| **Comparative Analysis** **Topic Modeling**  CSD300 - Major Project Department of Computer Science & Engineering, IIIT Kota   *Author*  **Group No. 9**  **Priyansh Jhalora(2020KUCP1038) | Divyam Choudhary (2020KUCP1098) |**  **Saurabh Dohaiya(2020KUCP1105)**  *Mentor*  **Dr. Priyanka Mishra**    **Date of Submission - May 24, 2023** |
| --- |

| **1. Abstract**  This report presents a comparative analysis of three prominent topic modeling approaches: Latent Dirichlet Allocation (LDA), BERT (Bidirectional Encoder Representations from Transformers), and Top2Vec. The analysis explores their methodologies, strengths, and limitations, considering interpretability, scalability, and performance on various datasets. The potential for integrating these techniques is also examined to leverage their individual advantages and synergies. The findings guide researchers and practitioners in selecting the most suitable technique for topic modeling, facilitating valuable insights extraction from textual data. Understanding the nuances of these advanced approaches enhances the field of topic modeling, unlocking opportunities for knowledge discovery.  **2. Introduction**  Topic modeling is a vital technique in natural language processing (NLP) for extracting hidden themes from large text datasets. This report focuses on three prominent approaches: Latent Dirichlet Allocation (LDA), BERT (Bidirectional Encoder Representations from Transformers), and Top2Vec. LDA is a probabilistic model known for its interpretability, while BERT is a powerful language model that captures contextual word representations. Top2Vec combines document clustering and word embeddings to enhance topic modeling's interpretability and robustness. This report provides a comparative analysis of these techniques, exploring their methodologies, strengths, and potential integration for improved topic modeling results. The insights gained will aid researchers and practitioners in choosing the most suitable approach for their topic modeling needs, unlocking valuable insights from textual data.  **3. Problem**  We have address the main problem statement and definition as follows :  **3.1 Problem statement:** Topic modeling plays a crucial role in extracting meaningful insights and understanding the underlying themes within large collections of textual data. However, existing topic modeling techniques face several challenges that hinder their effectiveness and efficiency. The problem at hand is the need for an advanced topic modeling approach that can handle increasing data sets, ensure robust data cleaning, and leverage the power of state-of-the-art models like BERT.  *Visualization:*   1. Describe visualization methods to represent the LDA results. 2. Explain techniques such as word clouds, topic-document matrices, or topic networks to aid in understanding and interpretation.   *Model Evaluation and Parameter Tuning:*   1. Discuss evaluation measures for LDA, such as topic coherence or perplexity. 2. Explain strategies for selecting the optimal number of topics (K) and tuning other hyperparameters.   *Post-processing and Topic Interpretation:*   1. Extract the discovered topics from the trained model. 2. Apply post-processing techniques to enhance the interpretability of the topics, such as removing noisy or irrelevant words. 3. Assign human-readable labels to each topic based on the most representative words or documents.   *Visualization and Analysis:*   1. Utilize visualization techniques to present and analyze the discovered topics, such as word clouds, topic-document matrices, or topic networks. 2. Conduct exploratory data analysis to gain insights into the relationships between topics and documents or any other relevant patterns.   *Iterative Refinement:*   1. Iterate through the previous steps to refine the BERT topic modeling model. 2. Adjust the preprocessing steps, fine-tuning process, or model architecture based on the analysis results and domain-specific requirements.   **4.2 BERT (Bidirectional Encoder Representations from Transformers)**  *Data Preprocessing:*   1. Gather and preprocess the dataset containing the textual data for topic modeling.   *Topic Extraction and Analysis:*   1. Extract the discovered topics from the trained Top2Vec model. 2. Use techniques such as hierarchical clustering or density-based clustering to group similar documents into clusters. 3. Analyze the clusters and examine the most representative documents and words within each cluster to gain insights into the topics.   *Topic Interpretation and Labeling:*   1. Assign meaningful labels to each topic based on the analysis of the representative documents and words. 2. Apply post-processing techniques to improve the interpretability of the topics, such as removing noisy or irrelevant terms. 3. Optionally, manually review and refine the topic labels for better accuracy and relevance.   *Visualization and Exploration:*   1. Utilize visualization methods to present and explore the discovered topics, clusters, and their relationships. 2. Use techniques such as word clouds, topic-document matrices, or t-SNE visualizations to aid in the interpretation and understanding of the topics.   *Evaluation and Iterative Refinement:*   1. Consider methods like cross-validation or held-out data evaluation to assess model performance.   **4.3 : Top2Vec Model**  *Data Preprocessing:*   1. Gather and preprocess the dataset containing the textual data for topic modeling. 2. Perform common preprocessing steps such as tokenization, lowercasing, and removing stop words. 3. Consider additional preprocessing techniques like stemming or lemmatization if relevant to the specific dataset.   model. It involves dividing the dataset into multiple subsets called folds. The model is trained on a combination of folds and evaluated on the remaining fold, iteratively repeating this process for each fold. The cross-validation score is then calculated by averaging the evaluation results across all the folds. It provides a more reliable estimate of a model's performance and helps to detect issues like overfitting or underfitting.  *Silhouette Score:* The silhouette score is a metric used to evaluate the quality of clustering algorithms and assess the cohesion and separation of clusters. It quantifies how close each sample in one cluster is to samples in neighboring clusters. The silhouette score ranges from -1 to 1, where a score close to 1 indicates well-separated clusters, a score around 0 indicates overlapping clusters, and a score close to -1 suggests misclassified or poorly separated clusters. The higher the silhouette score, the better the clustering algorithm's performance.  In table 1.1 and 1.2, we can clearly see the difference each model produces when trained with the same data. See also the result section for more.  **6. Results**  The provided data allows us to draw conclusions about the three models (LDA, Top2Vec, and BERT) based on their coherence scores, success rates, and silhouette coefficients. Unfortunately, the coherence score for LDA is not available. Top2Vec demonstrates a high coherence score of 0.9822, indicating meaningful and consistent topics. BERT achieves a perfect coherence score of 1.0, implying highly coherent and easily interpretable topics. In terms of success rates, Top2Vec achieves a high score of 0.980, accurately clustering and identifying topics. When considering silhouette coefficients, LDA shows moderate cohesion and separation of clusters with a coefficient of 0.3923. Top2Vec, on the other hand,exhibits poor cohesion and separation with a low coefficient of 0.021, while BERT achieves a  utilizing distributed computing frameworks or cloud-based solutions to ensure efficient data management. Additionally, a focus should be placed on refining data cleaning techniques to address challenges such as noise, duplicates, and missing values, ensuring the integrity and quality of the dataset. Furthermore, integrating advanced techniques like BERT and LDA holds promise for improving topic modeling results. By combining BERT's contextual embeddings and LDA's probabilistic modeling, it is possible to capture both semantic relationships and latent topics within the accurate clustering, and distinguishable clusters.  **9. Appendices**    **Fig 1.1 :** Representation of clusters in BERT for K=~80.    **Fig 1.2 :** Representation of clusters in BERT for K=5.   | Factors | LDA | Top2Vec | BERT | | --- | --- | --- | --- | | Coherence Score Actual |  | 0.9822 | 1.00000000002 | | Coherence  Score Reference | 0.3197 | – | 0.3249 |   **Table 1.2 :** Coherence score of actual and referenced model for LDA, Top2Vec and BERT.  **10 References**   1. For various models of topic modeling refer to Topic Modeling with BERT by Maarten Grootendorst . 2. Research paper on LDA arXiv:1711.04305 3. Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1772–1782, Online. Association for Computational Linguistics. 4. Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. arXiv preprint arXiv:2008.09470. | This entails addressing issues related to scalability, data quality, and the integration of contextual embeddings. By addressing these challenges, we aim to develop a topic modeling solution that can accurately uncover latent topics in large and complex datasets, enabling researchers and practitioners to gain valuable insights and make informed decisions based on the extracted knowledge.  **4. Methodology**  **4.1 Latent Dirichlet Analysis (LDA)**  *Data Preprocessing:*   1. Describe the steps involved in preparing the data for LDA analysis. 2. Include procedures such as tokenization, lowercasing, stop word removal, stemming or lemmatization, and document vectorization (e.g., Bag-of-Words representation).   *Model Initialization*:   1. Explain the process of initializing the LDA model. 2. Specify the number of topics (K) to be discovered. 3. Set hyperparameters such as the Dirichlet priors for topic distributions and word distributions   *Model Inference*:   1. Outline the algorithm used for LDA model inference, such as variational inference or Gibbs sampling. 2. Describe the iterative process of estimating the posterior distribution of latent variables, including topic assignments for each document and word distributions for each topic. 3. Provide details on convergence criteria, such as maximum number of iterations or desired convergence threshold.   *Post-processing and Topic Interpretation*:   1. Present techniques for post-processing and interpreting the LDA model results. 2. Explain methods to extract meaningful topics and assign human-readable labels to them. 3. Discuss metrics or heuristics that can be used to evaluate the quality of the discovered topics. 4. Perform typical preprocessing steps such as tokenization, lowercasing, and removing stop words. 5. If required, apply additional domain-specific preprocessing techniques like stemming or lemmatization.   *Fine-tuning BERT:*   1. Obtain a pre-trained BERT model (e.g., BERT-base or BERT-large) from a reliable source. 2. Fine-tune the BERT model on the specific topic modeling task using the preprocessed dataset. 3. Convert the textual data into appropriate input formats compatible with BERT, such as tokenized sequences or sentence embeddings.   *Training the Topic Model:*   1. Design the architecture for the topic modeling model, which will utilize the fine-tuned BERT as a component. 2. Incorporate techniques such as clustering, classification, or probabilistic modeling to extract topics from the fine-tuned BERT representations. 3. Determine the number of topics (K) to be discovered and consider additional hyperparameters specific to the chosen topic modeling technique.   *Model Training and Evaluation:*   1. Train the topic modeling model using the preprocessed dataset. 2. Employ appropriate evaluation metrics to assess the performance of the model, such as topic coherence or topic diversity. 3. Fine-tune the model and hyperparameters based on the evaluation results if necessary. 4. Instantiate the Top2Vec model, which combines document clustering and word embeddings. 5. Feed the preprocessed textual data and trained word embeddings into the Top2Vec model. 6. Specify the desired number of topics (K) to be discovered and other model-specific hyperparameters.Train the Top2Vec model to learn the topic representations and their associated documents.   *Training Word Embeddings:*   1. Use a word embedding model (e.g., Word2Vec, FastText, or GloVe) to learn vector representations for words in the dataset. 2. Train the word embeddings model on the preprocessed textual data. 3. Set the hyperparameters for the word embeddings model, such as the dimensionality of the word vectors and the training algorithm.   *Top2Vec Model Training:*   1. Assess the quality of the discovered topics using evaluation metrics specific to topic modeling, such as topic coherence or topic diversity. 2. Iterate through the previous steps to refine the Top2Vec model, including adjusting hyperparameters, retraining word embeddings, or exploring different clustering techniques. 3. Validate the results with domain experts or conduct manual evaluation to ensure the topics align with the desired goals and expectations.   **5. Comparative Analysis**  **5.1 Comparison Criteria**  *Coherence Score:* The coherence score is a measure used in natural language processing (NLP) and topic modeling to evaluate the quality of a set of topics generated from a text corpus. It measures the degree of semantic similarity between the words within each topic, providing an indication of how coherent and interpretable the topics are. A higher coherence score suggests that the topics are more coherent and representative of distinct themes in the data.  *Cross-Validation Score:* Cross-validation score is a technique used in machine learning to assess the performance and generalization ability of a predictive  higher coefficient of 0.4402, indicating better cohesion and Overall, BERT outperforms LDA and Top2Vec in terms of coherence, success rate, and cluster separation, providing highly coherent topics,  data. Future work can involve exploring methodologies for aligning preprocessing steps, fine-tuning hyperparameters, and evaluating the performance of integrated BERT and LDA models. This integration has the potential to provide a more comprehensive and interpretable understanding of topics, empowering researchers and practitioners with valuable insights from ever-growing and increasingly complex datasets.  **7. Conclusion**  In conclusion, based on the evaluated factors, BERT consistently outperforms LDA and Top2Vec. BERT achieves a perfect coherence score of 1.0, indicating highly coherent topics, while Top2Vec follows closely with a coherence score of 0.9822. Both models demonstrate impressive success rates, with Top2Vec achieving 0.980 and BERT achieving a perfect 1.0, showcasing their accuracy in clustering and identifying topics. When considering cluster cohesion and separation, BERT excels with a higher silhouette coefficient of 0.4402, surpassing Top2Vec's 0.021 and LDA's 0.3923. These results emphasize BERT's superiority in topic modeling and document clustering tasks, solidifying its position as a highly effective and reliable model to consider for such applications. However, it is important to note that further evaluation and analysis may be required to assess the models comprehensively, taking into account additional metrics and data.  **8. Future Work**  Several areas can be explored to enhance the topic modeling process, considering aspects such as increasing data sets, cleaning the dataset, and integrating BERT and LDA. As datasets continue to grow in size and complexity, it becomes necessary to develop scalable strategies for handling such increasing volumes of data. This may involve    **Fig 1.3 :** Representation of word cloud by Top2Vec.    **Fig 1.4 :** Representation of word cloud by BERT.   | Factors | LDA | Top2Vec | BERT | | --- | --- | --- | --- | | Coherence Score | - | 0.9822 | 1.00000000002 | | Success Rate | - | 0.980 | 1.00 | | Silhouette Coefficient | 0.3923 | 0.021 | 0.4402 |   **Table 1.1 :** Representation of coherence score, success rate, and silhouette coefficient for LDA, Top2Vec and BERT. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |