ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 13 (2024), e31448
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.31448

Hybrid Text Embedding and Evolutionary Algorithm Approach for Topic Clustering in Online Discussion Forums

Ibrahim Bouabdallaoui, Fatima Guerouate and Mohammed Sbihi

LASTIMI Laboratory EST Salé, Mohammed V University in Rabat, Avenue Prince Héritier, Salé, Morocco
ibrahim_bouabdallaoui@um5.ac.ma, fatima.guerouate@est.um5.ac.ma, mohammed.sbihi@est.um5.ac.ma

KEYWORDS

LDA; BERT; K-Means; Genetic Algorithms; Forum Analysis

ABSTRACT

Leveraging discussion forums as a medium for information exchange has led to a surge in data, making topic clustering in these platforms essential for understanding user interests, preferences, and concerns. This study introduces an innovative methodology for topic clustering by combining text embedding techniques—Latent Dirichlet Allocation (LDA) and BERT—trained on a singular autoencoder. Additionally, it proposes an amalgamation of K-Means and Genetic Algorithms for clustering topics within triadic discussion forum threads. The proposed technique begins with a preprocessing stage to clean and tokenize textual data, which is then transformed into a vector representation using the hybrid text embedding method. Subsequently, the K-Means algorithm clusters these vectorized data points, and Genetic Algorithms optimize the parameters of the K-Means clustering. We assess the efficacy of our approach by computing cosine similarities between topics and comparing performance against coherence and graph visualization. The results confirm that the hybrid text embedding methodology, coupled with evolutionary algorithms, enhances the quality of topic clustering across various discussion forum themes. This investigation contributes significantly to the development of effective methods for clustering discussion forums, with potential applications in diverse domains, including social media analysis, online education, and customer response analysis.

1. Introduction

Discussion forums serve as virtual platforms facilitating the exchange of ideas and communication amongst users with shared interests or concerns. Predominantly, such platforms involve communities of individuals interacting through message postings on a website. The structural organization of a forum is typically bifurcated into several categories and subcategories, each corresponding to a specific theme or topic. Each category houses individual threads or topics, initiated by a starter post (or 'Original Post') and followed by replies from other users. Threads can span over extended periods, often years, allowing for open-ended, dynamic dialogues. Depending on the forum rules, users might be required to register and create an account before participating, while others permit anonymous postings. Registered users often maintain profiles detailing personal interests, areas of expertise, geographical location, and more.

Discussion forums hold vast potential and provide numerous opportunities:

- Knowledge Exchange: Forums foster knowledge sharing and exchange of expertise, enhancing understanding of a given topic or issue.
- Community Building: Forums foster the formation of communities with shared interests or experiences, encouraging mutual support and engagement.
- Problem Solving: Collaborative problem solving and issue resolution are facilitated through forum discussions.
- Idea Generation: The interactive dialogue in forums often sparks creativity, leading to innovative ideas and perspectives.
- Peer Learning: The diversity of forum participants offers opportunities to learn from others' experiences and perspectives, thereby broadening one's understanding.
- Feedback Mechanism: Forums can serve as platforms for providing feedback on products, services, or ideas, thereby improving the quality of offerings and providing insights.

The complex and flexible structure of platforms such as Facebook Groups and Linkedin Groups has significantly leveraged these potentials, thereby generating more threads across a broad spectrum of domains and subdomains. This explosion of threads and posts, however, leads to a massive, unstructured data pool, presenting challenges in topic identification from discussion forums, thereby affecting the accuracy and utility of the results. Key challenges include:

- Language Ambiguity and Variability: Discussion forums are known for their diverse range of topics and subjects. Participants often use informal or colloquial language, and there is a high degree of variability in the way people express themselves. This variability can lead to language ambiguity, where the same word or phrase may have different meanings in different contexts. For example, the word "bank" can refer to a financial institution, a river bank, or the action of tilting to the side. In the context of discussion forums, such language ambiguity can make it challenging for natural language processing (NLP) techniques to accurately understand and categorize the content.
- Contextual Deficiency: Discussion forum posts often lack the rich context that is typically present in more formal or structured texts. Participants may assume that others are familiar with the background of a discussion, and as a result, they may omit important details or references. This contextual deficiency can lead to misinterpretation of the content. For instance, a post mentioning "the new update" may be clear to forum members who are aware of recent developments,

but it could be confusing to those without that context. NLP algorithms need to navigate this lack of context to accurately determine the topic or intent of a post.

- Noise and Irrelevant Content: Discussion forums can contain a significant amount of noise, which refers to content that is off-topic, spammy, or irrelevant to the main discussion. Participants may engage in off-topic conversations, share unrelated links, or use humor and sarcasm, which can further obscure the identification of meaningful topics. The presence of noise and irrelevant content poses a challenge for algorithms aiming to extract relevant and meaningful topics from forum discussions.
- Dynamic Forum Nature: Discussion forums are dynamic by nature, with threads evolving over time as new posts are added and discussions progress. This dynamic nature means that the topics being discussed can change rapidly, and the language used in the forum can evolve accordingly. NLP models need to adapt to these changes and provide up-to-date topic identification. The challenge lies in keeping topic modeling results current and relevant as discussions continue to evolve.
- Scarcity of Labeled Data: Supervised machine learning techniques often rely on labeled data for training. However, obtaining labeled data from discussion forums can be challenging and time-consuming. Labeling content typically requires human annotators to categorize posts, and the sheer volume of forum data makes this a daunting task. As a result, there may be a scarcity of labeled data available for training models, limiting the effectiveness of supervised learning approaches for topic modeling.

Consequently, accurately identifying related topics from a multitude of threads remains a persistent challenge in information retrieval due to the semantic complexities involved (Adams et al., 2008). This paper proposes a novel approach to address this challenge by clustering relevant discussion forum topics. The proposed method leverages the power of Genetic Algorithms and K-Means clustering applied on Latent Dirichlet Allocation autoencoders.

2. Related Work

As online forums continue to grow in popularity, researchers have turned to topic modeling to better understand the conversations and topics being discussed within these communities. In this related work section, we review several scientific papers that have explored the application of topic modeling in forum discussions. By examining the insights and findings from these previous studies, we can gain a better understanding of how topic modeling can be used to analyze online forum discussions and identify potential research areas.

2.1. Embedding and Clustering Topics

Topic models serve as potent instruments for scrutinizing document collections and unearthing inherent themes. Traditional methodologies rely on probabilistic topic models, premised on a generative story. An alternate proposition involves garnering topics via clustering of pre-trained word embeddings, integrating document information for weighted clustering, and re-ranking top words (Sia et al., 2020). An ensuing benchmark examination tested the efficacy of various combinations of word embeddings and clustering algorithms, analyzing their performance with dimensionality reduction in mind. Empirical evidence indicates that pre-trained word embeddings—both contextualized and non-contextualized—alongside tf-weighted K-Means and tf-based reranking, present a viable substitute to conventional topic modeling. Notably, this alternative method promises reduced complexity and execution time.

A novel methodology for scrutinizing social media data combines the K-Means algorithm with the Genetic algorithm and Optimized Cluster Distance method to create clusters within social media communities (Alsayat et al., 2016). This is based on parameters such as leadership and follower scores, as well as attitudes. The suggested algorithm, empirically evaluated, exhibits superior performance compared to its predecessors. An applied case of this method further illustrates user communities via self-explanatory labels assigned to each cluster. Evaluation using multiple cluster validation metrics indicates superior clustering outcomes compared to existing methodologies, thereby introducing a unique use-case of user community grouping based on their activities. This optimized and scalable approach holds potential for real-time clustering of social media data.

2.2. Clustering Topics in Discussion Forums

The principal query regarding the application of a clustering task on forum discussions revolves around the effective clustering of forum threads. It also considers the advantages of automated thread clustering and its efficiency enhancement implications for users and forum administrators alike. One proposed three distinct methods for post weight assignment within a forum thread, aiming to ensure a more precise thread representation (Pattabiraman et al., 2013). This study utilized data gathered from three distinct Linux forums to evaluate both within-forum and across-forum clustering. Although the state-of-the-art methods displayed competent performance, the results showed further improvements could be achieved by leveraging thread structures. Specifically, a parabolic weighting method that assigns higher weights to the beginning and end posts of a thread demonstrated consistent outperformance over a standard clustering method. The study aimed to tackle thread clustering issues using three cutting-edge methods typically utilized for text-clustering problems. The dual objectives included determining the effectiveness of standard clustering algorithms in this new context and identifying the best performing method, as well as establishing a robust baseline method that can be further developed to harness forum structures and augment thread clustering. In contrast to typical document clustering where text is unstructured and represented as a “bag of words,” this approach aims to consider the unique characteristics of forum threads.

In another context, an applied study sought to understand student learning experiences more profoundly by proposing the use of clustering models and interactive visualizations for qualitative analysis of graduate discussion forums (Gokarn Nitin et al., 2019). This approach was predicated on the belief that discussion forums contain valuable content that could be harnessed to construct a knowledge repository. By employing text mining techniques, the study aimed to identify sub-topics and topic evolutions within discussion forums, generate insights into topic analysis, and uncover students’ cognitive understanding within and beyond classroom learning settings. The proposed analysis model was developed and tested on a graduate course in Information Systems. The findings revealed that the proposed techniques were successful in extracting knowledge from forums and generating user-friendly visualizations. This methodology could assist faculty members in analyzing student discussions, understanding their strengths and weaknesses, and enhancing their cognitive knowledge of course topics.

2.3. Comparative Evaluation of Machine Learning Models for Topic Clustering

The table below (Table 1) provides a comparative evaluation of different machine learning models used for topic clustering:

Table 1. Comparative evaluation of Machine Learning models for Topic Clustering

Model	**Description**	**Strengths**	**Weaknesses**	**Comparison with LDA-BERT**
TF-IDF with K-Means (Bouabdallaoui et al., 2023)	Transforms text into a vector space, categorizing documents based on word frequency patterns.	Simple and efficient for large datasets.	Struggles with synonymy and polysemy; does not capture word order or context.	LDA-BERT captures contextual nuances and semantic relationships better, leading to more meaningful clusters.
Word2Vec with Hierachical Clustering (Wang, Bo, et al., 2017)	Converts words into vectors capturing semantic relationships, and creates a tree of clusters.	Captures semantic relationships and forms a topic hierarchy.	Computationally expensive; sensitive to distance metric choice.	LDA-BERT offers a more efficient balance between computational cost and clustering quality.
Doc2Vec with DBSCAN (Eklund, A., 2023)	Extends Word2Vec to document level and groups vectors based on density.	Effective for clusters of varying shapes and sizes.	Sensitive to parameter settings; may struggle with high-dimensional data.	LDA-BERT effectively reduces dimensionality while preserving semantic meaning, ensuring robust performance in high-dimensional text data.

Haut du formulaire
Bas du formulaire

3. Methodology

The methodology proposed in this paper is based on a pipeline of text processing techniques for topic generation purposes, as well as an improved clustering method using the combination of K-Means and genetic algorithms to define the main topics of a post. A similarity matrix is created based on the generated topics and K-Means centroids, to define relations between posts of a forum. This approach is benchmarked on 3 online web forums: TripAdvisor NYC, Ubuntu Forums and Photography-on-the.net forums (Obasa et al., 2016).

3.1. Dataset and Preprocessing

Three distinct online platforms were selected for this study: a travel-focused forum (TripAdvisor NYC), a technology-related forum (UbuntuForums), and a photography-oriented forum (Photography-on-the.net). Each platform facilitates user interaction and discussion on their respective areas of interest through numerous threads and posts. Active participation in these discussions generally requires user account creation.

TripAdvisor NYC host's discussions related to travel experiences and tips in New York city. UbuntuForums provides a space for users to discuss and troubleshoot issues related to a certain operating system. Photography-on-the.net serves as a creative space for individuals to share their work, discuss photography techniques, and solicit advice from the community.

We deployed scraping scripts designed to follow the structure and features of each website. A common feature set, including thread title, username, user location, post content, and postdate, was identified across the three forums. This information was used to define a normalized data structure that could be used to gather outputs from the scraping scripts.

The data collection phase resulted in the extraction of a vast number of posts and threads from each platform, covering various date ranges. This data is utilized for further analysis and experiments within our study.

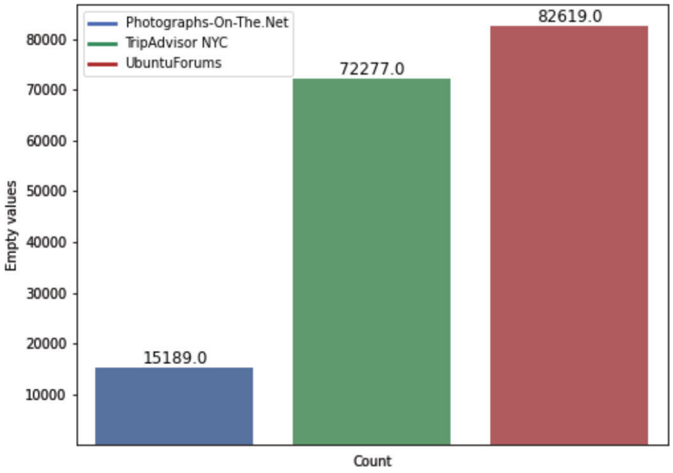

Figure 1. Empty values in the scrapped datasets

Fig. 1. provides an overview of the data sets under examination, highlighting several thousand null entries. These empty entries indicate threads that haven't received any responses from forum members, thus providing no meaningful content for our study. Consequently, these entries must be removed from the data sets. Another notable feature in the data sets is the presence of numerous invalid text entries. This is especially prominent in the UbuntuForums data set, which includes special keywords and punctuation marks commonly used in Linux commands.

To ensure consistency and uniformity in data handling, the data sets were first converted from their semi-structured JSON format to CSV. Post conversion, the empty entries were dropped, and punctuation was generally eliminated. However, it should be noted that the punctuation marks within the UbuntuForums data set were treated differently due to their significance in Linux commands, distinguishing this data set from the TripAdvisor NYC and Photography-on-the.net data sets.

Subsequently, the data underwent a series of text normalization procedures, starting with converting all text to lowercase. Other actions included identifying missing sentence delimiters, controlling for letter and non-word repetition, pinpointing words contained within parentheses (assumed to contribute less meaningful content), and finally, spotting phrase repetition. Words that were potentially misplaced or excessively repeated were stripped out to reduce noise and enhance the accuracy of text analysis (Saleem et al., 2017).

A crucial step involved filtering the text to remove numbers and special characters, this being a precursor to detecting nouns. The focus on nouns arises from our primary interest in the topics present within the data sets. An additional layer of refinement was the correction of detected nouns using a multilingual module. This module allowed for language detection and appropriate word selection based on language dictionaries. The final preparatory step involved filtering out stop words. We relied on a multilingual stop word list to effectively handle all potential stop words appearing in the forums.

Upon completion of data cleansing, each word in the datasets underwent stemming. This process, crucial for subsequent model processing, ensured reduced variance in the data without imposing an excessive computational burden (Gupta et al., 2018).

3.2. Encoding and Topic Generation

In this study, an innovative approach is employed which merges Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT) to discover underlying topics in each forum discussion. LDA is a probabilistic topic modeling technique typically used to discern topics within a corpus of text data, whereas BERT is a pre-trained language model designed to produce high-quality vector representations of text.

The primary aim is to leverage BERT to generate vector representations of text data, subsequently applying LDA on these representations in accordance with a specified parameter to detect intrinsic topics (Bouabdallaoui et al., 2022). The technical implementation involves creating a term dictionary from tokenized documents. The preprocessed tokens are converted into identifiers, which are then transposed into a document-term matrix, regarded as the corpus for this study. BERT Encoder and LDA are independently administered to the sentences and tokens, respectively, leading to the generation of two comprehensive vectors. Each vector comprises information pertaining to each token over the corpus.

The two vectors undergo ‘slice translation’, resulting in concatenation along the second axis. The LDA vector is scaled by a fixed parameter, represented as Γ, to confer a degree of relative importance to LDA. The equation to describe the vector concatenation process is as follows:

$$V_{LDA-BERT_{ij}} = \begin{cases} V_{LDA_{ij}} * \tau; if\ j < n \\ V_{BERT_{i(j-n)}}; if\ j \geq n \end{cases} \tag{1}$$

With V_{LDA} and V_{BERT} are two vectors of the same length n.

The $V_{LDA-BERT_{ij}}$ is the element in the i^{th} row and j^{th} column of the concatenated vector $V_{LDA-BERT}$, $V_{LDA_{ij}}$ is the element in the i^{th} row and j^{th} column of the first vector V_{LDA} multiplied by the value of Γ, and $V_{BERT_{(j-n)}}$ is the element in the i^{th} row and $(j-n)^{th}$ column of the second vector V_{BERT}.

The generated vector is considered as the vectorization representation of the corpus and can be mainly used for topic modeling purposes. A single layer autoencoder is implemented to reduce dimensionality and extract important features from the output vector, since it is based on an LDA representation which can help on extracting latent topics (Wang et al., 2020). The novelty in this work is that instead of training the single layer autoencoder over a bag-of-words vectors which is less efficient, we used BERT encoders so as the performance of the LDA would be improved by reducing efficiently the dimensionality and removing the irrelevant attributes to gain only the most important information.

3.3. Clustering Threads

The integration of a single-layer autoencoder is utilized in this study to diminish the dimensionality of BERT and LDA vectors, facilitating the employment of their compressed representations as inputs for a clustering algorithm. This algorithm serves to categorize similar documents based on their latent topics (Atagün et al., 2021). The k-means clustering algorithm is commonly used for this purpose, forming 'k' clusters from data points predicated on their similarity. Application of k-means clustering to the compressed representations derived from the single autoencoder requires specifying the number of clusters, represented as 'k'. Subsequently, the k-means algorithm iteratively assigns data points to the nearest cluster center (the mean of the data points in the cluster), continuously updating cluster centers until convergence is achieved. The clusters generated signify groups of documents sharing similarities in their latent topics, enabling topic analysis within different document subsets and the identification of patterns and relationships within the document collection (Curiskis et al., 2020).

In the context of this study, the forum data extracted encompasses a broad range of discussed topics, prompting the challenge of choosing an initial 'k'. This necessitates identification of frequently occurring topics within these discussion forums. As the number of topics increases, the clustering score diminishes, potentially leading to skewed results. To counter this, we have harnessed the capability of genetic algorithms in identifying optimal centroids. This is achieved by selecting a random population applicable within a centroid computation space, enabling it to seamlessly adapt to this space (Rahman et al., 2014). The algorithm proposed incorporates a genetic algorithm to execute clustering based on the autoencoder results. The algorithm is partitioned into three components: population construction, chromosome identification coupled with fitness calculation, and genetic operations.

The population is initiated with a crossover probability and a mutation probability, a generation of 'n', and a population size of 'N' (where 'n' and 'N' are positive integers). Chromosomes are discerned within each pool, indexed based on gene existence, and an adjacency matrix is subsequently formed using these indexes. The k-means algorithm is applied for chromosome clustering, with the Silhouette coefficient serving as the fitness function. Genetic operations encompass random parent selection using the tournament algorithm, crossover operation to create two offspring, and mutation operation for each gene. The children produced are appended to the offspring, and the generation of the parents is preserved. The best generation is fitted to the BERT and LDA vector, followed by the application of the k-means algorithm using the same 'k' parameter (i.e., number of topics to be clustered).

The amalgamation of these algorithms, trained on all data extracted from the forums, enables the management of a multitude of attributes whilst minimizing the risk of overfitting. A constraint emerges with the choice of topics for clustering using k-means; a higher number of topics corresponds to a lower Silhouette score, thereby encouraging the use of genetic algorithms to define the most suitable populations, yielding optimal clustering results. This combined approach is implemented for each post within each thread, facilitating topic prediction for each post based on the 'k' number ascertained in the Genetic Algorithms and K-Means model combination. It is crucial to note that a smaller number of topics may result in the elimination of certain posts, due to a lack of a defined topic relative to the basic model, or because the post contains only undesired words or characters.

3.4. Topic Similarity

Sentence similarity serves as a vital application in topic modeling, wherein the similarity score between two topic vectors is computed to identify shared topics. This mechanism aids in the detection of semantic issues and enhances the precision of recommendations (Jeong et al., 2019).

Regarded as a measure of likeness between two non-zero vectors within an inner product space, Cosine Similarity is commonly employed in information retrieval systems to identify documents exhibiting maximum similarity to a given query. In the realm of information retrieval, text data is often represented as a high-dimensional vector. Each dimension corresponds to a word present within the document collection. This metric quantifies the angle between two vectors in a high-dimensional space, disregarding the magnitude of the vectors and focusing solely on their directions. This characteristic renders Cosine Similarity particularly suited to text data. The magnitude of the vectors may exhibit significant variation, but the direction may hold substantial relevance in determining similarity (Kalhori et al., 2018).

Upon calculation of the Cosine Similarity between the query vector and each document vector, the documents are ranked in accordance with their similarity scores. The documents possessing the highest similarity scores are deemed most relevant to the query and are consequently returned as search results. This metric represents a rapid and scalable method for locating similar documents within extensive collections (Bisandu et al., 2019).

In order to quantify the resemblance between the deduced topics, a similarity matrix is constructed employing the cosine similarity of two topic vectors. In accordance with vector space modeling, we illustrate these two vectors as $V_{t_i} = (\alpha t_{i,1}, \ldots, \alpha t_{i,N})$ and $\mathrm{V}_{t_j} = (\alpha t_{j,1}, \ldots, \alpha t_{j,M})$. The weight denotes the relative term frequency of the i^{th} word in a sentence, which is computed as follows:

$$at_{s,k} = \frac{n_{s,k}}{\sum_{K}(n_{s,K})} \tag{2}$$

where $n_{s,k}$ denotes the number of times the i^{th} word appears in the sentence s, $\sum_{K}(n_{s,K})$ denotes the total number of occurrences of all words in a vector of a sentence, where K is the length of the sentence vector.

We measure the similarity between two sentences by computing the cosine similarity of their vector representations V_{t_i} and V_{t_j} as follows:

$$\cos\left(V_{t_i}, V_{t_j}\right) = \frac{\left\langle V_{t_i}, V_{t_j} \right\rangle}{\left|V_{t_i}\right|\left|V_{tj}\right|} \tag{3}$$

Where:

$$\left\langle V_{t_i}, V_{t_j} \right\rangle = \sum_{k=1}^{K} \sum_{s=i}^{j} at_{s,k}$$

And:

$$\begin{cases} \left|V_{t_i}\right| = \sqrt{\sum_{k=1}^{K} at_{i,k}} \\ \left|V_{t_j}\right| = \sqrt{\sum_{k=1}^{K} at_{j,k}} \end{cases}$$

The similarity matrix M is symmetric and has a diagonal of 1 since the similarity between two same vectors equals to 1, we can choose then only upper triangular matrix or the lower triangular matrix, where:

$$M(\cos(V_{t_i},V_{t_j})) = M(\cos(V_{t_i},V_{t_j}))^T \tag{4}$$

$$\text{with: } diag(M) = (1,\ldots,1)$$

This matrix is instrumental in constructing graphs wherein the nodes symbolize the posts and the edges represent the calculated cosine similarity scores between these posts.

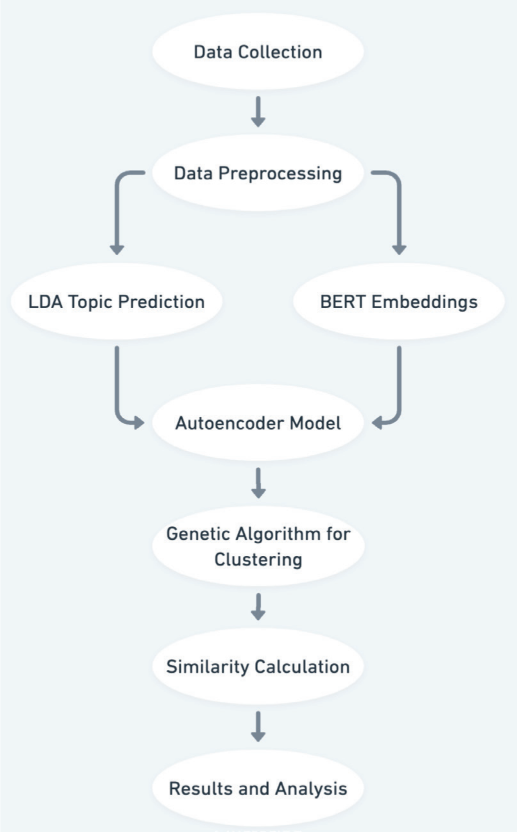

Figure 2. Flowshart of the pipeline

The schema above (Figure 2) depicts the methodology discussed in this section, beginning with data collection and preprocessing, followed by parallel processing using LDA for topic prediction and BERT for textual embeddings. These processes feed into an autoencoder model that reduces data dimensionali-ty, enhancing feature extraction. Subsequently, a genetic algorithm optimizes clustering, and similarity calculations assess the closeness of topics. The final stage involves analyzing the results to derive actionable insights, demonstrating an integrated approach to handling and interpreting complex textual data efficiently.

4. Experimental Setup

In the journey to refine and optimize our methodology for topic clustering, it is imperative to systematically analyze and understand the impact of various independent variables. This section delineates a comprehensive set of these variables, elucidating their potential roles and significance in the context of our methodology that synergistically combines Genetic Algorithms, K-Means, and LDA-BERT autoencoders.

- Genetic Algorithm Parameters:
    - Population Size: Investigating different population sizes offers insights into the trade-off between diversity of solutions and computational efficiency.
    - Mutation Rate: By tweaking the mutation rate, we can control the level of exploration in the solution space, impacting the algorithm's ability to escape local optima.
    - Crossover Rate: This variable influences the balance between exploitation of known good solutions and exploration of new potential solutions.
- K-Means Initialization:
    - Initial Centroid Selection: Analyzing different strategies for selecting initial centroids can unveil their impact on the convergence speed and quality of the final clusters.
    - Number of Clusters (K): Optimizing K is crucial, as it determines the granularity of the clustering. Genetic Algorithms can play a pivotal role in finding the optimal number of clusters.
- Text Embedding Quality:
    - Variance Retained in Autoencoder: This variable acts as a quality check, ensuring that the text representation retains the essential information for effective clustering.
    - Embedding Dimensionality: A careful balance is needed to ensure that the embeddings are of sufficient dimensionality to capture textual nuances, yet not so high-dimensional that they hinder the clustering process.
- Clustering Validation Metrics:
    - Silhouette Score: This metric provides a direct measure of the clustering quality, helping to validate the effectiveness of our methodology.
    - Intra-Cluster Distance: By minimizing intra-cluster distances, we ensure that the formed clusters are tight and well-defined.
- Computational Efficiency:
    - Runtime: Ensuring a swift convergence of the algorithm is paramount, especially when dealing with large and complex datasets.

- Memory Usage: Efficient memory usage ensures that our methodology remains scalable and accessible, even on hardware with limited resources.
- Data Characteristics:
  - Text Length: Understanding the impact of text length on the clustering process is vital for ensuring robust performance across diverse datasets.
  - Vocabulary Size: Managing the vocabulary size is crucial, as it affects both the text embedding stage and the clustering performance.

By meticulously analyzing and optimizing these independent variables, we not only enhance the performance of our methodology but also ensure its applicability and efficiency across diverse scenarios and datasets. This comprehensive approach ensures a robust validation of our methodology, paving the way for reliable and insightful topic clustering.

5. Results

This study was conducted to cluster topics of forums discussions and measure the topics similarity between topics. We used a mixed-methods approach for this study, which allowed us to gather both encoding techniques and evolutionary intelligence. The encoding techniques was performed by concatenating LDA and BERT word vectors with a parameter to equilibrate the concatenation, while the evolutionary intelligence was used to optimize the topic clustering of topics with the definition of the optimal generation.

5.1. Autoencoder Prediction

In this paper, we furnish a detailed examination of the results gleaned from our distinct methodological approach which encompasses the training of a single autoencoder model on three distinct forum discussion datasets. This analytical process was designed with a dual purpose: to assess the model's capability to accurately predict the topics under discussion within these forums and to gauge the proficiency of the single autoencoder model in performing this task.

Our model's training was predicated on Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT) vectors, following the preprocessing operations conducted on the three datasets. The preprocessing was executed with a Γ value of 35, and we established a comprehensive learning period comprising 300 epochs.

In the succeeding sections, we present graphical representations elucidating the performance of the single autoencoder model. These plots distinctly illustrate the error score progression of the model throughout its engagement with the three datasets. These visualizations serve to provide a more concrete understanding of the efficacy of the single autoencoder model in the context of predicting discussion topics within the examined forum datasets.

The outcomes of our empirical investigation revealed a commendable level of precision in topic prediction, courtesy of the autoencoder model. Evaluating the TripAdvisor NYC dataset (Figure 3), we observed congruity between the training and testing curves, both of which culminated in a Mean Squared Error (MSE) score under 0.2. Contrastingly, upon examining the Photography-on-the.Net dataset (Figure 4), it was discerned that the testing curve failed to align with the training curve, demonstrating a propensity for descending towards lower score values should the quantity of learning epochs be extended. The error curve generated by the UbuntuForum dataset (Figure 5) mirrored the results elicited from the TripAdvisor NYC dataset, exhibiting almost identical score values. The table below shows the accuracy of the three approaches:

Table 2. Performances of autoencoder model

Epochs	**Accuracy score**			**F1-score**		
	TripAdvisor NYC	***UbuntuForums***	***Photography-on-the.NET***	***TripAdvisor NYC***	***UbuntuForums***	***Photography-on-the.NET***
100	0.981	0.992	0.965	0.78	0.87	0.76
200	0.996	0.994	0.982	0.81	0.89	0.78
300	0.997	0.995	0.989	0.82	0.9	0.79

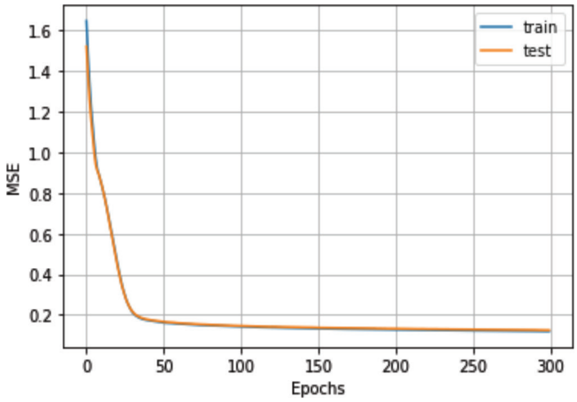

Figure 3. Learning curves for the TripAdvisor NYC Dataset

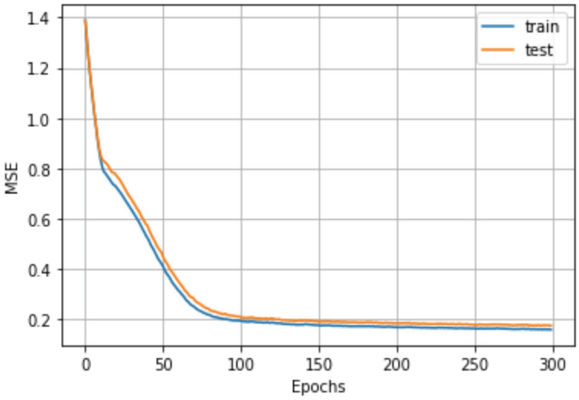

Figure 4. Learning curves for the Photography-on-the.Net Dataset

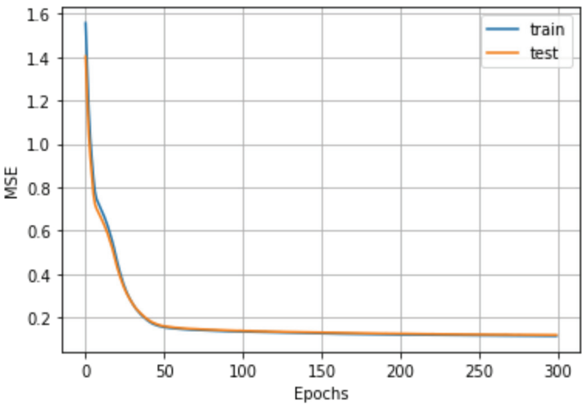

Figure 5. Learning curves for the UbuntuFormus Dataset

The aforementioned results imply a significant promise in the autoencoder model for large-scale topic modeling in expansive forum discussions. In essence, this indicates our capability to execute comprehensive topic detection encompassing a diverse range of subjects within a forum discourse. Consequently, this modeling approach could be harnessed in numerous pragmatic contexts, including monitoring social media trends (Jiang et al., 2021) and conducting market research (Hilmi et al., 2020).

5.2. Evolutionary Clustering

In our approach, we initially utilized K-means to aggregate data points of similar topics into clusters, taking distance measurements into account. However, it is important to acknowledge the limitations inherent to the K-means algorithm, such as its sensitivity to the initial selection of centroids and susceptibility to local optima (Ikotun et al., 2022). To surmount these constraints, we integrated a genetic algorithm that capitalizes on principles of natural selection and genetic operations to refine the cluster centroids. Following a comparative performance evaluation of both algorithms across the three textual datasets, it was determined that the genetic algorithm surpassed K-means in terms of both clustering precision and resilience.

In a more technical context, we preset a crossover probability of 0.8 and a mutation probability of 0.3, which are considered as the standard probability parameters in Genetic Algorithms. In addition, to optimize computational efficiency, we established 4 clusters, defined 10 generations, and a population size of 10.

The operational framework of the proposed algorithm is divided into three sections. The initial section entails generating a population for the genetic algorithm, incorporating specific parameters and data structures. The subsequent section is devoted to identifying chromosomes within the population, which are then used to construct an adjacency matrix. This matrix undergoes clustering using the K-Means algorithm, with the fitness determined using the Silhouette coefficient. The final section executes genetic operations, including parent selection via a tournament algorithm, crossover and mutation operations, and the addition of new offspring to the population. The algorithm also incorporates a parameter known as selection pressure, which influences the convergence rate. The resultant offspring, in conjunction with the parent generation, is preserved for future analysis.

We utilized the silhouette coefficient as a metric for evaluating the effectiveness of clustering executed using the Genetic Algorithm and K-Means algorithm across the aforementioned three forums. The subsequent plots depict the silhouette score progression across generational iterations of the genetic algorithm.

After evaluating the silhouette score over the selected generations (Figures 6, 7 and 8), we fit the optimal generation on the LDA-BERT vector and perform K-Means clustering to evaluate the silhouette score and the quantization error. Table 3 describes these scores obtained from the best generation:

Table 3. Silhouette score and quantization

Datasets	**Scores**	
	Silhouette Score	***Quantization Error***
TripAdvisor NYC	**0.7102**	5.3612
Photography-on-the.Net	0.6289	8.3299
UbuntuForums	0.4975	**1.9555**

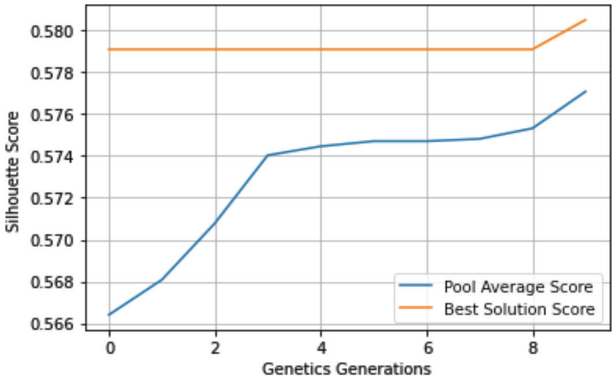

Figure 6. Silhouette score following gentrics generations for the Photography-on-the.Net Dataset

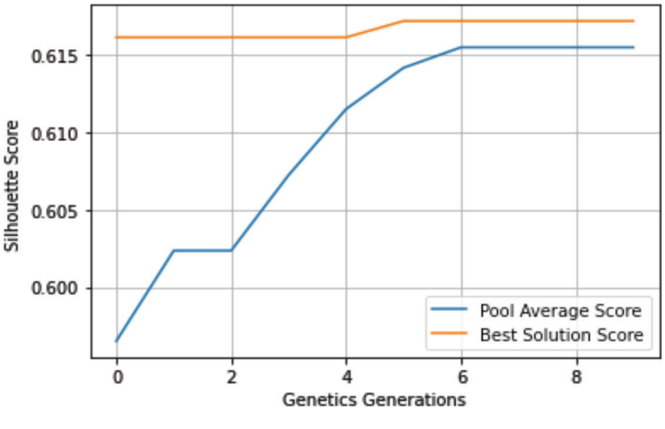

Figure 7. Silhouette score following gentrics generations for the TripAdvisor NYC Dataset

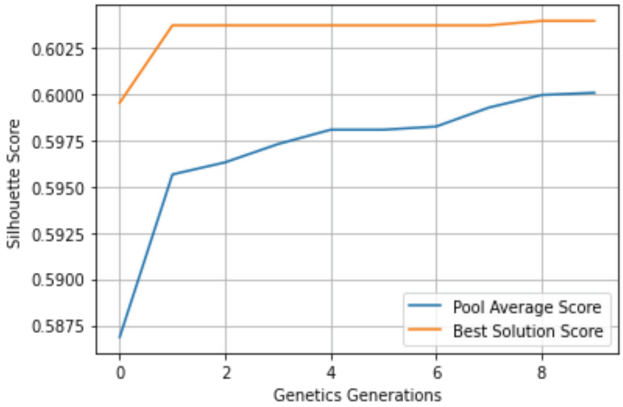

Figure 8. Silhouette score following gentrics generations for the UbuntuForums Dataset

In the provided table, it can be observed that the dataset from Photography-on-the.Net returns the highest quantification error score, followed closely by the TripAdvisor NYC dataset. Both of these datasets are substantial in size, boasting a significant volume of posts and threads in comparison to the Ubuntu forums dataset, which registers the lowest quantification error score.

However, an alternate perspective is provided when we examine the silhouette values. The TripAdvisor NYC dataset exhibits the highest silhouette value at 71%, signifying an accuracy score for clusters of topics pertinent to tourism. The Photography-on-the.Net forum, despite its vast assortment of photography-centric topics, manages a respectable silhouette value of 62%. The UbuntuForums, a forum dedicated to technical resolutions and problem-solving, delivers a silhouette value of 49.75%. This lower silhouette score can be attributed to the complexity of topic detection in a forum rife with intricate commands and proposed user solutions (Yang et al., 2021), where punctuation is integral and cannot be removed, making it challenging for the encoders to comprehend.

Such outcomes propose that genetic algorithms could serve as a beneficial substitute to k-means in the context of topic clustering, particularly when dealing with complex or noisy data (Santhanam et al., 2015).

5.3. Similarity Evaluation

The similarity matrix delineated in the methodology section offers a quantifiable measure of likeness between pairs of topic vectors, as well as facilitates the detection of patterns and associations among distinct data points. The matrix renders an efficacious graphical illustration of the similarity degree between each topic vector pair, with higher values suggesting greater similarity and lower values indicating reduced likeness. Examination of the similarity matrix outcomes allows us to glean insights into the forum data's inherent structure, thereby enabling informed decisions concerning further analysis.

Assessment of a similarity matrix in the context of topic modeling ideally requires a synthesis of intrinsic and/or extrinsic evaluation measures (Qiu et al., 2018), coupled with a visual examination of the findings (Cao et al., 2016), to procure a comprehensive comprehension of the topic model's quality and applicability. For the scope of this paper, we engage intrinsic evaluation measures, which involve juxtaposing the topic model's output against a set of predetermined topics or gold-standard data (Qiu et al., 2018). Intrinsic evaluation measures might incorporate metrics such as coherence, gauging the quality and uniqueness of the model-generated topics (Costa et al., 2021).

To employ intrinsic evaluation to our forecasted data, we calculated the coherence metric between the K-Means centroids (most frequent topics) and the predicted labels to ascertain word similarity, subsequently comparing the results with the similarity matrix to discern any potential correlation between the cosine similarity and coherence. To streamline visualization and computational processes, we normalized the analysis across the three datasets, examining similarities and coherence for the first ten samples (Wu et al., 2016). Table 4 respectively describes the coherence value means and cosine similarity value means of these ten samples from each dataset.

Table 4. Coherence values

Datasets	**Scores**	
	Mean of Coherence scores	***Mean of Cosine Similarity***
TripAdvisor NYC	0.5244	0.4828
Photography-on-the.Net	0.3271	0.3733
UbuntuForums	0.4497	0.3946

As shown in the table above, we can notice that the means are quite close between each other for the TripAdvisor NYC and Photography-on-the.Net, while for the UbuntuForums the scores are distant, this means that the silhouette score obtained for the Ubuntu forums impacted the similarity between topics, while for the other datasets, we notice that there is a balance between the cosine score and the coherence.

Furthermore, to assign a logical sequence to the results of the similarity matrix, we modeled the results with a directed graph where the nodes contain the predicted topics and the edges are the calculated similarity values. Thus, we assigned letters from A to J (to refer to the 10 samples of each dataset), as follows:

Table 5. 10 first topics of each predicted dataset

Datasets	**Scores**	
	Sample	***Topics***
TripAdvisor NYC	A	[‘hotel’, ‘employe’, ‘luggag’, ‘piec’, ‘staff’, ‘bag’, ‘member’, ‘staff’, ‘hail’, ‘cab’, ‘buck’]
	B	[‘nyx’, ‘day’, ‘peopl’, ‘room’, ‘day’, ‘husband’, ‘tip’, ‘bag’, ‘hotel’]
	C	[‘thank’, ‘everyon’]
	D	[‘bag’, ‘peopl’, ‘work’, ‘day’, ‘buck’, ‘cab’]
	E	[‘wow’, ‘peopl’, ‘taxi’, ‘buck’, ‘drive’, ‘driver’, ‘dollar’, ‘day’, ‘maid’, ‘dollar’, ‘day’, ‘figur’, ‘maid’, ‘room’, ‘day’, ‘dollar’, ‘day’, ‘tip’, ‘pay’]
	F	[‘head’, ‘youtub’, ‘elf’]
	G	[‘housekeep’, ‘anyth’, ‘room’, ‘day’, ‘comic’, ‘taxi’, ‘hour’, ‘ga’, ‘price’, ‘danger’, ‘job’]
	H	[‘chariti’, ‘appel’, ‘tip’, ‘youtub’]
	I	[‘thank’, ‘stay’, ‘place’, ‘tip’, ‘tip’, ‘restaur’, ‘outdoor’, ‘season’, ‘sign’, ‘appreci’, ‘employe’, ‘luggag’, ‘room’, ‘anyon’, ‘anyon’, ‘hail’, ‘taxi’, ‘year’, ‘dollar’, ‘tip’, ‘bag’, ‘sens’, ‘time’, ‘day’, ‘day’, ‘suit’, ‘anyon’, ‘read’, ‘tip’, ‘night’, ‘hotel’, ‘night’]
	J	[‘parent’, ‘decad’, ‘job’, ‘time’, ‘room’, ‘tip’, ‘time’, ‘cross’, ‘home’, ‘tip’, ‘hospit’, ‘home’, ‘comic’]

(*continued*)

Table 5. 10 first topics of each predicted dataset (*continued*)

Datasets	**Scores**	
	Sample	***Topics***
Photography-on-the.Net	A	[‘porn’, ‘member’, ‘million’, ‘dollar’, ‘year’, ‘accessori’, ‘trip’, ‘beer’, ‘tripod’, ‘car’, ‘someth’, ‘let’, ‘share’, ‘feel’, ‘member’, ‘visitor’, ‘part’, ‘money’, ‘chariti’, ‘choic’, ‘holiday’, ‘season’, ‘distress’, ‘lack’, ‘food’, ‘water’, ‘care’, ‘medicin’, ‘thank’, ‘holiday’, ‘imag’, ‘forum’, ‘pukka’, ‘share’, ‘qualiti’, ‘preview’, ‘pleas’, ‘log’, ‘qualiti’, ‘stuff’, ‘imag’]
	B	[‘thank’, ‘remind’, ‘boss’, ‘joy’, ‘store’, ‘bell’, ‘ringer’, ‘dollar’, ‘lot’, ‘peopl’, ‘ladi’, ‘night’, ‘visit’, ‘centrepiec’, ‘holiday’, ‘kitchen’, ‘tabl’, ‘christma’, ‘holiday’, ‘everyon’, ‘porn’]
	C	[‘year’, ‘donat’, ‘marin’, ‘corp’, ‘reserv’, ‘campaign’]
	D	[‘toy’, ‘honor’, ‘marin’, ‘salvat’, ‘armi’, ‘spin’, ‘bifid’, ‘foundat’, ‘honor’, ‘tim’]
	E	[‘tag’, ‘year’, ‘christma’, ‘angel’, ‘tree’, ‘mall’, ‘salvat’, ‘armi’, ‘station’, ‘someth’, ‘kettl’]
	F	[‘pm’, ‘calendar’, ‘thing’]
	G	[‘famili’, ‘share’, ‘time’, ‘year’]
	H	[‘enough’, ‘year’, ‘vet’, ‘goodwil’, ‘holiday’, ‘food’, ‘bank’, ‘groceri’, ‘store’, ‘thing’, ‘packag’, ‘price’, ‘peopl’, ‘food’, ‘item’, ‘year’, ‘comput’, ‘center’, ‘children’]
	I	[‘organ’, ‘year’, ‘harder’]
	J	[‘salvat’, ‘armi’, ‘chang’, ‘back’, ‘kettl’, ‘salvat’, ‘armi’, ‘minist’, ‘head’, ‘honcho’, ‘live’, ‘presid’, ‘organ’, ‘figur’, ‘lot’]
Ubuntu Forums	A	[‘use’, ‘kernel’, ‘boot’, ‘paramet’]
	B	[‘use’, ‘kernel’, ‘boot’, ‘paramet’]
	C	[‘recoveri’, ‘grub’, ‘menu’, ‘grub’, ‘menu’, ‘recoveri’, ‘usernam’, ‘instal’, ‘command’, ‘code’, ‘grub’, ‘menu’, ‘est’, ‘machin’, ‘shift’, ‘legaci’, ‘bio’, ‘case’, ‘usernam’, ‘uppercas’, ‘type’, ‘password’, ‘password’, ‘recoveri’, ‘step’, ‘time’]
	D	[‘keyboard’, ‘layout’, ‘keyboard’, ‘layout’, ‘usernam’, ‘layout’, ‘password’, ‘key’, ‘password’, ‘usernam’, ‘charact’]
	E	[‘recoveri’, ‘grub’, ‘menu’, ‘grub’, ‘menu’, ‘recoveri’, ‘usernam’, ‘instal’, ‘command’, ‘code’, ‘grub’, ‘menu’, ‘est’, ‘machin’, ‘shift’, ‘legaci’, ‘bio’, ‘case’, ‘usernam’, ‘uppercas’, ‘type’, ‘password’, ‘password’, ‘recoveri’, ‘step’, ‘time’, ‘menu’, ‘login’, ‘prompt’, ‘whirl’, ‘thank’, ‘thing’, ‘case’, ‘user’, ‘name’, ‘input’, ‘password’, ‘mismatch’, ‘warn’, ‘setup’, ‘mistak’]
	F	[‘updat’, ‘reinstal’, ‘keyboard’, ‘acc’, ‘password’, ‘book’, ‘password’, ‘correct’, ‘server’, ‘login’, ‘password’, ‘modem’, ‘password’, ‘user’, ‘password’, ‘client’, ‘server’, ‘someth’, ‘thank’, ‘grub’, ‘instal’, ‘grub’, ‘prompt’, ‘read’, ‘help’, ‘guid’, ‘grub’, ‘troubl’, ‘boot’, ‘record’, ‘tip’, ‘cheer’]

(*continued*)

Table 5. 10 first topics of each predicted dataset (continued)

Datasets	**Scores**	
	Sample	***Topics***
	G	[‘recoveri’, ‘grub’, ‘menu’, ‘grub’, ‘menu’, ‘recoveri’, ‘usernam’, ‘instal’, ‘command’, ‘code’, ‘grub’, ‘menu’, ‘est’, ‘machin’, ‘shift’, ‘legaci’, ‘bio’, ‘case’, ‘usernam’, ‘uppercas’, ‘type’, ‘password’, ‘password’, ‘recoveri’, ‘step’, ‘time’]
	H	[‘keyboard’, ‘layout’, ‘keyboard’, ‘layout’, ‘usernam’, ‘layout’, ‘password’, ‘key’, ‘password’, ‘usernam’, ‘charact’]
	I	[‘recoveri’, ‘grub’, ‘menu’, ‘grub’, ‘menu’, ‘recoveri’, ‘usernam’, ‘instal’, ‘command’, ‘code’, ‘grub’, ‘menu’, ‘est’, ‘machin’, ‘shift’, ‘legaci’, ‘bio’, ‘case’, ‘usernam’, ‘uppercas’, ‘type’, ‘password’, ‘password’, ‘recoveri’, ‘step’, ‘time’, ‘menu’, ‘login’, ‘prompt’, ‘whirl’, ‘thank’, ‘thing’, ‘case’, ‘user’, ‘name’, ‘input’, ‘password’, ‘mismatch’, ‘warn’, ‘setup’, ‘mistak’]
	J	[‘updat’, ‘reinstal’, ‘keyboard’, ‘acc’, ‘password’, ‘book’, ‘password’, ‘correct’, ‘server’, ‘login’, ‘password’, ‘modem’, ‘password’, ‘user’, ‘password’, ‘client’, ‘server’, ‘someth’, ‘thank’, ‘grub’, ‘instal’, ‘grub’, ‘prompt’, ‘read’, ‘help’, ‘guid’, ‘grub’, ‘troubl’, ‘boot’, ‘record’, ‘tip’, ‘cheer’]

The following three graphs show the links between the topics defined in Table 5 according to each forum.

In Figure 9, we observe that the majority of topic vectors exhibit relatively lower relationships, where cosine values are lesser. An exception is the correlation between Topic B and Topic E, where similarity surpasses 50%. This pattern can be attributed to the diverse responses often encountered in travel tourism forums, which may result in a high semantic rate within a thread (Colladon et al., 2020). This poses a challenge in scenarios where users require precise answers.

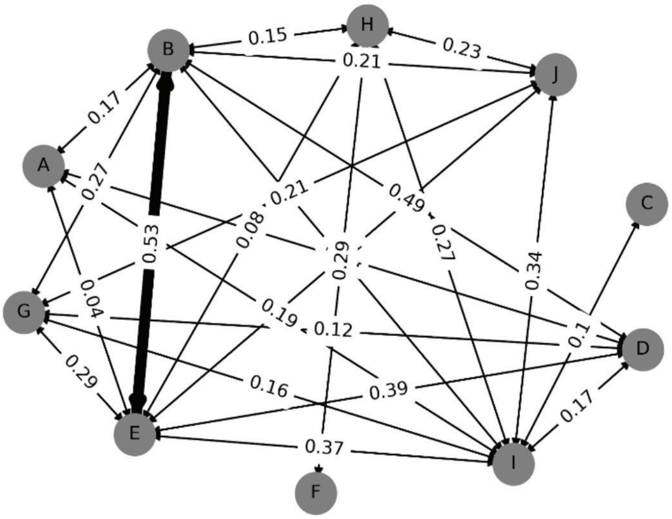

Figure 9. Topic similarity graph of the TripAdvisor NYC samples

The Ubuntuforums graph depicted in Figure 10 reveals several similarities between the topic vectors, occasionally reaching 100% similarity, implying identical vectors. This pattern can be explained by the technical or scientific content typical of this forum type. The solutions to a technical problem discussed in a thread can often be resolved with a limited or similar set of solutions, thus explaining the correlation between similarity values for technical topics (Jia et al., 2021).

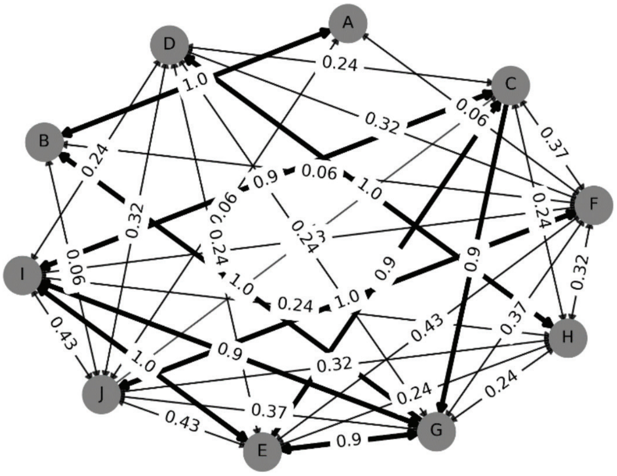

Figure 10. Topic similarity graph of the UbuntuForums samples

In Figure 11, the similarity values are considerably lower, with a lesser weighted proportion than that of TripAdvisor NYC. We note that the maximum similarity value is 35%, linking Topic B and Topic J, with nodes E and I demonstrating multiple connections with other topics.

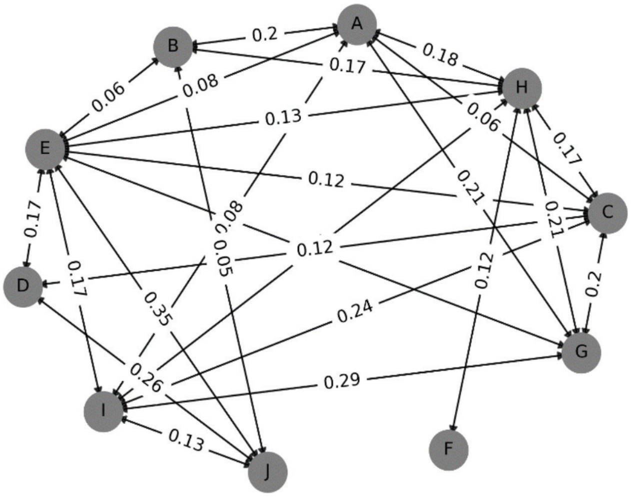

Figure 11. Topic similarity graph of the Photography-on-the.Net

The results from the graphs indicate that Ubuntuforums exhibits more relevant topics, given its consistency and lowest quantization score, compared to the other datasets. The Photography-on-the.net forum registers the highest quantization score, which adversely impacts the similarity results, as demonstrated in the preceding plot.

6. Discussion

This study has provided a comprehensive analysis of topic clustering and similarity in forum discussions across three different domains. Utilizing a combination of LDA and BERT for topic prediction, an autoencoder for topic clustering, and a genetic algorithm for optimization, the study aimed at understanding the intricacies of topic similarities in expansive forum discussions.

The autoencoder model demonstrated a commendable level of precision in topic prediction, especially notable in the TripAdvisor NYC and UbuntuForums datasets. The learning curves and Mean Squared Error (MSE) scores suggest that the model was able to effectively capture the underlying topic structures in these forums. However, the divergence in the Photography-on-the.Net dataset's testing curve indicates potential overfitting or a need for additional epochs for convergence.

The evolutionary clustering with a genetic algorithm showcased an enhanced performance over the traditional K-means algorithm. The silhouette scores from the genetic algorithm provided evidence of well-defined cluster separations, particularly in the TripAdvisor NYC dataset. This implies the potential of this approach for topic modeling in forums with diverse and expansive discussions.

The similarity matrices, modeled as directed graphs, provided a visual representation of topic relationships within the forums. The TripAdvisor NYC dataset displayed varied topic relationships, with occasional high similarity between topics such as B and E. This can be attributed to the diverse nature of travel-related discussions, where threads might span a wide array of subjects yet maintain semantic coherence.

In contrast, the UbuntuForums dataset demonstrated a high degree of similarity across several topics, reflecting the technical nature of discussions where solutions to problems may be consistently similar or related. This consistency is further supported by the forum's lower quantization error, indicating a higher relevance and precision in the clustered topics.

The Photography-on-the.Net dataset presented a scenario with generally lower similarity values, aligning with its highest quantization error among the three forums. This suggests a potential dilution of topic relevance, possibly due to the vast range of photography-related discussions.

The findings from this study have significant implications for the development of topic modeling and information retrieval systems. The varying degrees of topic similarity across different forum types underscore the necessity for adaptive algorithms that can cater to the unique characteristics of discourse in diverse online communities.

Future research could delve deeper into refining the autoencoder model and evolutionary clustering approach, perhaps exploring alternative optimization techniques or modifications in the encoding process to better handle diverse datasets. Additionally, a more extensive evaluation of topic similarity, incorporating extrinsic evaluation measures and larger sample sizes, could provide a more robust understanding of the models' performance and the nature of topic relationships in online forums.

7. Conclusion

In conclusion, this work has presented an in-depth analysis of the application of autoencoder models and genetic algorithms in topic modeling of large forum discussions. Through the careful application of methodologies and rigorous testing across three different forums, we demonstrated that the autoencoder model exhibits high accuracy in predicting topics, especially when combined with genetic algorithms for clustering. Particularly, the model proved effective in large, diverse datasets such as TripAdvisor NYC and Photography-on-the.Net, despite the intricacies involved with semantic variances.

However, there were some limitations in our study that provide avenues for future research. First, the challenge of handling technical forums, such as UbuntuForums, was evident where unique identifiers like command sequences played crucial roles in topic determination, but were not fully captured by the encoders. Moreover, the study also revealed the sensitivity of K-means to initial centroid placements and its propensity for local optima, which while addressed through the use of genetic algorithms, opens the door for further exploration into other optimization techniques.

Looking forward, it would be worthwhile to explore how other models, such as transformer-based models, fare in topic prediction. There's also a need to better handle specialized forums, potentially through additional preprocessing steps or more specialized models. Further research could also include exploring the application of this methodology in real-time scenarios and evaluating its performance. Additionally, examining how well these techniques scale to much larger datasets and identifying ways to improve efficiency will also be valuable for future research.

8. Acknowledgements

We would like to address our acknowledgements to the Moroccan National Center for Scientific and Technical Research (CNRST) and the Moroccan Institute for Scientific and Technical Information (IMIST) for giving us the access to use their High-Performance Computing (MARWAN).

9. References

Adams, P. H., & Martell, C. H. (2008). Topic detection and extraction in chat. *In 2008 IEEE International Conference on Semantic Computing* (pp. 581-588). https://doi.org/10.1109/ICSC.2008.61

Alsayat, A., & El-Sayed, H. (2016). Social media analysis using optimized K-Means clustering. In *2016 IEEE 14th Inter-national Conference on Software Engineering Research, Management and Applications (SERA)* (pp. 61-66). https://doi.org/10.1109/SERA.2016.7516129

Atagün, E., Hartoka, B., & Albayrak, A. (2021). Topic Modeling Using LDA and BERT Techniques: Teknofest Example. In *2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 660-664). https://doi.org/10.1109/UBMK52708.2021.9558988

Bisandu, D. B., Prasad, R., & Liman, M. M. (2019). Data clustering using efficient similarity measures. *Journal of Statistics and Management Systems, 22*(5), 901-922. https://doi.org/10.1080/09720510.2019.1565443

Bouabdallaoui, I., Guerouate, F., Bouhaddour, S., Saadi, C., & Sbihi, M. (2022). A hybrid Latent Dirichlet Allocation-BERT approach for topic discovery of market places. https://doi.org/10.21203/rs.3.rs-1674353/v1

Bouabdallaoui, I., Guerouate, F. & Sbihi, M. (2023). Combination of genetic algorithms and K-means for a hybrid topic modeling: tourism use case. *Evol. Intel. 17*, 1801–1817. https://doi.org/10.1007/s12065-023-00863-x

Cao, N., & Cui, W. (2016). *Introduction to text visualization*. Springer. https://doi.org/10.2991/978-94-6239-186-4

Colladon, A. F., Grippa, F., & Innarella, R. (2020). Studying the association of online brand importance with museum vis-itors: An application of the semantic brand score. *Tourism Management Perspectives, 33*, 100588. https://doi.org/10.1016/j.tmp.2019.100588

Costa, G., & Ortale, R. (2021). Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors. *Information Sciences, 563*, 226-240. https://doi.org/10.1016/j.ins.2021.01.019

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic model-ling in two online social networks: Twitter and Reddit. *Information Processing & Management, 57*(2), 102034. https://doi.org/10.1016/j.ipm.2019.04.002

Eklund, A., Forsman, M., & Drewes, F. (2023). An empirical configuration study of a common document clustering pipe-line. *Northern European Journal of Language Technology (NEJLT), 9*(1). https://doi.org/10.3384/nejlt.2000-1533.2023.4396

Gokarn Nitin, M., Gottipati, S., & Shankararaman, V. (2019). Clustering models for topic analysis in graduate discussion forums. In *Proceedings of the 27th International Conference on Computers in Education*. https://ink.library.smu.edu.sg/sis_research/4516

Gupta, R., & Jivani, A. G. (2018). *Analyzing the stemming paradigm. In Information and Communication Technology for Intelligent Systems (ICTIS 2017) -Volume 2* (pp. 333-342). Springer. https://doi.org/10.1007/978-3-319-63645-0_37

Hilmi, M. F., Mustapha, Y., & Omar, M. T. C. (2020). Innovation in an Emerging Market: A Bibliometric and Latent Di-richlet Allocation Based Topic Modeling Study. In *2020 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 882-886). https://doi.org/10.1109/DASA51403.2020.9317278

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences, 622*, 178-210. https://doi.org/10.1016/j.ins.2022.11.139

Jeong, B., Yoon, J., & Lee, J.-M. (2019). Social media mining for product planning: A product opportunity mining ap-proach based on topic modeling and sentiment analysis. *International Journal of Information Management, 48*, 280-290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009

Jia, J., Tumanian, V., & Li, G. (2021). Discovering semantically related technical terms and web resources in Q&A discus-sions. *Frontiers of Information Technology & Electronic Engineering, 22*(7), 969-985. https://doi.org/10.1631/FITEE.2000186

Jiang, L. C., Chu, T. H., & Sun, M. (2021). Characterization of vaccine tweets during the early stage of the COVID-19 outbreak in the United States: topic modeling analysis. *Jmir Infodemiology, 1*(1), e25636. https://doi.org/10.2196/25636

Kalhori, H., Alamdari, M. M., & Ye, L. (2018). Automated algorithm for impact force identification using cosine similari-ty searching. *Measurement, 122*, 648-657. https://doi.org/10.1016/j.measurement.2018.01.016

Obasa, A. I., Salim, N., & Khan, A. (2016). Hybridization of bag-of-words and forum metadata for web forum question post detection. *Indian Journal of Science and Technology, 8*(32), 1-12. https://doi.org/10.17485/ijst/2015/v8i32/92127

Pattabiraman, K., Sondhi, P., & Zhai, C. (2013). Exploiting forum thread structures to improve thread clustering. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval* (pp. 64-71). https://doi.org/10.1145/2499178.2499196

Qiu, Y., Li, H., Li, S., Jiang, Y., Hu, R., & Yang, L. (2018). Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19--21, 2018, Proceedings 17* (pp. 209-221). https://doi.org/10.1007/978-3-030-01716-3_18

Rahman, M. A., & Islam, M. Z. (2014). A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowledge-Based Systems, 71*, 345-365. https://doi.org/10.1016/j.knosys.2014.08.011

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces.

Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer Science, 47*, 76-83. https://doi.org/10.1016/j.procs.2015.03.185

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! https://doi.org/10.18653/v1/2020.emnlp-main.135

Wang, B., Liakata, M., Zubiaga, A., & Procter, R. (2017). A hierarchical topic modelling approach for tweet clustering. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceed-ings, Part II 9* (pp. 378-390). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_30

Wang, C., Zhang, H., Chen, B., Wang, D., Wang, Z., & Zhou, M. (2020). Deep relational topic modeling via graph pois-son gamma belief network. *Advances in Neural Information Processing Systems, 33*, 488-500. https://proceedings.neurips.cc/paper/2020/hash/05ee45de8d877c3949760a94fa691533-Abstract.html

Wu, Y., Cao, N., Archambault, D., Shen, Q., Qu, H., & Cui, W. (2016). Evaluation of graph sampling: A visualization perspective. *IEEE transactions on visualization and computer graphics, 23*(1), 401-410. https://doi.org/10.1109/TVCG.2016.2598867

Yang, Z., Zhang, W., Yuan, F., & Islam, N. (2021). Measuring topic network centrality for identifying technology and technological development in online communities. *Technological Forecasting and Social Change, 167*, 120673. https://doi.org/10.1016/j.techfore.2021.120673