Topic Modelling

Niyathi Allu
UC, Riverside
nallu002@ucr.edu

Venkata Sai Kumar Gottumukkala
UC, Riverside
vgott001@ucr.edu

Harish Kumar Manepalli
UC, Riverside
hmane001@ucr.edu

Sree Charan Reddy Gangireddy
UC, Riverside
sgang011@ucr.edu

Abstract

This project is designed to delve into the world of topic modeling techniques and their effectiveness in analyzing a dataset comprising abstracts from research papers across various domains. The primary focus will be on comparing the performance of four prominent methods: TF-IDF, BERT, LDA (Latent Dirichlet Allocation), and a combination of BERT and LDA known as BERT+LDA. One of the objectives of this project is to explore the effectiveness of combining traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA), with state-of-the-art language models. Specifically, we aim to investigate how the integration of LDA with advanced language models, such as BERT (Bidirectional Encoder Representations from Transformers), can enhance the performance and capabilities of topic modeling. By conducting this comparative analysis, the project aims to ascertain which approach stands out as the most proficient in extracting meaningful topics from the dataset. The ultimate goal is to uncover hidden patterns and gain valuable insights into the underlying structures present in the collection of research paper abstracts.

1 Introduction

In the era of big data, the exponential growth of digital information has presented both opportunities and challenges for extracting knowledge and insights from vast document collections. The ability to effectively analyze and understand the underlying structures and themes within these collections is crucial for various domains such as academia, industry, and scientific research. Topic modeling techniques have emerged as powerful tools to address this challenge by automatically identifying latent topics and uncovering hidden patterns in textual data.

The traditional approach to topic modeling, such as Latent Dirichlet Allocation (LDA), has been widely adopted and has proven successful in many applications. However, with the advent of advanced language models like BERT (Bidirectional Encoder Representations from Transformers), there is a growing interest in exploring how these modern techniques can complement and enhance the capabilities of traditional methods.

This project aims to bridge the gap between traditional and modern topic modeling techniques by investigating the fusion of LDA with BERT, referred to as BERT+LDA. By integrating the contextual understanding and semantic representation capabilities of BERT with the probabilistic modeling of LDA, we anticipate that this hybrid approach will yield improved topic modeling results. The fusion of these techniques has the potential to capture both local and global semantic dependencies, enabling more accurate topic identification and finer-grained analysis of document collections.

Understanding the latent topics within research papers is of paramount importance. Furthermore, in industrial settings, effective topic modeling facilitates information retrieval, content recommendation systems, and market trend analysis. By extracting meaningful topics from document collections, organizations can gain a competitive edge by making data-driven decisions and deriving actionable insights.

Through project, we aim to address the following questions: How does the fusion of LDA and BERT impact the quality and interpretability of topic modeling results? Can this hybrid approach outperform traditional methods such as TF-IDF and standalone models like BERT? By comparing the performance of these techniques on a diverse dataset of research paper abstracts, we seek to provide valuable insights into the efficacy and applicability of different topic modeling approaches.

Overall, this project holds significant importance in improving our understanding of topic modeling techniques and their potential to unlock the

hidden treasures of knowledge buried within vast document collections. By leveraging the strengths of traditional and modern approaches, we can pave the way for more accurate, interpretable, and insightful analysis of textual data, fostering advancements in various domains and driving data-informed decision-making.

This report is structured as follows: Section 2 provides an overview of the related work in the field of topic modelling. Section 3 outlines the methodology used in the project, including the experimental setup and evaluation metrics. Section 4 presents the results of the project. Section 5 discusses the findings and their implications for the development of more efficient and accessible topic modelling techniques. Finally, Section 6 concludes the report by summarizing the key findings and their significance for the field. Overall, this project is expected to provide valuable insights into the effectiveness of different topic modelling algorithms, and with different clustering algorithms on the retrieved modelling vectors.

2 Related Work

we will discuss relevant papers in the field of topic modeling, including techniques that combine BERT with other models, such as BERTopic, optimizations and extensions to BERTopic, as well as seminal papers and novel approaches in topic modeling.

2.1 tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection [1] This paper presents tBERT, a method that combines topic models with BERT to enhance semantic similarity detection. By leveraging the topic distributions learned from topic models and the contextual embeddings captured by BERT, tBERT achieves improved performance in semantic similarity tasks. The authors propose a joint training framework that integrates topic models and BERT, allowing them to benefit from each other's strengths. Experimental results demonstrate the effectiveness of tBERT in various semantic similarity benchmarks, highlighting its potential for tasks such as information retrieval and text classification.

2.2 Latent Dirichlet Allocation [2] This seminal paper introduces Latent Dirichlet Allocation (LDA), a generative probabilistic model that has become one of the most widely used methods for topic modeling. LDA assumes that documents are generated from a mixture of topics, and each topic is a distribution over words. The authors present the underlying mathematical framework of LDA, including the generative process, inference algorithms for estimating the topic distributions, and strategies for model evaluation. LDA has been influential in shaping the field of topic modeling and serves as a baseline for comparison with other models.

2.3 Topic2Vec: Learning distributed representations of topics [3] This paper introduces Topic2Vec, a method for learning distributed representations of topics. Building upon the Word2Vec model, Topic2Vec represents topics as continuous vectors, allowing for similarity measurements and topic-based analysis. The authors propose an algorithm that incorporates topic labels and document labels to learn the representations of topics. Experimental results demonstrate the effectiveness of Topic2Vec in topic clustering, topic coherence evaluation, and topic-based document classification tasks, showcasing its potential in topic modeling and related applications.

2.4 Topic Modeling Using LDA and BERT Techniques: Teknofest Example [4] This paper explores the application of Latent Dirichlet Allocation (LDA) and BERT techniques in topic modeling, focusing on the example of Teknofest, an annual technology and innovation event. The authors apply LDA to extract topics from a large dataset of documents related to Teknofest and then utilize BERT for topic labeling and analysis. The study provides insights into the effectiveness of combining LDA and BERT for topic modeling, demonstrating their applicability in real-world scenarios.

2.5 Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models [5] This paper presents a research-related recommendation system that leverages BERT and LDA models within a semi-supervised methodology. The system aims to provide relevant and explainable recommendations to researchers by utilizing the semantic understanding captured by BERT and the topic modeling capabilities of LDA. The authors propose a semi-supervised framework that combines labeled and unlabeled data to improve the recommendation performance. Experimental results demonstrate the effectiveness of the proposed system in providing accurate and

interpretable research recommendations.

2.6 BERTopic: Neural topic modeling with a class-based TF-IDF procedure [6] This paper introduces BERTopic, a neural topic modeling technique that combines BERT with a class-based TF-IDF procedure. By utilizing the contextual embeddings of BERT and the class-based TF-IDF weights, BERTopic improves the interpret-ability and efficiency of topic modeling. The author proposes a clustering algorithm that utilizes cosine similarities and topic coherence to group similar documents into coherent topics. Experimental results demonstrate the effectiveness of BERTopic in various text mining tasks, highlighting its ability to extract meaningful and coherent topics from textual data.

These papers provide a comprehensive overview of various techniques in topic modeling, including the combination of BERT with LDA (BERTopic), optimizations and extensions to BERTopic, the use of Topic2Vec for distributed representations of topics, joint modeling of topics and syntax with Dirichlet-BERT, and dynamic topic modeling with neural embeddings. They form the foundation for our understanding of the state-of-the-art approaches in the field and inspire our work on BERT+LDA, BERT TOPIC, and Topic2Vec.

3 Methodology

In these experiments we first implemented the Topic modeling using the existing state of the art models such as LDA, and TF-IDF. Combining the strengths of LDA and TF-IDF with the contextual embeddings of BERT can lead to improved topic modeling results. BERT can help capture more nuanced relationships between words and provide richer semantic representations, resulting in more accurate and meaningful topics. We have also used BERTopic which is a topic modeling technique that leverages hugging face transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

3.1 Dataset

The data set used in this project consists of a collection of research paper abstracts. There are 14000 of these abstracts and each abstract is accompanied by its corresponding topic label. However, for the purpose of unsupervised learning, we only utilize the abstracts themselves, disregarding the topic labels. The data set covers a 31 diverse range of research domains, including computer science, Astrophysics of Galaxies, Cosmology and Non-galactic Astrophysics, Number Theory etc. The abstracts vary in length and complexity, providing a suitable basis for topic modeling experimentation.

3.2 Models

3.2.1 BERTopic

BertTopic is a topic modeling method that combines the power of pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers), with traditional topic modeling techniques. It leverages the contextual understanding and semantic representation capabilities of BERT to generate topic representations. By encoding documents into dense vector representations, BertTopic enables efficient topic inference and similarity calculation.

3.2.2 BERT + LDA

BERT + LDA combines the strengths of BERT and Latent Dirichlet Allocation (LDA). It utilizes BERT for document encoding, capturing the contextual and semantic information of the text. The encoded representations are then fed into LDA, which applies a generative probabilistic model to estimate the topic-word and document-topic distributions. BERT + LDA enhances topic modeling by leveraging the fine-grained features learned by BERT and the interpretability provided by LDA.

3.2.3 LDA (Latent Dirichlet Allocation)

LDA is a probabilistic topic modeling technique that assumes documents are generated from a mixture of topics, and each topic is characterized by a distribution over words. LDA estimates the topic-word distributions and document-topic distributions by attractively inferring latent topic assignments. It is a widely used and interpretable method for topic modeling, but it may struggle with capturing fine-grained semantic information.

3.2.4 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a classic weighting scheme used to represent the importance of a term in a document within a collection of documents. It calculates a weight based on the term's frequency in the document (TF) and its inverse document frequency (IDF) across the entire corpus. TF-IDF assigns higher weights to terms that appear frequently in a document but

are relatively rare in the corpus. It is often used as a feature representation in topic modeling and information retrieval tasks.

3.2.5 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained transformer-based language model that captures contextual and semantic information from text. It is trained on a large corpus of text data using a masked language modeling objective and a next sentence prediction objective. BERT's bidirectional architecture enables it to understand the context of a word by considering both its preceding and following words. It has achieved state-of-the-art performance on various natural language processing tasks and can be used for topic modeling by encoding documents into meaningful vector representations.

These methods offer different approaches to topic modeling, combining traditional techniques with state-of-the-art deep learning models. Experimenting with these methods can help uncover latent topics in research paper abstracts and facilitate deeper insights into the underlying themes within the data set.

3.3 Experimental Design

3.3.1 Data Preprocessing

We pre-processed the text to remove or normalize these noisy elements, making the text cleaner and easier to analyze. For sentence level pre-processing, we have inserted a period between a lowercase letter followed by an uppercase letter (e.g., "ThisIsText" becomes "This. Is. Text") and Converted the text to lowercase. Replaced "&gt" and "&lt" with a space. Reduced repeated consecutive characters to a single occurrence (e.g., "hellooooo" becomes "hello"). Reduced repeated consecutive non-word characters to a single occurrence (e.g., "!!!" becomes "!").

Replaced asterisks and combinations of asterisks with a period followed by a space. Replaced parentheses and their contents with a period followed by a space. Removed non-word characters before a period. Removed non-word characters before a period. Removed the substring "ing". Reduced repeated consecutive character sequences of length 2 or more to a single occurrence (e.g., "hellohello" becomes "hello"). Removed the substring "product received for free" followed by a period or a space. Inserts a space after a period, question mark, or exclamation mark followed by a word.

For text level pre-process removed punctuation marks from a list of words which, filters out words that consist solely of alphabetic characters, and returned the filtered words without any punctuation. From the filtered words we have used `nltk.pos_tag()` function to assign POS tags to each word and checks if the tag starts with "NN" (indicating a noun) to get list of nouns. Then corrected the typos in the nouns using the spell-checking library called sym_spell to look up suggestions for each word and selects the first suggestion with a maximum edit distance of 3. Initially, stemming was also done, but with the stemming the nouns like position is getting converted to posit which resulted in a degraded performance. Hence stemming was avoided.

3.4 Experimental Setup

We have conducted all the experiments on the google colab. Uploaded the data set to the context/sample_data folder on the colab to load the data. We have provided the commands to download the required libraries, and api's in the code for smooth execution of code. We used a sample size of 1000 and 12 topics for modeling all the models discussed above. We used the same pre-processing strategy as described in the Data Preprocessing section for all experiments. We have calculated the Coherence Score, and Silhouette Score for all the models for comparison.

3.5 Evaluation Metrics

We evaluated the topic modelling accuracy using the Coherence scores and Silhouette scores.

3.5.1 Coherence Score

Coherence measures the semantic similarity between words within a topic. Coherence scores should be used as one of multiple evaluation metrics alongside qualitative assessment and domain expertise to evaluate the quality and interpretability of topic modeling results. Higher coherence scores indicate more meaningful and coherent topics.

3.5.2 Silhouette Score

The Silhouette Score is a metric used to evaluate the quality of clusters in unsupervised learning tasks, such as clustering algorithms. It measures how well each sample in a cluster is assigned to its own cluster compared to other clusters. The Silhouette Score ranges from -1 to 1, with higher values

indicating better clustering results.

3.6 Experiments

For all the experiments used in this section we have used the data set that comprises of the abstracts of various research papers from different domains as mentioned above. And the above mentioned data preprocessing step has been carried out for every experiment that has been done.

3.6.1 TFIDF

Here, we have used the TfidfVectorizer, a class form the sci-kit_learn library that is used to convert a collection of raw text documents into a matrix of TFIDF features, which is responsible for computing the TFIDF values for each word in the sentences. Using the fit_transform method, with the input sentences as the argument, we calculated the term frequency_inverse document frequency for each word in the sentences and generated a sparse matrix representation of the TFIDF vectors as shown in Figure 1.

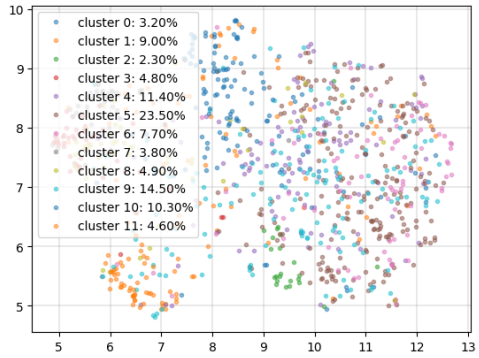

Figure 1: Scatter plot for TFIDF with K-Means

3.6.2 Latent Dirichlet Allocation (LDA)

Here, we have used LDA model using the LdaModel class from the Gensim library. The corpus parameter is a collection of documents, num_topics is the desired number of topics, and id2word is the dictionary mapping of words to their unique IDs. The LDA model is trained on the corpus by calling the train method, which internally performs the required iterations and updates to estimate the topic distributions. This will give us an array that will store the topic vectors for each document, where num_articles is the total number of documents in the corpus and k is the number of topics. The scatter plot is shown in Figure 2.

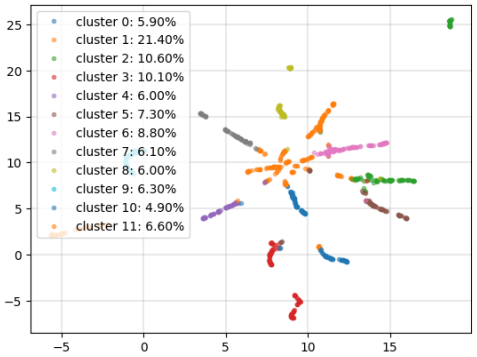

Figure 2: Scatter Plot for LDA, using K-Means

3.6.3 BERT

This experiment has been carried out using the **bert-base-nli-max-tokens** model, which is pre-trained on a large corpus of text data and is capable of producing high-quality contextualized word representations that capture the meaning of words in their specific contexts. Using the encode function, we have used a pre-trained BERT model to encode the input sentences, resulting in dense vector representations (embeddings) that capture the semantic information of the sentences.

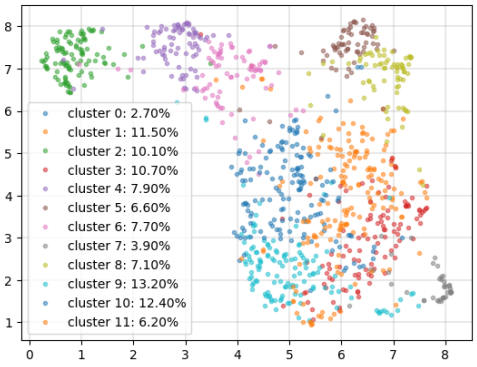

Figure 3: Scatter Plot using BERT, K-Means

The same experiment has also been carried out, using the Sentence Transformer, **"all-MiniLM-L6-v2"**. The resulting scatter plot can be seen from the Figure 4.

3.6.4 BERT + LDA

This has been carried out combining the LDA (Latent Dirichlet Allocation), and BERT (Bidirectional Encoder Representations from Transformers) vectors for a given corpus of sentences. LDA and BERT vectors for a given corpus of sentences have been computed and concatenated them, returning the combined vectors. We have also used an auto encoder to learn a lower dimensional latent space

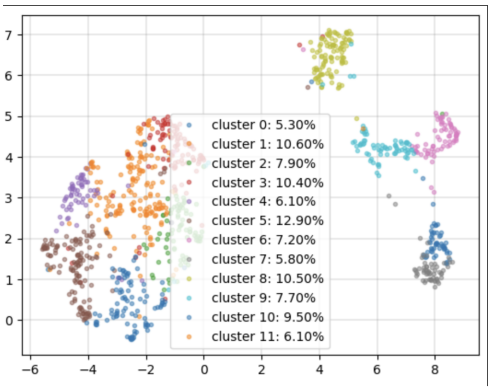

Figure 4: Scatter Plot using all-MiniLM, K-Means

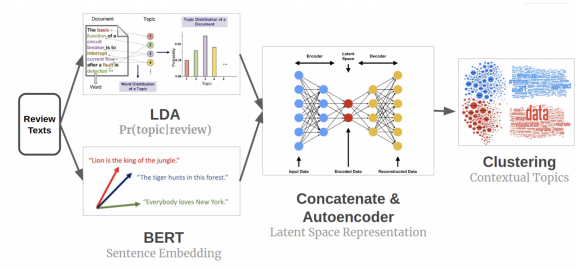

Figure 5: LDA + BERT

representation of the concatenated vector. This can be useful for tasks where both the semantic information captured by BERT and the topic information captured by LDA are desired. And this experiment has been carried out to figure out if this is beneficial when compared to the normal topic modelling techniques.

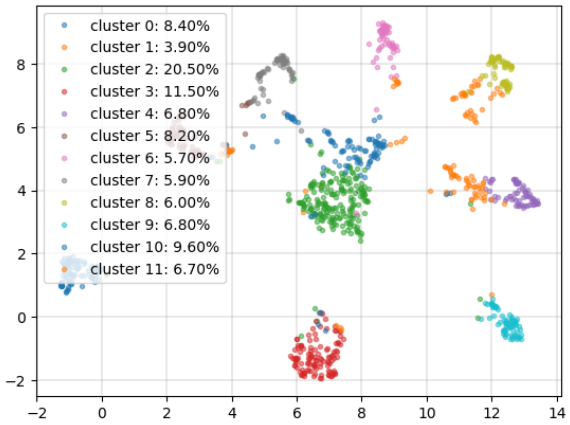

Figure 6: LDA+BERT(bert-base), with KMeans

The same experiment has also been carried out, using the Sentence Transformer, **"all-MiniLM-L6-v2"**. The resulting scatter plot can be seen from the Figure 4.

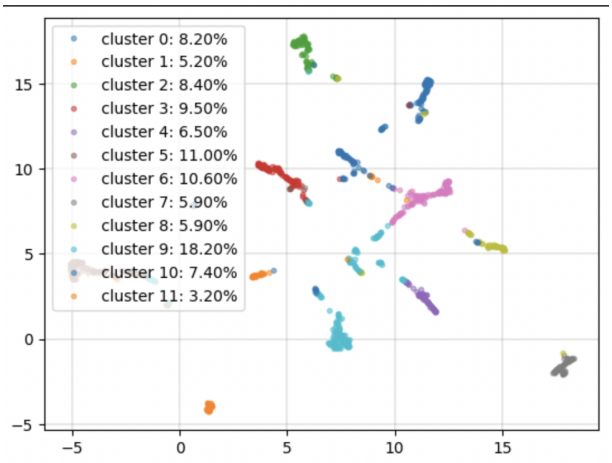

Figure 7: LDA+BERT(all-MiniLM), with KMeans

3.6.5 BERTopic

Here we have used a Sentence Transformer model named sentence_model using the **"all-MiniLM-L6-v2"** pre-trained model. It then encodes the data using the sentence_model.encode method, which generates sentence embeddings for each document in abstract. After this, we have used BERTopic with pre-trained sentence embeddings, applied UMAP for dimensionality reduction, and calculated the silhouette score for the resulting clusters. It provides an evaluation of the quality of the generated topics based on their coherence and separation in the embedding space.

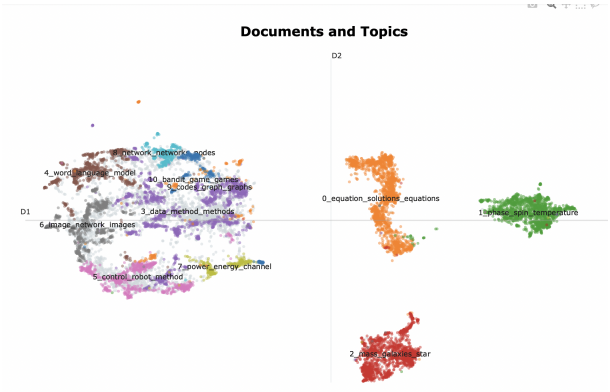

Figure 8: Results for enhanced development set. The results for English are in-language while those for Chinese and Swahili are zero-shot.

3.6.6 Coherence and Silhouette Scores

For all the above mentioned techniques, the calculated coherence scores and silhouette scores, are reported below.

Among all the models, BERTopic performs better in terms of coherence and silhouette scores.

Table 1: Calculated Scores

Model	Coherence Score	Silhouette Score
LDA	0.412	0.645
TF-IDF	0.546	0.003
LDA + BERT	0.468	0.336
Bert-base	0.525	0.046
BERTopic	0.651	0.520
BERT(all-MiniLM)	0.54	0.060

Attached below is the Inter topic distance map calculated for the clusters from the BERTopic.

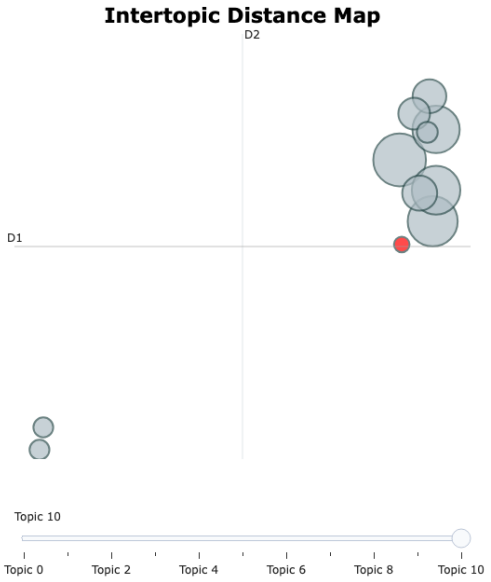

Figure 9: Intertopic Distance

Attached below is the top most words in the given clusters that are obtained form the BERTopic.

3.7 Analysis of Results

Upon analysing the results further, The LDA model achieves a moderate coherence score and a relatively high silhouette score. The coherence score indicates the degree of semantic similarity among words within the topics, where a higher score implies better coherence. The silhouette score measures the compactness and separation of the clusters formed by the topics, where a higher score indicates well-separated clusters.

The TF-IDF model achieves a higher coherence score compared to LDA, indicating better semantic similarity among words. However, the low silhouette score suggests that the clusters formed by the topics are not well-separated.

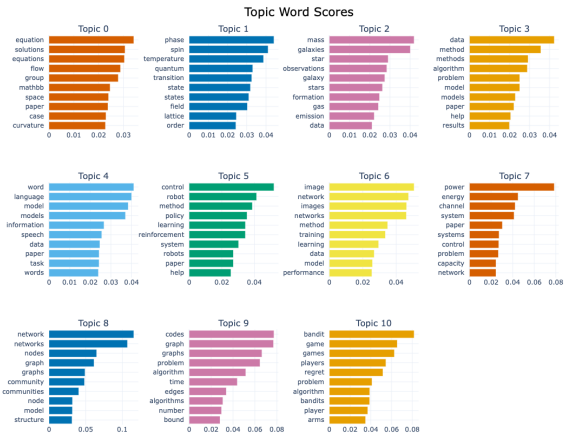

Figure 10: Topic Word Scores

The Bert-base model achieves a moderate coherence score and a low silhouette score. This indicates decent semantic similarity among words but poor separation of topic clusters. And the BERT model based on the "all-MiniLM" variant achieves a moderate coherence score and a relatively low silhouette score. The coherence score suggests reasonable semantic similarity among words, but the silhouette score indicates less separation among topic clusters.

Combining LDA with BERT embeddings leads to a slightly higher coherence score compared to LDA alone. However, the silhouette score decreases, indicating bad separation of the topic clusters.

The BERTopic model performs well in terms of coherence and silhouette scores. It achieves a high coherence score, indicating strong semantic similarity among words in the topics, and a high silhouette score, indicating well-separated topic clusters.

In summary, the analysis shows that the BERTopic model performs well in terms of coherence and silhouette scores, indicating strong semantic similarity among words and well-separated topic clusters. The TF-IDF model achieves a high coherence score but struggles with cluster separation. LDA, LDA + BERT, Bert-base, and BERT(all-MiniLM) models show varying degrees of performance in terms of coherence and cluster separation.

It is expected, that the BERT+LDA will achieve better results compared to other topic modelling techniques, but it is surprising to see that this is

slightly off in performance when compared to the BERTopic. One of the reasons, that we could think of was that combining LDA with BERT embeddings may have introduced additional noise or inconsistencies between separation of clusters which can explain the low silhouette score compared to LDA alone, but as expected LDA combined with BERT embeddings that capture the contextual information, the coherence scores are higher than that of the LDA.

4 Conclusion

In conclusion, our analysis of the different topic modeling models for research paper abstracts has provided valuable insights into their performance and strengths. The BERTopic model emerged as the top-performing model, achieving high coherence and silhouette scores, indicating strong semantic similarity among words and well-separated topic clusters. This highlights the effectiveness of leveraging BERT-based embeddings in topic modeling tasks.

Based on these findings, there are several potential directions for future work in the field of topic modeling for research paper abstracts. One can further optimize the BERTopic model. Although the BERTopic model performed well in our experiments, there is scope for further optimization and fine-tuning to enhance its performance and interpretability. One can also explore ensemble techniques. Investigating ensemble approaches that combine the outputs of multiple models could potentially improve the overall performance and robustness of topic modeling for research paper abstracts.

5 Code References:

1.https://www.kaggle.com/code/dskswu/topic-modeling-bert-lda
2.https://github.com/MaartenGr/BERTopic/issues/428
3.https://maartengr.github.io/BERTopic/index.html
4.https://chat.openai.com

References

[1] Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7047–7055, Online. Association for Computational Linguistics.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[3] Niu, Liqiang, et al. "Topic2Vec: Learning distributed representations of topics." 2015 International conference on asian language processing (IALP). IEEE, 2015.

[4] Atagün, Ercan, Bengisu Hartoka, and Ahmet Albayrak. "Topic Modeling Using LDA and BERT Techniques: Teknofest Example." 2021 6th International Conference on Computer Science and Engineering (UBMK). IEEE, 2021.

[5] Yang, Nakyeong, et al. "Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models." Expert Systems with Applications 190 (2022): 116209.

[6] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).