Sports, Business, and Science/Technology.

DBpedia is an ontology-based classification dataset developed by selecting specific classes from DBpedia's knowledge framework by Zhang et al. [50]. Each text piece represents an entity's descriptor, with its tag being the ontological class. The dataset includes 14 distinct, non-overlapping categories: Company, Educational Institution, and Athlete.

The Reuters-21578 dataset, initially curated by the Carnegie Group and Reuters, comprises 21,578 documents spread across 135 categories. Notably, this dataset is imbalanced; while some categories have thousands of documents, others barely have a few. Following the experimental setup of Guan et al. [11], we created two new datasets, R2 and R5, encompassing the two and five most populous categories, respectively. R2 contains the 'earn' and 'acq' categories, while R5 includes 'earn', 'acq,' 'crude', 'trade', and 'money-fx.'

4.2 Baselines

In the related work Section 2.5, we carefully selected and briefly introduced the baselines. Now, in this Section, we explore into the comprehensive training process of our baselines (D.T., L.R., and N.B.).

Firstly, we employed wikification as explained in the methodology, Section 3.4.1. This process extracts Wikipedia concepts from the documents. These concepts serve as the foundation for constructing a bag of words model using a count vectorizer in the baseline models. By providing the baselines with a bag of concepts derived from the document, similar to our proposed method, instead of the entire document, we make the comparison even between the baseline and the proposed method. Furthermore, it ensures that the explanations generated by each model are made up of meaningful concepts. Our preliminary experiments showed that when an entire document is provided to the baselines, without wikification, they tend to learn noise within the data, resulting in less faithful explanations.

Similarly to the proposed method, the baselines utilize the predictions from the black-box model as pseudo-class labels (e.g., cluster 0, cluster 1). With these Wikipedia concepts as input and pseudo-class labels, the baselines learn to map the relationship between documents and clusters.

Despite rigorous hyperparameter tuning via grid search, we found that the performance with optimized parameters closely resembled that achieved with default settings. Consequently, we used default settings across various datasets to prevent overfitting and maintain consistency. This approach ensures uniform conditions for evaluating the models under different data scenarios, making it easier to compare performance outcomes.

After training, the baselines evaluate the importance of the document concepts for each cluster via their conditional probabilities (Naive Bayes) or coefficients (Logistic Regression). We utilize these learned weights to create visualizations that illustrate the contribution of features/concepts, as depicted in Fig. 3.10. We derive the explanation model's predictions for each document by applying a softmax function to each cluster's linear aggregation of feature weights.

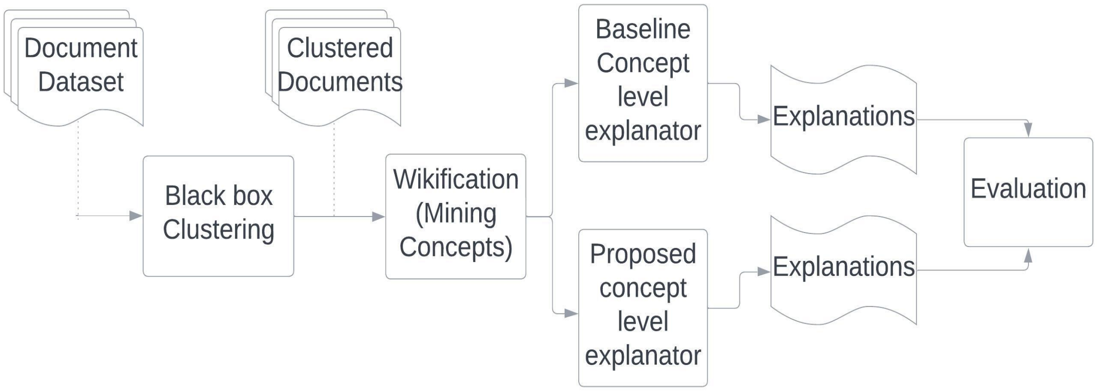

Document Dataset
Black box Clustering
Clustered Documents
Wikification (Mining Concepts)
Baseline Concept level explanator
Proposed concept level explanator
Explanations
Explanations
Evaluation

Figure 4.1: Experimental setup for evaluating the explanation for black-box clustering generated by the proposed approach and baseline methods.

4.3 Evaluation

Adopting a comprehensive approach to evaluation is crucial in assessing Explainable Artificial Intelligence (XAI) systems. Such an evaluation must assess essential properties for an explanation model to demystify the black-box for users effectively. Firstly, the model must accurately and truthfully explain the workings of the complex, often opaque A.I. system (emphasizing faithfulness). Secondly, the explanation should be presented in a way that is easy to understand and convincing for users, addressing their needs for clarity and belief. To assess three fundamental properties—faithfulness, plausibility, and readability—we evaluate each using specific metrics analyzed in quantitative and qualitative sections. The quantitative section measures these properties with statistical and computational metrics (like predictive power evaluation and Fidelity), while the qualitative section evaluates user perception through studies and questionnaires. This approach ensures a comprehensive assessment of each property from technical and user-centric perspectives. Each of these aspects plays a unique role in determining how effective an explanation is. Together, they ensure that the assessment of XAI explanations is technically accurate, comprehensible, and useful to the user. This all-encompassing evaluation approach is essential to ensure that A.I. system explanations are technically sound and meet user's diverse needs and expectations.

4.4 Quantitative Metrics and Evaluation

This Section discusses the quantitative metrics and methodologies used to evaluate the effectiveness of explanation models in capturing and reproducing the behaviour of black-box models (BBMs). We focus on two key metrics, Fidelity and Predictive Power, which collectively assess the internal consistency and generalizability of the explanation models.

4.4.1 Fidelity

Fidelity measures how accurately an explanation model can replicate the behaviour of a black-box predictor, as defined by Guidotti [13]. A low fidelity score could indicate that the explanation model is too simplistic and fails to capture the essential logic of the BBM. In contrast, a high fidelity score suggests that the explanation model effectively mirrors the complex model. High fidelity scores on training data alone do not confirm that the model is faithful, as it might just be memorizing the data. At best, it adds to the plausibility of the explanations to the user.

Experimental Setup for Fidelity Evaluation

The setup for assessing Fidelity involves performing document clustering on the dataset, and then the explanation models learn to assign a document to the cluster predicted by the BBM. Evaluating Fidelity involves making the explanation models predict the cluster label for each document. This prediction of the explanation model is compared with the cluster assignment of the black-box model in terms of accuracy and F1. High-fidelity performance means that the explanation accurately aligns with the prediction made by the BBM.

4.4.2 Predictive Power Evaluation Metric

After establishing Fidelity, we extend our evaluation to the predictive power of the explanation models, which tests the ability of the explanation model to predict the decision of the BBM on unseen samples. The more accurate the explanation model is, the higher the faithfulness. Intuition is that if the explanation model has encapsulated the decision-making logic of the BBM, it will be able to accurately predict the behaviour of the BBM [26]. The underlying assumption of this metric is that if an explanation leads to different decisions than those made by the model it explains, it is unfaithful [18]. This metric serves as a test for generalization and confirms the model's utility in practical scenarios.

Experimental Setup for Predictive Power Evaluation

The setup for predictive power evaluation involves using unsupervised clustering with our BBM over 90% of the data. We then provide the predictions of the BBM, along with intermediate results like cluster centres, to the explanation models. These models learn the association between documents and cluster predictions. We use the remaining 10% of the data to observe how the BBM and explanation models assign clusters to unseen documents. We report the agreement between them in terms of F1 score and accuracy.

Note on Metric Selection

In this study, we apply F1 and accuracy metrics, typically used in classification, to evaluate clustering explanations for both Fidelity and predictive power evaluation. Fidelity, assessed on the training dataset, measures the explanation model's internal consistency. This verification ensures that under known, controlled conditions, the explanation model can accurately replicate the prediction of the BBM. On the other hand, the predictive power evaluation metric, applied to unseen validation or test data, assesses the generalizability of the explanation model. It determines how effectively the explanation model captures and applies the BBM's underlying logic to new, varied scenarios. Together, these metrics offer a robust assessment of an explanation model's accuracy in mimicking the original model under familiar conditions and its ability to extend this mimicry accurately to novel situations.

4.5 Quantitative Results

We have detailed our approach for evaluating the Fidelity and predictive power of the explanation models in previous sections (4.4.1,4.4.2). These evaluations are conducted on previously observed and unseen data to ascertain scores for Fidelity and predictive power evaluation metrics, respectively. The results of these metrics are depicted in Table 4.1 and Table 4.2, and further statistical analysis is provided in the discussion section. These observed values tell us how faithful the explanation methods are with respect to the BBM, using predictive power evaluation and Fidelity, respectively.

Table 4.1: The Table presents a fidelity evaluation using accuracy and F1 metrics, where higher values indicate greater alignment of the explanation model with the black-box prediction over clustered data.

Model	AG News		DBpedia		R2		R5	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CDCEM	**96.29**$_{\pm 0.21}$	**0.96**$_{\pm 0.0}$	**94.93**$_{\pm 0.19}$	**0.95**$_{\pm 0.0}$	92.38$_{\pm 3.95}$	0.92$_{\pm 0.04}$	82.54$_{\pm 0.39}$	0.81$_{\pm 0.02}$
N.B.	91.14$_{\pm 0.18}$	0.91$_{\pm 0.0}$	86.34$_{\pm 0.71}$	0.86$_{\pm 0.01}$	87.04$_{\pm 0.17}$	0.87$_{\pm 0.0}$	85.78$_{\pm 0.53}$	0.83$_{\pm 0.01}$
D.T.	92.34$_{\pm 0.52}$	0.92$_{\pm 0.01}$	91.56$_{\pm 0.43}$	0.92$_{\pm 0.01}$	**95.15**$_{\pm 0.25}$	**0.95**$_{\pm 0.0}$	**91.04**$_{\pm 0.26}$	**0.9**$_{\pm 0.0}$
L.R.	93.78$_{\pm 0.45}$	0.94$_{\pm 0.0}$	91.37$_{\pm 0.18}$	0.92$_{\pm 0.0}$	94.37$_{\pm 0.13}$	0.94$_{\pm 0.0}$	88.26$_{\pm 0.47}$	0.86$_{\pm 0.0}$

The fidelity evaluation results demonstrate that different explanation models perform best in specific datasets as summarized from Table 4.1. Notable observations from fidelity results are listed below:

- **AG News and DBpedia**: The proposed explanation model achieves the highest scores in both accuracy and F1, recording a value of 0.96. This superior performance indicates an excellent alignment with the black-box predictions, particularly suitable for news-related and structured data contexts.

- **R2 and R5**: The D.T. model excels, with both accuracy and F1 scores reaching 0.95. Its effectiveness is apparent in datasets involving complex or hierarchical data structures, highlighting its suitability for such contexts.

- **General Performance**: L.R. models display consistent Fidelity but do not achieve the high performance of the proposed or D.T. models. Their scores range from 0.85 to 0.94 across all datasets with low variance. N.B. however displays low performance.

We observe significantly higher predictive power of CDCEM compared to the baselines as depicted in Table 4.2. These results demonstrate the robustness and effectiveness of the proposed model, consistently and significantly ($p < 0.05$) outperforming baseline methods across all examined datasets, thus reinforcing its utility

Table 4.2: Table illustrating enhanced predictive power of proposed explanation model compared to baselines over unseen samples.

Model	AG News		DBpedia		R2		R5	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CDCEM	$92.31_{\pm 0.24}$	$0.92_{\pm 0.00}$	$77.09_{\pm 0.62}$	$0.71_{\pm 0.03}$	$91.41_{\pm 4.39}$	$0.91_{\pm 0.04}$	$76.21_{\pm 2.19}$	$0.74_{\pm 0.01}$
N.B.	$67.69_{\pm 1.14}$	$0.68_{\pm 0.01}$	$64.32_{\pm 1.33}$	$0.62_{\pm 0.03}$	$78.16_{\pm 1.49}$	$0.78_{\pm 0.01}$	$70.53_{\pm 2.38}$	$0.66_{\pm 0.01}$
D.T.	$56.44_{\pm 5.18}$	$0.56_{\pm 0.04}$	$51.41_{\pm 1.66}$	$0.52_{\pm 0.03}$	$76.91_{\pm 3.31}$	$0.77_{\pm 0.03}$	$58.93_{\pm 1.67}$	$0.54_{\pm 0.01}$
L.R.	$68.75_{\pm 1.37}$	$0.68_{\pm 0.02}$	$63.09_{\pm 0.67}$	$0.64_{\pm 0.01}$	$82.14_{\pm 3.69}$	$0.82_{\pm 0.04}$	$66.08_{\pm 2.79}$	$0.63_{\pm 0.01}$

in providing faithful explanations. Furthermore, L.R. stands out as the most faithful baseline with consistent Fidelity among the options available. Therefore, we consider the proposed method and L.R. for our subsequent qualitative evaluation.

4.6 Qualitative Metrics and Evaluation through User Study

Unlike the quantitative section that explores how the explanation models interact with the BBM, this part evaluates how the user engages with the explanation model. We conducted a qualitative evaluation through a user study called "Visual System Evaluation for Explanation of Document Clustering". In this study, we present participants with explanations generated by the proposed method and a high-performing baseline, L.R., and collect their feedback via integrated questionnaires and tasks. The aim of this study is twofold: first, to compare the explanations from the baseline and proposed method to determine which one better satisfies users and effectively articulates the explanation; second, to evaluate the usefulness of the generated visualization tool.

We assess readability and plausibility by using questionnaires that explore several essential aspects of an explanation: understandability, satisfaction, detail sufficiency, completeness, usefulness, accuracy, and trustworthiness. The study includes specific questions and tasks to evaluate how well local and global explanations convey the rationale behind clustering and its decision-making processes.

The following section outlines the user study setup, including the design, tasks, and questionnaire. We detail the online setting, participant criteria, tools used, and feedback questionnaires, providing an overview of the study's methodology and evaluation process.

4.6.1 Study Setting

We conducted the study online using a visualization tool hosted on a server with participants accessing it via a link we provided.

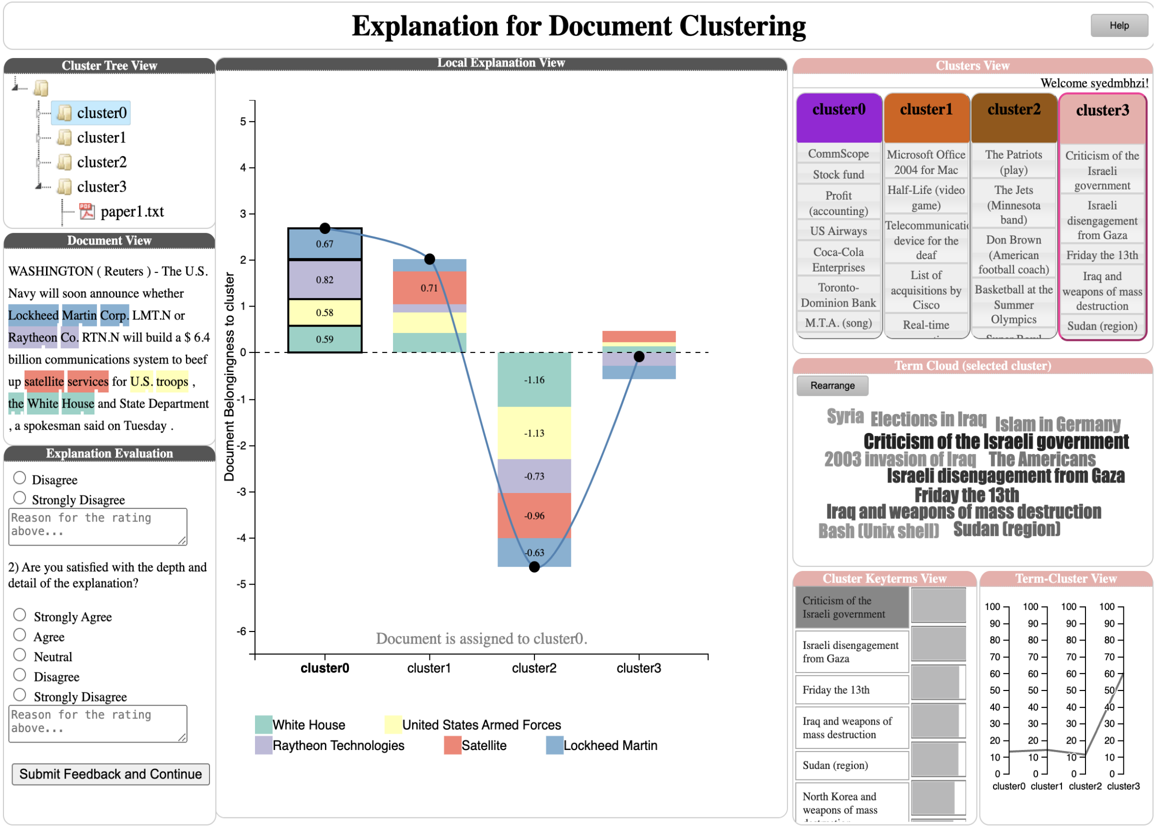
Participants Eighteen participants participated in this study, ensuring a balance between the feasibility and stability of the metrics. We divided the participants into two subsets: one experienced only the baseline explanations, and the other experienced only the proposed method's explanations, though we did not inform them of this distinction. We required participants to have access to a laptop, fluency in English, and knowledge of document clustering. Most participants were computer science graduate students selected for their familiarity with web interfaces and probable understanding of document clustering.

Tool Used We originally adapted the source code for the interactive document clustering tool by Sherkat et al. [41]. They developed the tool with Python CGI script and JavaScript hosted on our server, as illustrated in Fig. 4.2. Similar tools have been used for clustering interactively [35, 6, 42]. Since the developers of such tools did not design the tool for explanations, it did not include local explanations or global explanations aligned with its local counterpart.

4.6.2 Tasks and Procedure

Initially, we assigned participants the role of editors at an online news portal, presenting them with a simulated scenario depicted in Fig. 4.3. We then provided them with 20 randomly selected documents from the AG News dataset, one at a time, and required them to provide feedback on a) the logic of assigning documents to a cluster and b) their satisfaction with the depth and detail of the explanation. The setup encouraged participants to actively engage with and analyze the explanations, making the study more interactive.

At first, the tool only offered local explanations to mimic the typical document clustering scenario, where users know each document's cluster assignment but not the themes of each cluster. We designed this to explore global explanation's importance



Explanation for Document Clustering

WASHINGTON (Reuters) - The U.S. Navy will soon announce whether Lockheed Martin Corp. LMT.N or Raytheon Co. RTN.N will build a $ 6.4 billion communications system to beef up satellite services for U.S. troops , the White House and State Department , a spokesman said on Tuesday .

○ Disagree
○ Strongly Disagree

Reason for the rating above...

2) Are you satisfied with the depth and detail of the explanation?

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

Reason for the rating above...

Document is assigned to cluster0.

Welcome syedmbhzi!

Figure 4.2: The user study tool features various views, with each view's name in its header, to help users understand AI-driven clustering. The interface begins with the Cluster Tree View, organizing documents into clusters. Below it is the Document View, displaying one document at a time with highlighted text indicating annotated phrases. These highlighted phrases are used to construct the Local Explanation View in the middle. For each highlighted phrase in the Document View, you will find the corresponding concept in the legend at the bottom of the Local Explanation View, colour-coded to match. The Local Explanation View features a stacked bar graph representing the impact of each concept on the document's cluster assignment. On the right, views provide a global explanation for the selected cluster (cluster 3), marked by pink headers. First, the Cluster View offers an overview of the main ideas or topics within a cluster. Below, the Term Cloud View visualizes keywords associated with the selected cluster, with darker shades indicating higher importance. Further down, the Cluster Key Terms View and Term-Cluster View depict the importance of concepts within a cluster and across the clusters, respectively. Finally, The bottom-left view requests feedback on the current document explanation and moves to the next one.

Who are you?

Imagine stepping into the role of a lead editor at an online news portal. Your new task involves utilizing an innovative tool designed to organize news articles into distinct, coherent categories. This tool, powered by advanced AI, not only clusters articles but also provides explanations for their categorizations, highlighting the unique patterns of each cluster.

Figure 4.3: The simulation puts the participants in the shoes of a role that cares about why the explanations must be right. The positions that need to understand and evaluate the explanations generated for the AI system to build trust and reliability in the system.

and test the hypothesis that both local and global explanations are crucial for comprehensive document clustering explanations. After the first two questions, we showed participants global and local explanations and asked them to name each cluster. This step aimed to determine if global explanations effectively conveyed the cluster themes to the users. After the participants had named the clusters and provided feedback on each explanation for 20 documents, we presented them with the three questionnaires at the end of the user study. Therefore, we receive feedback from the users in three ways:

- **Embedded user study questionnaire**: In our study, participants were asked to respond to two questions for each explanation they reviewed. This design, which embeds questions within the study, allows us to capture real-time feedback from users immediately after they analyze an explanation, making the study more interactive and engaging. Participants rated their responses on a Likert scale from 1 to 5, with 1 indicating strong disagreement and 5 indicating strong agreement. The two questions posed to the participants are as follows:

  - Question 1: “Does the local explanation logically support the document’s assignment to `<cluster Label>`?”

  - Question 2: “Are you satisfied with the depth and detail of the explanation?”

- **Post user study questionnaires**: After interacting with our tool roughly for an hour, having analyzed 20 explanations of either baseline of the proposed explanation model depending on the subset, they are presented with three Likert scale questionnaires as stated below:

  - Questionnaire for Evaluating the Effectiveness of the Proposed Explainability: This questionnaire determines how well the visualization method provides explainability to users. Specifically, we asked whether each component of the local and global explanations was useful and accurate from the user's viewpoint. Such questions helped us demonstrate the need for detailed clustering explanations using both local and global insights. Additionally, we assessed the effectiveness of the overall software.
  - User Satisfaction Questionnaire: This questionnaire was proposed to assess several crucial aspects of explanations, including understandability, satisfaction, detail sufficiency, completeness, usefulness, accuracy, and trustworthiness (Hoffman et al., [16]). We used it to measure the readability (ease of understanding the explanations) and plausibility (how convincing the explanations are).
  - Software Usability Questionnaire: This questionnaire evaluates the usability of visualization systems, particularly how users interact with our explainable clustering system. It has been widely used in many user studies to assess the usability of tools and systems since it was proposed by John Brooke [5]. We employed this questionnaire to evaluate the software used in our user study for visualizing explanations in unsupervised clustering.

- **Feedback Comments:** Participants provided comments for their ratings through an embedded user study questionnaire and offered additional remarks at the end of the post-study questionnaires.

4.7 Qualitative Results

In this Section, we discuss the qualitative responses gathered from the participants of our user study. This detailed analysis provides insights into the subjective experiences and perceptions of users interacting with our system.

4.7.1 Embedded user study questionnaire

The results discussed in this section are derived from questionnaires embedded within the user study. These questions were presented to users for each document and its explanation.

Table 4.3: We presented users with 20 documents and their explanations for cluster assignment explanation. We asked users to rate logic and depth of explanation for each question on a Likert scale (1-5). 5 being strong agreement. The table presents the distribution of the user's mean responses across all the questions. t and p show the trend and significance of improvement.(*p<0.1)

	Question	Baseline	CDCEM	t	p
1	Does the local explanation logically support the document's assignment to a cluster?	4.02 ± 0.35	4.33 ± 0.22	2.16	0.06*
2	Are you satisfied with the depth and detail of the explanation?	4.05 ± 0.58	4.31 ± 0.3	1.08	0.31

The p-value for (Question 1 4.3) is 0.06, slightly above the 0.05 threshold. This results suggest that the proposed method may offer users a more logical and supportive explanation. However, the lack of strong statistical significance prevents us from definitively asserting the superiority of the proposed method based on this data alone.

Regarding satisfaction with the depth and detail of the explanation (Question 2 4.3) the data is inconclusive, indicating no significant difference between the baseline and proposed methods.

4.7.2 Post user study questionnaires

This Section is structured to systematically present findings from three different questionnaires. Each part begins with a description of the questionnaire's purpose, followed by the presentation of statistical analyses. The results are detailed in tables summarizing survey responses and accompanied by figures that visually illustrate the data. This consistent pattern ensures a clear and organized presentation of the post-study evaluation.

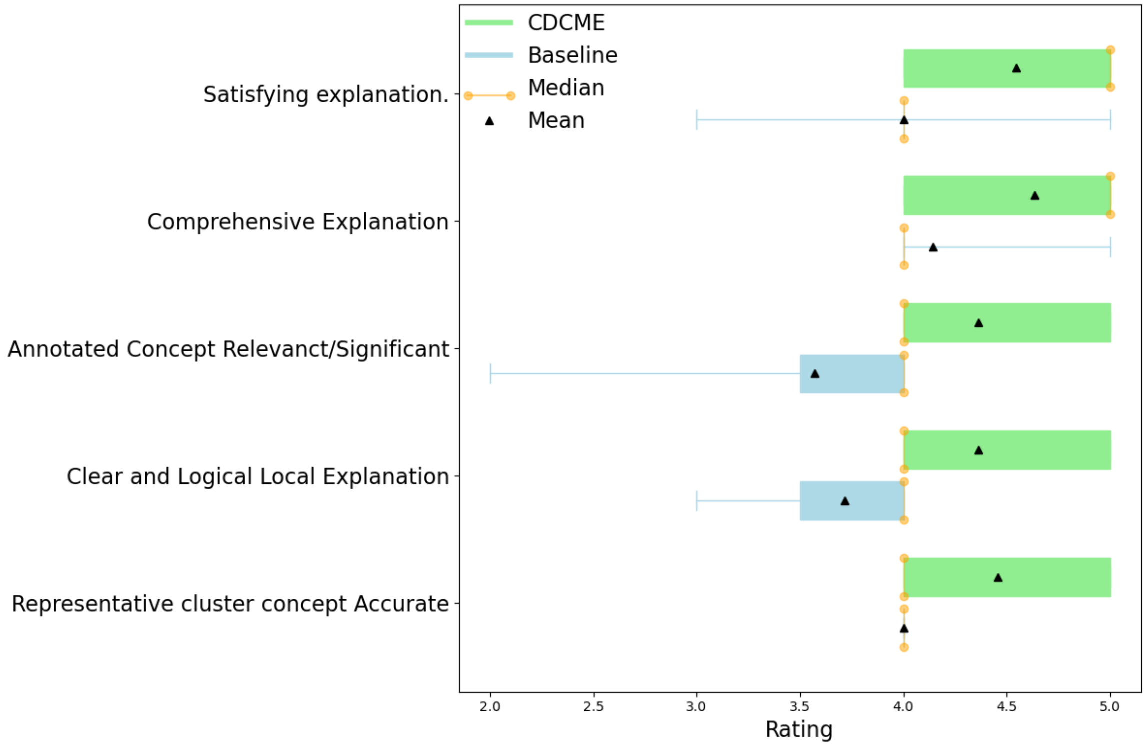


Figure 4.4: The bar chart shows the post-user study questions where the proposed method significantly outperformed the baseline method (L.R.). The proposed method received better responses for logical and global explanations, overall comprehensiveness, and satisfaction.

Questionnaire for Evaluating the Effectiveness of the Proposed Explainability.

After conducting a statistical hypothesis test on the responses from the effectiveness questionnaire, we observed significant improvements for the proposed explanation method (CDCEM) over the baseline method (L.R.) in several key areas. Notably, there was a significant improvement ($p < 0.05$) in the perceived accuracy of representative concepts describing the content of clusters (Q1, Fig 4.4). These representative concepts form the main component of the global explanations. Additionally, users found the logic and clarity in feature-based local explanations for cluster assignment significantly better (Q12, Fig 4.4). The annotated document concepts used to generate these local explanations were deemed relevant and significant for explaining clustering (Q13, Fig 4.4). Overall, users found that the explanations, composed of both local and global parts, were collectively effective and comprehensive for explaining the decisions involved in document clustering (Q13, Fig 4.4).

Moreover, in the questions like Q2, Q3, Q6 and Q8, enquire if the views were helpful. For such questions, both subsets of users agree, with the mean of the proposed method being higher; this means that the view was useful for both, but it was slightly more useful because the proposed method was accurate. These questions are not essentially comparative, and having high scores in both means the nature of explanations is deemed helpful and useful to the users.

Table 4.4: Summary of responses to the custom effectiveness questionnaire. We divided user study participants into two subsets: one presented with the baseline explanation and the other with the proposed explanation. The baseline and proposed columns show Likert scale responses. t and p values indicate trends and significance (** `p<0.05`).

	Question	Baseline	CDCEM	t	p
1	Cluster View: Did the key concepts presented as representative of each cluster accurately convey the underlying content of the documents within that cluster?	4.0 ± 0.0	4.45 ± 0.52	2.87	0.02**
2	Local Explanation: Did you find it useful to see the contribution of individual concepts within the documents when assessing their assignment to particular clusters?	4.29 ± 0.76	4.55 ± 0.52	0.79	0.45

	Question	Baseline	CDCEM	t	p
3	Local Explanation: The depiction of how much each concept of documents contributes to cluster assignment was helpful.	4.57 ± 0.53	4.36 ± 0.67	-0.74	0.47
4	Local Explanation: Do you believe that the importance attributed to the document's concepts in the cluster assignment accurately reflects their relevance?	4.0 ± 0.58	4.09 ± 0.3	0.38	0.71
5	Local Explanation: Does the local explanation offer a clear and logical basis for the document's placement in a cluster?	3.71 ± 0.49	4.36 ± 0.5	2.72	0.02**
6	Was the feature that allows you to select a concept and assess its significance across different clusters in the Term-Cluster View beneficial for your understanding?	4.0 ± 0.58	4.27 ± 0.79	0.83	0.42
7	Do you find the displayed probabilities accurately represent the association between the concepts and their respective clusters in the Term-Cluster View?	4.14 ± 0.69	4.18 ± 0.75	0.12	0.91
8	Cluster Keyterm View: Was the significance indicator for concepts in the Cluster Keyterms View helpful for understanding each cluster?	4.29 ± 0.95	4.18 ± 0.75	-0.26	0.80
9	Cluster Keyterm View: Do the bars in the Cluster Keyterms View accurately represent concept relevance within the clusters?	4.0 ± 0.58	4.36 ± 0.67	1.21	0.25

	Question	Baseline	CDCEM	t	p
10	Global Explanation - Right Side Panels: Do the right-side panels (Cluster view, Term Cloud, Cluster Keyterms View, Term Cluster view) provide vital information and context to understand document clustering?	4.43 ± 0.53	4.45 ± 0.69	0.07	0.95
11	Global Explanation - Right Side Panels: Do the right-side panels (Cluster view, Term Cloud, Cluster Keyterms View, Term Cluster view) offer accurate explanations for each cluster?	4.0 ± 0.58	4.45 ± 0.52	1.67	0.12
12	Document View: Are the emphasized terms in the explanation relevant and significant in clustering documents?	3.57 ± 0.79	4.36 ± 0.5	2.36	0.04**
13	All panels/views collectively and effectively provide a comprehensive explanation for document clustering, validating the allocation of documents to their respective clusters.	4.14 ± 0.38	4.64 ± 0.5	2.40	0.03**
14	The explanation for document clustering provides enough detail for you to construct a mental model of how the clustering algorithm works.	4.29 ± 0.49	4.64 ± 0.5	1.47	0.17

User Satisfaction Questionnaire

From the satisfaction questionnaire, the evidence is inconclusive regarding whether the baseline or the proposed explanation method is superior. With a p-value of 0.06, the proposed method appears to potentially offer a more satisfying explanation for users on how clustering works. The lack of statistical significance means we cannot definitively assert that the proposed method is better based on this data alone. However, the higher means suggest potential areas where the proposed method could enhance user understanding and satisfaction. These areas include comprehension of the clustering algorithm, satisfaction with the explanation, perceived detail and completeness of the explanation, and its usefulness to user's goals. Future studies with larger sample sizes or more sensitive measures might confirm these trends as significant improvements.

Table 4.5: Summary of responses to the User satisfaction questionnaire [16]. We divided user study participants into two subsets: one presented with the baseline explanation and the other with the proposed explanation. The baseline and proposed columns show Likert scale responses. t and p values indicate trends and significance (* p<0.1).

	Question	Baseline	CDCEM	t	p
1	From the explanation, I understand how the clustering algorithm works.	4.14 ± 0.69	4.36 ± 0.67	0.67	0.52
2	This explanation of how the clustering algorithm works is satisfying.	4.0 ± 0.58	4.55 ± 0.52	2.04	0.06*
3	This explanation of how the clustering algorithm works has sufficient detail.	4.0 ± 0.82	4.55 ± 0.52	1.58	0.15
4	This explanation of how the clustering algorithm works seems complete.	3.86 ± 0.69	4.09 ± 0.54	0.75	0.47
5	This explanation of how the clustering algorithm works tells me how to use it.	4.14 ± 1.07	4.27 ± 0.65	0.29	0.78
6	This explanation of how the clustering algorithm works is useful to my goals.	4.29 ± 0.76	4.45 ± 0.69	0.45	0.66
7	This explanation of the clustering algorithm shows me how accurate the clustering algorithm is.	4.14 ± 1.07	4.18 ± 0.6	0.09	0.93
8	This explanation lets me judge when I should trust and not trust the clustering algorithm.	4.43 ± 0.53	4.27 ± 0.9	-0.47	0.64

Software Usability Questionnaire

Unlike the previous two questionnaires, this one focuses on the responses of the entire user set (from both the proposed and baseline groups) to evaluate the tool we developed to visualize feature-based explanations. We computed the SUS (System Usability Scale) score for this questionnaire as described by its author [5]. The System Usability Scale (SUS) score is computed based on the responses to a ten-item Likert scale questionnaire. We assign a score to each statement on a scale from one to five, where one represents strong disagreement, and five indicates strong agreement by the user. We substract the scores for odd-numbered questions by 1, while subtract those for even-numbered questions from 5. The adjusted scores are then summed and multiplied by value 2.5 to obtain the final SUS score, which ranges from 0 to 100.

we get an average score of 74.31 with a standard deviation of 11.50, as shown in Fig. 4.5. This places the score in the 70-79 percentile range according to the grading scale interpretation of the SUS score by Soura and Lewis. The individual scores for the baseline and proposed methods were 73.21 (SD = 5.72) and 75.0 (SD = 14.27), respectively. Statistical analysis revealed no significant difference between the two groups.

In addition to the overall user score, we analyzed the rating distribution for each question across all users to identify potential areas for improvement, as illustrated in table 4.6. The analysis highlighted the following strengths and weaknesses:

Strengths: The system scored high on questions about frequency of use, ease of use, integration of functions, and user confidence. These results indicate that users generally find the system easy to use and well-integrated.

Weaknesses: We observe scores in questions concerning unnecessary complexity, the need for technical support, inconsistency, cumbersomeness, and learning requirements. These areas require further improvement to enhance the overall usability of the system.

Table 4.6: Summary of responses to the Software Usability Scale (SUS) Questionnaire [5]. The Table presents the mean responses by the users per question. Analyzing these values allows us to understand the strengths and weaknesses of our explanation visualization tool.

	Question	Mean Score	STD Score
1	I think that I would like to use this system frequently.	4.00	0.77
2	I found the system unnecessarily complex.	2.11	0.83
3	I thought the system was easy to use.	4.00	0.49
4	I think that I would need the support of a technical person to be able to use this system.	1.89	0.90
5	I found the various functions in this system were well integrated.	4.39	0.61
6	I thought there was too much inconsistency in this system.	2.00	0.91
7	I would imagine that most people would learn to use this system very quickly.	3.83	0.71
8	I found the system very cumbersome to use.	2.22	0.88
9	I felt very confident using the system.	4.00	0.59
10	I needed to learn a lot of things before I could get going with this system.	2.28	1.02

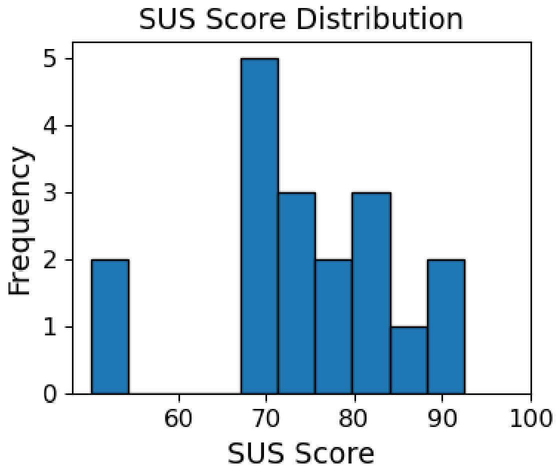


Figure 4.5: The histogram shows the Likert score received for the Software Usability Questionnaire. A Likert score above 68 is considered acceptable for software.

4.7.3 Feedback Comments:

This Section presents a detailed analysis of user feedback regarding their experiences and perceptions of our system. By examining their comments, we aim to identify specific areas for improvement and validate the effectiveness of our proposed methodologies. The summarized feedback highlights key concerns and positive observations across various aspects of our system, which are categorized as follows:

- Interface: The low score for the software usability questionnaire stems from concerns about the user interface (U.I.). Despite recognizing that user interface improvements may not be the primary objective of the research, several users have expressed a desire for a better U.I. One user suggested that the software should provide a more detailed explanation wherever the document partially belongs to multiple clusters. In summary, users have consistently noted that the U.I. could be improved to enhance their overall experience.

- Wikifer: We use Wikifer annotator to generate explanations using meaningful document concepts instead of document words to prevent the inclusion of stop words and focus on meaningful concepts, enhancing the relevance and significance of the highlighted concepts. Users responded positively to the effectiveness questionnaire on the annotated concepts (Q12). However, feedback indicated that the annotation could be improved. Users felt that it sometimes failed to capture all critical keywords from the documents, resulting in incomplete explanations due to missed relevant terms. As one user noted, “It sometimes lacks in some areas where it fails to pick up some important keywords from documents.” This feedback suggests enhancing the concept extraction process to ensure more comprehensive explanations.

- Global and local explanation: Initially, users were given only local explanations; they guessed the clusters until global explanations were provided. Some comments were, “The document seems to be related to geopolitics.” and “The document is regarding terrorism and the U.K.” After revealing it to them, they appreciated the explanations even more: “It got so much easier when you showed the topics of the cluster.” These comments suggest that global explanations are essential to comprehending local explanations and, in turn, understanding the clustering. It was also observed from the effectiveness questionnaire responses (Q1)

- Comparison of baseline and proposed explanation method:

    1. User Satisfaction and Trust: It is strange because users do not trust the baseline, as they disagree even when the baseline explanation is similar to the proposed method explanation. The user found the information sufficient and accurate in the proposed method. Furthermore, if users are unsatisfied with the explanation, they may ask for more details. If they agree, they will find the information enough. For example, both baseline and proposed have the same features. The difference is feature contribution in the explanations. Despite the same features, the baseline feels incomplete, whereas the proposed do.

    2. Handling Nuances and Subjective Perspectives: The proposed method handled nuanced content and subjective perspectives better. When a document partially belonged to one category but more to another, users appreciated the proposed method’s nuanced explanation. As a result, users were more satisfied with the explanation of the proposed method. One user stated: “I was hesitating between C0(cluster 0) and C1(cluster 1) because of the words satellite and service. I liked that model also caught that”.

3. Contextual Inaccuracy: The baseline method, relying on non-contextual embeddings, often fails to capture the correct context of terms. This failure was evident when the explanation only highlighted terms corresponding to one cluster but did not pick up concepts for another cluster, leading to poorer performance than the proposed method.

4. Engagement and Detailed Feedback: Participants using the proposed method were more engaged and provided more detailed feedback. We observed that the users presented with the proposed method seemed more involved as they provided more comments than those from the baseline subset.

Chapter 5

Results and Discussion

In this Section, we present and discuss the results of our experiments, evaluating the performance of various explanation models in terms of fidelity and predictive power. Additionally, we provide insights from a user study to assess the practical effectiveness and user satisfaction with our proposed explanation models.

5.1 Quantitative Analysis

In fidelity evaluations comparing explanation models to black-box predictions, measured by accuracy and F1 scores, distinct performance patterns emerged across different methods and datasets. The Proposed CDCEM method excelled particularly in A.G. News and R5, achieving top metrics (0.96 accuracy, 0.95 F1 in both). Conversely, the D.T. method outperformed others in DBpedia and R2, with leading scores (0.95 accuracy, 0.90 F1).

The performance of N.B. was less robust, peaking at 0.91 for both accuracy and F1 on A.G. News. At the same time, Logistic Regression displayed steady results across all datasets, notably scoring 0.94 in both metrics for DBpedia, with minimal variance (±0.01).

The superior performance of the proposed model aligns with our initial claim that utilizing both intermediate and final outcomes enhances the faithfulness of explanations. We developed the proposed explanation method to closely align with a specific architecture that uses centroid-based clustering, making the explanation more faithful. Furthermore, the document representations used within the system must also be accessible. Due to this, in pursuit of faithfulness, we sacrifice the proposed method's ability to apply to a wide range of clustering methods.

In the analysis of predictive power, where the performance of BBMs was evaluated against an explanation model over unseen data using accuracy and F1 scores, the CDCEM model demonstrated statistically superior faithfulness across all datasets with a p-value below 0.01%. Improvements in accuracy ranged from 8% to 34% and in F1 scores from 11% to 35% over the next best results achieved by Logistic Regression, which was the most robust baseline.

Specifically, in the A.G. News dataset, the CDCEM achieved a 35% and 34% improvement in both accuracy and F1 scores compared to the scores obtained by Logistic Regression. In the DBpedia dataset, CDCEM outperformed Logistic Regression by 11% in accuracy and 22% in F1. For the R2 dataset, we noted improvements of 11% in both metrics over Logistic Regression. In the R5 dataset, while the enhancements were modest, the CDCEM still showed notable increases of 11% in accuracy and 8% in F1 score over N.B.

These findings support our research question regarding the development of explanation models that reflect the accurate reasoning process of the BBM. This empirical evidence suggests that baseline explanation models perform well on fidelity over seen data but have poor predictive power over unseen data. This discrepancy indicates that these models may be overfitting the data, capturing patterns that appear convincing but do not generalize well to unseen data. Rather than accurately interpreting and explaining the BBM's reasoning, they create their own rationale for categorizing documents. Our proposed models address this issue by not solely focusing on mimicking the BBM's predictions but also aligning the explanation process with the BBM by using BBM's intermediate results. This approach leads to our superior performance in meeting faithfulness criteria.

5.2 User Study

In the user study, we received subjective feedback on the effectiveness of global and local explanations, user satisfaction, and software usability from the questionnaires administered.

The explanations generated were praised for providing clear and logical local insights and accurate global explanations with representative cluster concepts based on relevant and significant terms. These components collectively offered effective and comprehensive explanations. The distribution of responses is shown in Fig. 4.4 However, specific elements of the global explanations, namely the term cluster view and cluster key term view, were neither helpful nor accurate, adding unnecessary complexity to the system.

User satisfaction with the proposed explanations was generally higher. This satisfaction was gauged in terms of accuracy, trust, usefulness, completeness, level of detail and satisfaction. The response to the satisfaction question showed the most improvement, with a p-value of 0.06, which is slightly above the conventional 0.05 threshold. Despite this, there was no statistical improvement in other areas, indicating that these aspects require further attention. Particularly, the completeness of the explanations can be improved by working on recall of the relevant feature or concept extraction. Furthermore, this explanation method might not be suitable for recent research papers that usually contain novel concepts yet to be introduced as Wikipedia concepts.

Regarding software usability, feedback indicated that the system was consistent and well-integrated. Users could confidently use it frequently, and most found it easy without technical assistance. However, many users felt that the system needed to be quicker to learn, requiring them to understand many aspects before effective use.

Summarizing the findings from the quantitative analysis and user study reveals the effectiveness of the Proposed method, CDCEM. Quantitatively, it achieves superior accuracy and F1 scores, demonstrating its robustness. Qualitatively, users praised the explanations for their clear and logical insights and found the system consistent and well-integrated. This alignment of solid performance metrics with positive user feedback reinforces the method's overall validity and practical applicability, showing that it delivers reliable and user-approved explanations. Although we cannot broadly generalize these insights due to the small number of participants, these results give preliminary insights encouraging enough to perform a more extensive user study. We further need to address the system's learnability and enhance completeness by improving the feature extraction process and user experience and satisfaction.

Chapter 6

Conclusion and Future Work

The primary outcome of our research is the development of the Conceptual Document Clustering Explanation Model (CDCEM). We learn the explanation model using documents as inputs and treat the final cluster predictions of the black-box model as the ground truth. This model also utilizes the black-box model's intermediate results, such as embeddings and k-means, to align the explanation generation process with that of the black-box model. This alignment aims to ensure that the explanations accurately reflect the clustering process of the black box, i.e., be faithful. The faithfulness is checked by the intuition that if the explanation model has learned the black box model's decision mechanism, the explanation-based prediction will match the prediction of the black box model on unseen data. Therefore, we compare the cluster assignment of the black box model on unseen data with the prediction made based on the explanation model using metrics like F1 and accuracy. We found that CDCEM maintains faithfulness to the underlying black-box model by aligning the explanation generation closely with the prediction process of the black box. This faithfulness is evidenced by significant improvements in accuracy, ranging from 8% to 34%, and F1 scores, from 11% to 35%, over four baseline models across four datasets. Furthermore, a user study with eighteen computer science students comparing CDCEM and logistic regression (the next best baseline) reveals that proposed explanations generate significantly clearer and more logical local explanations with accurate global explanations. Therefore, this new faithful method generates reliable, comprehensive, and user-approved explanations for document clustering. We designed CDCEM to align closely with specific A.I. architectures, enhancing its faithfulness compared to model-agnostic baselines, though its applicability is limited to specific architectures.

While we showed improvements in faithfulness, perceived accuracy, clarity, logic, and comprehensibility, further research can enhance its performance in several areas. Rapid advancements in large language models (LLMs) suggest that using even more recent LLMs could offer superior clustering performance than the one we explained. We aim to adapt and apply our explanation model to work effectively with these more contemporary clustering approaches.

To improve the completeness of our explanations, we need to incorporate more concepts or features in explaining the decisions of the black box model. Users indicated that additional document features should be used to explain cluster assignments, suggesting a need for higher recall in feature extraction. While maintaining current precision, we should aim for annotations with higher recall. Using a domain-specific knowledge graph-backed annotator instead of general-purpose tools like Wikifier can provide more accurate and contextually relevant annotations, enhancing both recall and precision and thereby leading to more complete and reliable explanations.

Another promising direction involves refining the selection of concepts used in our model. Currently, our model considers all identified concepts within the text. Future iterations could benefit from a mechanism that prioritizes concepts based on their significance, effectively capturing the essence of the document more succinctly and meaningfully. We encountered documents with concepts about one category (like sports), but the underlying essence was about another category (like finance) without prominent features representing the latter. Addressing this issue could enhance the alignment of features with the true content of the documents, leading to more accurate explanations.

Finally, refining the user interface can improve the readability and user satisfaction of the explanations. Several users pointed out the need for more detailed explanations, while others requested that the complexity and overwhelming effect of the current layout be reduced. A potential solution could be a more dynamic and responsive design that adjusts to user preferences. This design could include customizable views that allow toggling between detailed and simplified explanations. Along with enhancing readability, these improvements would also improve the overall usability of the software.

Therefore, in future work, we intend to work on the comprehensiveness, readability, and user satisfaction of the explanations and applications of more contemporary black box models and evaluate them with other metrics and extensive user studies with more participants. Such work would extend our model's capabilities and improve the explainability of document clustering models.

Bibliography

[1] Omar M Ayad, Abd-El Fattah Hegazy, and Ahmed Dahroug. A proposed model for loan approval prediction using xai. *Nile Journal of Communication and Computer Science*, 6(1):1–11, 2023.

[2] Szymon Bobek, Michal Kuk, Maciej Szelek, and Grzegorz J Nalepa. Enhancing cluster analysis with explainable ai and multidimensional cluster prototypes. *IEEE Access*, 10:101556–101574, 2022.

[3] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*, 472, 2017.

[4] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees.* CRC press, 1984.

[5] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[6] Eric M Cabral, Sima Rezaeipourfarsangi, Maria Cristina F Oliveira, Evangelos E Milios, and Rosane Minghim. Addressing the gap between current language models and key-term-based clustering. In *Proceedings of the ACM Symposium on Document Engineering 2023*, pages 1–10, 2023.

[7] Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

[8] Eoghan Cunningham and Derek Greene. Surrogate explanations for role discovery on graphs. *Applied Network Science*, 8(1):28, 2023.

[9] Finale Doshi-Velez and Been Kim. Considerations for evaluation and generalization in interpretable machine learning. *Explainable and interpretable models in computer vision and machine learning*, pages 3–17, 2018.

[10] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[11] Renchu Guan, Hao Zhang, Yanchun Liang, Fausto Giunchiglia, Lan Huang, and Xiaoyue Feng. Deep feature-based text clustering and its explanation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3871–3872. IEEE Computer Society, 2023.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[14] Pratiyush Guleria, Parvathaneni Naga Srinivasu, Shakeel Ahmed, Naif Almusallam, and Fawaz Khaled Alarfaj. Xai framework for cardiovascular disease prediction using classification techniques. *Electronics*, 11(24):4086, 2022.

[15] Sadeq Heydarbakian and Mehran Spehri. Interpretable machine learning to improve supply chain resilience, an industry 4.0 recipe. *IFAC-PapersOnLine*, 55(10):2834–2839, 2022.

[16] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[17] David W Hosmer Jr and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2 edition, 2004.

[18] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

[19] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics, 2020.

[20] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[21] Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[22] Freddy Lecue. On the role of knowledge graphs in explainable ai. *Semantic Web*, 11(1):41–51, 2020.

[23] Le Li, Jianjun Yang, Yang Xu, Zhen Qin, and Honggang Zhang. Documents clustering based on max-correntropy nonnegative matrix factorization. In *2014 International Conference on Machine Learning and Cybernetics, Lanzhou, China, July 13-16, 2014*, pages 850–855. IEEE, 2014.

[24] Yutong Li, Juanjuan Cai, and Jingling Wang. A text document clustering method based on weighted bert model. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 1426–1430, 2020.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[26] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–70, 2024.

[27] Vivek Mehta, Mohit Agarwal, and Rohit Kumar Kaliyar. A comprehensive and analytical review of text clustering techniques. *International Journal of Data Science and Analytics*, pages 1–20, 2024.

[28] Christoph Molnar. *Interpretable Machine Learning.* Self-Published, 2020.

[29] Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, 2021.

[30] Enea Parimbelli, Tommaso Mario Buonocore, Giovanna Nicora, Wojtek Michalowski, Szymon Wilk, and Riccardo Bellazzi. Why did ai get this one wrong?—tree-based explanations of machine learning model predictions. *Artificial Intelligence in Medicine*, 135:102471, 2023.

[31] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[32] Brett Poulin. Visual explanation of evidence in additive classifiers. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence—Volume 2*, n.d. Accessed: October 14, 2023.

[33] Rafael Poyiadzi, Xavier Renard, Thibault Laugel, Raul Santos-Rodriguez, and Marcin Detyniecki. Understanding surrogate explanations: the interplay between complexity, fidelity and coverage. *arXiv preprint arXiv:2107.04309*, 2021.

[34] Jayasree Ravi and Sushil Kulkarni. Text embedding techniques for efficient clustering of twitter data. *Evolutionary Intelligence*, 16(5):1667–1677, 2023.

[35] Sima Rezaeipourfarsangi, Ningyuan Pei, Ehsan Sherkat, and Evangelos Milios. Interactive clustering and high-recall information retrieval using language models. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, pages 1–5, 2022.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[37] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pages 45–50, 2021.

[38] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[39] J. Rožanec, E. Trajkova, I. Novalija, P. Zajec, K. Kenda, B. Fortuna, and D. Mladenić. Enriching artificial intelligence explanations with knowledge fragments. *Future Internet*, 14(5):Article 5, 2022.

[40] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE access*, 9:135392–135422, 2021.

[41] Ehsan Sherkat, Evangelos E Milios, and Rosane Minghim. A visual analytics approach for interactive document clustering. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(1):6, 2019.

[42] Ehsan Sherkat, Seyednaser Nourashrafeddin, Evangelos E Milios, and Rosane Minghim. Interactive document clustering revisited: a visual analytics approach. In *23rd International Conference on Intelligent User Interfaces*, pages 281–292, 2018.

[43] T. Shi, X. Zhang, P. Wang, and C. K. Reddy. Corpus-level and concept-based explanations for interpretable document classification. *ACM Transactions on Knowledge Discovery from Data*, 16(3):48:1–48:17, 2021.

[44] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 519–528. Springer, 2021.

[45] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 519–528. Springer, 2021.

[46] Alvin Subakti, Hendri Murfi, and Nora Hariadi. The performance of bert as data representation of text clustering. *Journal of big Data*, 9(1):15, 2022.

[47] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[48] Ka Yee Yeung and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinform.*, 17(9):763–774, 2001.

[49] Jianhua Yin and Jianyong Wang. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636. IEEE, 2016.

[50] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[51] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.