TR-GPT-CF: A Topic Refinement Method using GPT and Coherence Filtering

Ika Widiastuti * and Hwan-Seung Yong *

Posted Date: 11 January 2025

doi: 10.20944/preprints202501.0825.v1

Keywords: Topic Refinement; Misaligned Word detection; Coherence Enhancement

Article

TR-GPT-CF: A Topic Refinement Method using GPT and Coherence Filtering

Ika Widiastuti * and Hwan-Seung Yong *

Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, Republic of Korea;
* Correspondence: ika@ewhain.net (I.W.); hsyong@ewha.ac.kr (H.-S.Y.)

Abstract: Traditional topic models are effective at uncovering patterns within large text corpora but often struggle with capturing the contextual nuances necessary for meaningful interpretation. As a result, these models may produce incoherent topics, making it challenging to achieve consistency and clarity in topic interpretation—limitations that hinder their utility for real-world applications requiring reliable insights. To overcome these challenges, we introduce a novel post-extracted topic refinement approach that uses z-score centroid-based misaligned word detection and hybrid semantic-contextual word replacement with WordNet and GPT to replace misaligned words within topics. Evaluations across multiple datasets reveal that our approach significantly enhances topic coherence, providing a robust solution for more interpretable and semantically coherent topics.

Keywords: Topic Refinement; Misaligned Word detection; Coherence Enhancement

1. Introduction and State of the Art

Topic modeling is an area of natural language processing (NLP) that employs statistical techniques to identify hidden topics or themes in documents. It is widely utilized in various disciplines to aid in extraction of patterns from large quantities of texts and documents. As a result of its ability to assist the processing of enormous amounts of text data, topic modeling is a beneficial tool in a variety of sectors, including social media analysis, market research and healthcare. This versatility across domains is one of the most notable advantages of topic modeling [1]. There is a lot of information available on the internet or other social media site, especially an immense quantity of user-generated content that can be difficult to examine [2]. Topic modeling enables businesses to better understand consumer preferences, opinions, and perceptions of their product by extracting topics of interest to a brand or company [3]. Topic modeling also helps substantially in market research and advertising campaigns since it enables businesses discover consumer behavior pattern [4], identify new trends, and customize consumer messaging to certain target markets [5].

In healthcare, topic modeling has been employed to facilitate text classification, where clinical notes are analyzed as collections of topics to enhance the categorization process [6]. Research [7] utilizes topic modeling to analyze and categorize the diverse public perceptions and sentiments related to statins. Topic modeling systematically identifies and categorizes prevalent themes and issues related to statins that have arisen from discussions. The topics encompass adverse effects, dietary interactions, hesitancy regarding statin use, and perceptions of bias within the pharmaceutical industry. Identifying these themes aids in understanding the factors that influence public acceptance and resistance to statins.

For these applications of topic modeling, the coherence and interpretability of topics are crucial. In the text classification of electronic health records, topic models enable the selection of topics as features for predictive task, which increases the interpretability of these classification models. This enhanced interpretability is critical for the clinical domain, particularly for decision-making [6]. Effective applications in information retrieval rely on a comprehensive understanding of the result derived from topic modeling [8].

 doi:10.20944/preprints202501.0825.v1

The primary topic modeling technique utilized in these applications is Latent Dirichlet Allocation (LDA) [9]. LDA is strong statistical model that identifies latent themes within a corpus by assuming that documents are mixtures of topics and topics are mixtures of words [10]. This method provides a brief overview of the document collection as well as guided exploration of the identified topics. A topic is intuitively characterized by a collection of words. Different topic may share words, and a document might be associated with multiple topics. A user can explore the themes to gain an overview of the corpus without reading all the documents and can focus on a specific topic by exploring only the text that are closely related to it. For instance, a topic with words, "computer, model, algorithm, data, mathematical" may be correlated with documents regarding computation [11].

In addition to LDA, numerous additional topic modeling methodologies are commonly employed, each possessing distinct characteristics. Non-negative Matrix Factorization (NMF) provides a linear algebraic method that generally yields more interpretable subjects by restricting topic weights to non-negative values. The Correlated Topic Model (CTM) [12] enhances LDA by permitting correlations among topics, hence providing a more accurate representation of real-world data characterized by overlapping themes. Furthermore, BERTopic utilized advanced language models like as BERT [13] to improve semantic comprehension of text, yielding highly coherent and contextually pertinent topics, hence rendering it exceptionally successful for intricate dataset with nuanced language.

However, despite their wide range of applications, these topic modeling systems frequently suffer from low coherence and inconsistent topic generation [14]. These issues are especially evident in traditional models, which have several major flaws. First, the themes provided frequently lack coherence, making it difficult for users to derive significant conclusion. Additionally, executing a single model numerous times with the same input document can produce different topics [15]. Furthermore, these models fail to capture the contextual nuance of language, which are critical for comprehending the deeper meanings inherent in texts. Finally, the restrictions in interpretability for real-world applications reduce their usefulness, as stakeholders want clear and actionable outputs in order to make informed decisions.

These issues could raise problems such as semantic irrelevance, where topics contain words that do not meaningfully relate, lack of thematic unity, where topics are broad or ambiguous and fail to represent a coherence subject, and the inclusion of outliers, where irrelevant or rare words skew topic interpretation. For instance, a topic model might incorrectly group medical term with irrelevant engineering jargon, causing confusion and decreasing the value of the research in clinical situations. Additionally, the identified topics may not necessarily correlate with human evaluations of topic quality, and may be perceived as poor from the point of view of an end user [16].

Numerous studies have been performed in order to overcome these problems, with a substantial emphasis placed on the process of refining the topics that are generated as a result of topic modeling methodologies. The majority of these studies are typically model-agnostic, which means that they do not rely on any one topic modeling technique. This model-agnostic feature enables these methods to be versatile and applicable across various contexts utilizing different topic modeling techniques. The following are major examples of such studies, first [17], explores how non-expert user engage with and refine topic models, revealing the gap between user requirements and the capabilities of present interactive topic modeling system. To learn how non-experts assess, interpret, and update topic models, the researchers performed two user studies—an in-person interview and an online crowdsourced study. They found numerous topic model modifications users wanted. User often sought to add words to explain and accentuate a topic's theme. To improve topic clarity and relevancy, irrelevant or generic terms were deleted. Changes to the word order were also considered to properly reflect the topic's concept. Users also consolidated similar topics to remove repetition and split broad topics into more specialized ones to make the model more detailed and useful. This study highlights the necessity of developing topic modeling tools that are more intuitive and correspond more closely with the methods by which non-experts typically evaluate and adjust topics. This

 doi:10.20944/preprints202501.0825.v1

approach enhances the usability of topic models while simultaneously improving their quality, aligning them more closely with user's understanding and needs.

An innovative method proposed by [18] addresses refining topic model. This approach incorporates word-embedding projections with interactive user input within the context of visual analytics framework. The method uses word embedding projections to build a visual representation of the themes. These projections are useful for displaying the semantic relationship between distinct words within a topic, providing a clearer grasp of their interconnection. By allowing users engage with the visual representations of topics, a visual analytics framework enhances this method. This interactivity enables users change these models to improve the topic depending on their personal knowledge and interpretation. This engagement including adding, removing or repositioning words within the concept space. Another study [19] also leverage user feedback to refine topic. This approach focuses on a mixed-initiative approach where both the user and the system collaborate to refine topic models dynamically. The refining process starts with the first creation of a topics applying conventional topic modeling methods. Users engage with the model by making changes and offering comments on the produced topics after the first generation of models. The system then analyzes this user feedback to learn their preference and adjust its algorithms based on user preference model. The system has six agents to support various refinement operations—combine similar topics, which identifies and combines the topics that are most similar to one another. On the other hand, the Split agent divides a topic into two distinct new topics. And the remaining agents are removing topic, reinsert small topic, reinsert outlier, and reinsert worst topic.

Additionally, research [20] presents a unique approach to refining topic models by emphasizing keyphrases, instead of specific words or whole documents. The system initially extracts keyphrases from the documents using RNN based encoder-decoder model. LDA is used as initial topic model to obtain topic words. Remove keyphrase and add keyphrase refinement function was proposed to identify documents that should be add or removed from the specified topic. Rather than changing the list of top words, the proposed method directly modifies the document-topic association by considering the keyphrases as a representative overview of the documents.

Unlike the research discussed above, the study [21] outlines a novel technique for enhancing topic models customized for short texts. This approach introduces topic refinement, a model-agnostic mechanism that utilizes the functionalities of large language model (LLMs) to enhance topics post-extraction. The method generates prompts to ask LLMs systematically over every topic. It gradually chooses a word as the possible intruder word in a topic, while the other words represent that topic. Then evaluates whether the word aligns with the semantic expression of the other words. Once alignment is verified, the word is retained, otherwise the coherence word will be provided as candidates to substitutes for the intruder word.

In this work, we introduce TR-GPT-CF, a novel approach for post-extraction topic refinement to improve coherence and interpretability. This method does not engage in the preliminary modeling of topics but focuses on enhancing topic post-extraction. Our strategy focuses on detecting the word that are not semantically related to the other word in the topic, and we named it 'misaligned word'. Then substituting them with alternatives suggested by WordNet and GPT model. The process starts by extracting topics using a topic model method such as LDA or BERTopic. From these generated topics, we detect the misaligned word using a z-score based cosine similarity of each topic word from the topic centroid. Subsequently, select the word most similar to the topic centroid as the basis for generating alternative words from WordNet and GPT model. Words with the highest coherence score will choose as candidate to replace the misaligned word. Evaluation across several datasets demonstrate that our method significantly improves topic coherence, offering an effective solution for achieving more interpretable and semantically coherent topics.

The remainder of this work is organized as follows: Initially, the proposed topic refinement method is explained, detailing the methods implemented in the study. Subsequently, performance results and discussion are presents in Section 3, and the final section concludes with an explanation of future directions.

 doi:10.20944/preprints202501.0825.v1

2. Materials and Methods

To overcome the shortcoming of conventional topic models in generating coherent and interpretable topics, we present an innovative topic refinement framework. A high-level overview of the framework is illustrated in Figure 1. Preprocessing, topic extraction, and topic refinement compose the three primary phases. Our proposed method focuses on topic refinement named TR-GPT-CF which consist of two main stage, misaligned word detection and misaligned word replacement. Misaligned word detection intends to identify the word that are not semantically related to the other word in topic. Misaligned word replacement intends to generate alternative word that are selected based on their coherence. This framework is model-agnostic, allowing it to utilize any topic modeling method and dataset. Because of flexibility, it can be used for a variety of purpose, including academic research and industry-specific analysis. Being independent of a particular model allows it to adjust to different data requirements and characteristic, increasing its usefulness and scalability across domains.

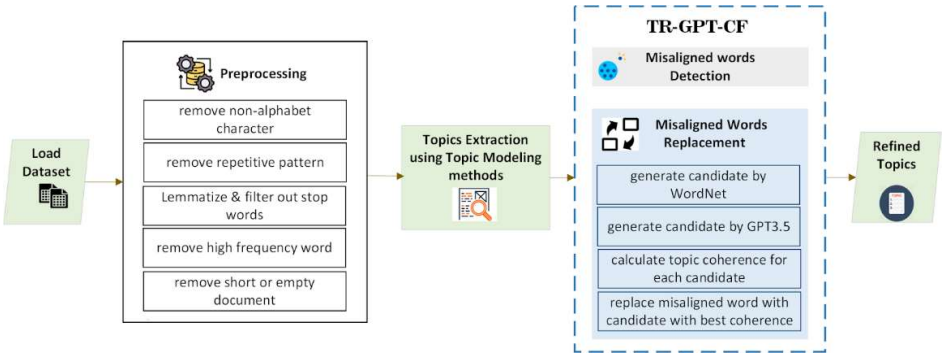

Figure 1. Comprehensive overview of the proposed topic refinement framework.

In this research, we utilize various datasets as shown in Table 1. We employ multiple datasets to evaluate the effectiveness of our refinement method across different contexts.

Table 1. Dataset summary.

Dataset	Description	Content Type	Average Length (word)	Size (document/article)
AGNews	News of articles across major topics	News articles	30	120,000
SMS Spam	Message labeled as spam or ham	Short text messages	10-20	5,574
TagMyNews	English news articles	News headlines	15-20	32,000
YahooAnswer	User-generated Q&A	Question and answer pairs	100	1,400,000
20Newsgroup	Newsgroup posts across 20 topics	Full news posts and threads	200	18,000
Kaggle's Research Article	Research articles for topic modeling exercises	Title and Abstract of Research Article	200	20,972

We sourced certain datasets from standard scikit-learn libraries, specifically AGNews, YahooAnswer, and 20Newsgroup. Specifically, this experiment utilizes the 'train' data from the AGNews and 20Newsgroup datasets, and only 20,000 training samples from the YahooAnswer dataset. The remaining three datasets—TagMyNews from [22], SMS Spam from [23], and Kaggle's Research Article from [24]—were obtained from their respective sources. These datasets provide a diverse range of topics and formats, enabling comprehensive analysis across different domains. By leveraging their distinct characteristics, we aim to derive meaningful insights and enhance our understanding of the underlying patterns within the data.

In this work we applied numerous preprocessing standards, including the elimination of stop words and lemmatization, to ensure that our model focused on the most relevant words, thereby

 doi:10.20944/preprints202501.0825.v1

enhancing the quality of the topics generated. Further preprocessing techniques such as normalization and tokenization were utilized to refine the dataset before analysis. Given the vast amount of unstructured data on the internet, particularly from social media, which predominantly consists of user-generated content, we tailored our preprocessing further. Building on the basic steps previously mentioned, we customized our dataset optimization with several specific actions, as inspired by [25], including:

- Removing high-frequency words: Identifies and removes the most frequent words in the corpus, based on TF-IDF scores. This procedure prevents the model from focusing on very common words (e.g., "the", "and")
- Eliminating repetitive pattern: Eliminate redundant sequences such as like "hahahaha" or "hihihhihi" that contribute unnecessary noise to the corpus. This procedure uses regular expression to replace any character repeated three or more times with a single occurrence.
- Removing non-alphabet character: Employs regular expression to eliminate non-alphabetic characters.
- Eliminating very short document: Exclude documents that are excessively brief or potentially consist of solely of nonsensical content, which might not be meaningful for topic modeling.

The overall pre-processing step is described in the following Algorithm 1.

Algorithm 1. Pre-processing

Input: Corpus D, stop words SW, Minimum word count min_word_count
Output: Preprocessed corpus $Dprocessed$

For each document $dm \in D$:
  convert dm to lowercase
  remove non-alphabetic characters from dm
  remove repetitive pattern from dm
  tokenize dm into words
  lemmatize each word w
  remove stop words $w \in SW$
end for
Compute high-frequency word
  vectorize D, using TF-IDF to obtain BoW
  calculate word frequencies $freq(w)$ for all words $w \in BoW$
  identify most frequent word HF
  remove high frequency words $w \in HF$
Filter short or empty document
  remove dm if $|\text{words}(dm)| < min\_word\_count$
Return preprocessed corpus $Dprocessed$

Following the preprocessing step outlined previously, we extracted a set of topics from a specific dataset using a base topic model. To facilitate this extraction, various topic modeling techniques were employed in this experiment as shown in Table 2, including LDA [9], NMF [26], BERTopic [27], and Gaussian-BAT [28]. Additionally, our framework is designed to adapts to other techniques, thereby broadening its application across a variety of scientific and industrial fields. This adaptability not only enhances the versatility of our framework but also enables customization of our studies to meet specific requirements and accommodate different datasets.

Table 2. Topic Modeling Techniques summary.

Aspect	LDA	NMF	BERTopic	G-BAT
Type of Model	Probabilistic	Matrix Decomposition	Neural (Embedding+Clustering)	Neural (VAE+Adversarial)
Input representation	Bag of Words	TF-IDF Matrix	Contextual Embeddings	Pre-trained Embeddings
Output	- Topic as word distribution	- Topic as word distributions	- Topics as ranked words based on	- Topics as latent Gaussian distributions

 doi:10.20944/preprints202501.0825.v1

	- Document topic proportions	- Document topic matrices	embeddings and clustering	- Document embeddings
Topic Representation	Topic-word	Topic-word	Cluster centers and their representative words	Latent space clusters
Strength	- Easy to implement - Interpretable	- Easy to implement - Fast and scalable	- Captures semantic relationship - Dynamic topic reduction	- Captures complex latent patterns - Robust through adversarial learning
Weakness	- Loses word order - Struggle with short or sparse text	- Loses contextual relationship - Requires TF-IDF preprocessing	- Computationally expensive - Depends on embedding quality	- Computationally expensive - Complex to train
Best Use Cases	- Long documents - Large datasets - Traditional NLP - pipelines	- Moderate size datasets	- Semantic topic modeling - Dynamic topic reduction	- Complex pattern in data - Short texts with sparse information
Application	- News categorization - Large-scale document analysis - Academic research	- Product review - Survey analysis	- Social media analysis - Short-text classification - Customer feedback analysis	- Domain specific document analysis

We then use the extracted topics from such topic modeling techniques as input in our proposed topic refinement method, TR-GPT-CF which consist of two main function, misaligned word detection which semantically detects words that deviate significantly from the centroid of the topic and misaligned word replacement, which leverages WordNet [29] and GPT-3.5 (model *gpt-3.5-turbo-1106*) from OpenAI API [30] to provide contextually appropriate replacements. The integration of these components guarantees that the refined topics exhibit enhanced coherence while preserving their interpretability for practical applications. To quickly understand the method, please see the pseudo-algorithm outlined in Algorithm 2.

2.1. Misaligned Word Detection

Prior research indicates that words unrelated to others within a topic are recognized as intruders [16, 31]. Researcher [16] conducted intruder identification by presenting users with a collection of words. The users were instructed to identify the word that was unrelated to other words in the topics. For instance, 'banana' is rapidly identified as intruder in a collection of {lion, tiger, elephant, banana, giraffe, zebra} as the other words— {lion, tiger, elephant, giraffe, zebra}—all represent animals. Conversely, in a set like {bike, professor, kangaroo, swift, green, Brazil}, which lacks this thematic unity, identifying the outlier becomes more challenging. The research [32] presents a method for detecting intruder words through the utilization of semantic similarity measures in an expanded corpus. More external data is added to the initial corpus to include a larger vocabulary, word embeddings are created to capture semantic meanings, and documents are grouped together to uncover hidden topics. Each cluster's centroid in the embedding space signifies its corresponding topic, and the low cosine similarity of intruder words to the topic centroid indicates semantic divergence.

 doi:10.20944/preprints202501.0825.v1

Although the previous paper referred to the irrelevant word as an 'intruder', in this work, we prefer to use term 'misaligned word'. This new terminology emphasizes the broader applicability and unique methodology of our technique, which extend beyond the constrains of previously established frameworks.

In contrast to the methodologies proposed in previous studies, our TR-GPT-CF mechanism utilizes the *bert-base-uncased* model of BERT embeddings [33], which are contextualized to capture the specific meaning of a words based on its surrounding words. BERT embeddings provide greater flexibility and depth in representing text, especially for task like intruder word detection in complex or ambiguous contexts. Unlike static embedding, which struggle with polysemy—for example 'bank' as financial institution versus 'bank' as the side of a river—BERT contextualized approach allows for dynamic representations that adapt to different contexts.

Our work employs the Hugging Face transformer library to apply the *bert-base-uncased* version of the BERT model, which features 12 layers, 768 hidden units. This model is pre-trained on large corpus of English text where all input words are converted to lowercase, making it case-sensitive.

Algorithm 2. TR-GPT-CF

Input: A set of topics $\boldsymbol{T} = \{t_1, t_2, \ldots, t_K\}$, Embedding model $\boldsymbol{M}$, Corpus $\boldsymbol{C}$, Large Language Model $\boldsymbol{L}$, WordNet $\boldsymbol{W}$, Z-score threshold $\boldsymbol{\theta_z}$, Inverse document frequency threshold $\boldsymbol{\theta_f}$
Output: Refine topics $\boldsymbol{T'} = \{\{t'_1, t'_2, \ldots, t'_K\}$.

Initialize the set of refine topics $\boldsymbol{T'} \leftarrow \emptyset$
For each topic $t_i \in \boldsymbol{T}$ **do**
  Initialize the refined topic $t'_i \leftarrow t_i$
    Compute the topic centroid using word embedding from $\boldsymbol{M}(t_i)$
    For each word $w_j \in t'_i$:
      Compute the centroid $c \leftarrow$ mean $\boldsymbol{M}(t_i)$
      Compute the cosine similarity s between w_j and the centroid c
      Compute Z-score z of similarities s
      Compute IDF value $IDF\ (w_{ij}, \boldsymbol{C})$
      if $z_{ij} < \boldsymbol{\theta_z}$ and $IDF\ (w_{ij}) > \boldsymbol{\theta_f}$ then
      mark w_{ij} as a misaligned word $w_{misaligned}$
    for each detected misaligned word $w_{misaligned}$ **do**
      initialize WordNet $\boldsymbol{W} \leftarrow \emptyset$
      select w_c most similar to the centroid c
      retrieve a hypernym or hyponym of w_c from WordNet $\boldsymbol{W}$
      generate via prompt $\boldsymbol{L}$ to provide alternative for w_c in the context of t_i
      combine all candidates: $\boldsymbol{W}_K \cup \boldsymbol{L}_K$
      calculate the coherence score of all candidates
      check if replacement word improves overall coherence score
        replace $\boldsymbol{w}_{misaligned}$ in t_i with $\boldsymbol{w}_{highest\_coherence\ score}$
      else
        retain $\boldsymbol{w}_{misaligned}$
    Update the refined topic $\boldsymbol{t'_i}$
    End if no further improvement in coherence is observed
End for (for all topics in $\boldsymbol{T}$)
Return the set of refined topics $\boldsymbol{T'}$

In this work, misaligned word detection is designed to analyze the semantic and statistical properties of words within a given topic in order to identify misaligned words. It starts by generating embeddings $\boldsymbol{M}$ for all words in the topic $\boldsymbol{T}$ using the *bert-base-uncased* model of BERT embeddings, and calculates a centroid embedding $\boldsymbol{M}(t_i)$, which represents the semantic center of the topic. Cosine similarity s is then computed between each word's embedding w_j and the centroid c to measure how closely each word aligned with the overall topic. Instead of relying solely on cosine similarity to detect outliers as suggested in [32], we utilize Z-score to standardize the similarity values.

 doi:10.20944/preprints202501.0825.v1

Using only cosine similarity present challenges in defining a universal threshold to classify a word as misaligned word. The range of similarity score can vary widely across different topics and datasets. For example, in some topics, all words may naturally exhibit low similarity scores due to the nature of the embeddings, complicating the identification of true misaligned word. Additionally, misaligned words often appear as 'relative' outliers—that is, their similarity to the centroid is significantly lower than that of other words in the topic. However, cosine similarity alone fails to account for the distribution of similarity scores within a topic.

Cosine similarity alone measures how semantically related each word is to the topic centroid, yet it struggles with relative comparisons across topics. Z-score address this issue by identifying words that significantly deviate from the norm within the context of the topic. This combination enables more reliable and context-aware detection of misaligned word, which are essentially the 'outlier' in the similarity distribution.

The Z-score, also known as the standard score, is a statistical measure that quantifies the number of standard deviations a data point is from the mean of a data set. It is calculated using the following formula [34]:

$$Z = \frac{(x - \mu)}{\sigma} \quad (1)$$

The numerator (x- μ) calculate the difference between the individual data point (x) and the mean (μ). This determines how far the data point is from the average value of the dataset. The standard deviation (σ) represents the typical amount by which data points differ from the mean.

In this experiment, Z-score quantify the deviation of a specific value from the mean of the similarity scores, normalized by the standard deviation. In this context, the value refers to the cosine similarity of a word to the centroid. By standardizing the similarity scores with Z-score, we normalize the variability in cosine similarity across different topics. This standardization allows us to apply the same Z-score threshold consistently, regardless of the overall range of similarity values within a given topic. We implement the Z-score using the scipy.stats.zscore function from the SciPy Python library.

To facilitate understanding of the variability in cosine similarity scores within a topic, Table 3 illustrates the scores for each word compared to the topic centroid, along with their Z-score value, assuming a standard deviation (σ) of 0.20. Word E, with a score of 0.40, exhibits significantly lower similarity to the centroid than other words. Relying exclusively on cosine similarity presents difficulties in establishing an appropriate threshold for misaligned word detection. This challenge is particularly evident when the topic demonstrates low similarities or when the dataset exhibits different behavior.

Table 3. Cosine similarity and Z-score.

Word	Cosine similarity	Z-score
Word A	0.85	0.40
Word B	0.87	0.50
Word C	0.83	0.30
Word D	0.90	0.65
Word E	0.40	- 1.85

Using Equation (1), the Z-score value is calculated to measure the extent to which each word's cosine similarity deviates from the 'average' similarity score, relative to the overall distribution of similarities. Form Table 3, we observe that most of the words have Z-score close to zero, indicating that their cosine similarities are near the mean. However, Word E has a Z-score of – 1.85, significantly lower than others, if Z-score threshold of – 1.5 is applied, Word E would be flagged as a misaligned word because its Z-score falls well below this threshold.

For several reason, Z-score provide a superior method for identifying misaligned word within topics. Firstly, they provide relativity by comparing each word's similarity to the other words in in the topic, rather than using arbitrary threshold, such as requiring a 'cosine similarity must be > 0.7'.

 doi:10.20944/preprints202501.0825.v1

This feature enhances the adaptability of Z-scores to their respective context. Secondly, the flexibility of Z-scores enables the effective application of the same threshold (e.g., - 1.5) across various topics, accommodating the unique distribution of similarities in each topic. Lastly, Z-score are particularly effective in emphasizing misaligned word like Word E, which stand out due to their significant deviations from the norm, regardless of whether the topic naturally exhibits low or high similarity values.

To further refine the detection of misaligned words, we also calculate the Inverse Document Frequency (IDF) of each word using a TF-IDF vectorizer to account for word importance. IDF introduces an element of statistical rarity, ensuring that the identified misaligned word is not only semantically distant but also unusual within the overall corpus. IDF helps emphasize less common, more relevant words while minimizes frequent, less significant ones.

Finally, we apply a threshold condition. We flag a word as misaligned word if its Z-score falls below a predefined threshold, indicating low alignment with the topic centroid, and its IDF score surpasses another threshold, indicating the word's uncommonness and potential contextual significance.

2.2. Misaligned Word Replacement

This function is designed to improve topic coherence by replacing an identified misaligned word in a topic with better alternative. It employs a dual approach for generating replacement candidates. First, it uses WordNet to identify synonyms and semantically related words based on the centroid word of the topic. Second, it leverages GPT to generate replacements that are more contextually relevant. These candidates are then combined into a unified set to ensure their uniqueness. The function evaluates each candidate by temporarily replacing the misaligned word in the topic and calculating the resulting topic coherence score. If candidate produces a higher coherence score than the original topic, it is selected as the best replacement. Additionally, the function prevents redundancy by ensuring that duplicate words are not added to the topic.

WordNet is a large lexical database of the English language that organizes words into sets of synonyms called synsets [29]. Each synset contains lemmas that are synonym or closely related terms. WordNet facilitates the exploration of lexical relationships, providing synonyms, hypernyms, and hyponym for the centroid word. Hypernyms are more general term, while hyponyms are more specific. We utilize it to generates candidates for a given centroid word. It extracts synonyms and lemmas for the centroid word to create a list of potential replacement. By querying WordNet for all synsets associated with the centroid word, we retrieve various semantic context for that word.

Consider the following workflow as an illustration: the input consists of topic words: ['education', 'learning', 'knowledge', 'teaching'], with 'learning' identified as the centroid word. WordNet's synsets for 'learning' include ['learning.n.01', 'learning.n.02'], and the lemmas derived from these synsets are ['learning', 'acquisition', 'education', 'study']. Therefore, the output for WordNet candidates is ['learning', 'acquisition', 'education', 'study'].

We extended our search for candidate words by also leveraging GPT-*3.5-turbo-1106* in addition to using WordNet. We develop a function that can be invoked within the misaligned word replacement function to generate alternative words by constructing a clear and explicit GPT's prompt. This prompt instructs GPT to provide alternative words for the centroid word, ensuring that the generated output is structured as a comma-separated list.

Our framework operates on Google Colab's T4 GPU with all necessary libraries installed. We leverage OpenAI's API version 1.55.3, setting the temperature parameter to 0 for deterministic responses to ensure minimal randomness and consistent results, and max_tokens to 100 to prevent overly long responses. We used the following prompt to query GPT-3.5 for alternative words:

"Provide alternative words for *'{centroid_word}'* in the context of the topic: *{topic_word}*. Please separate words with commas".

The function sends the prompt to GPT via the OpenAI's API using the client.chat.completions.create method.

 doi:10.20944/preprints202501.0825.v1

As previously discussed in this section, we will select the candidate with the highest coherence score to replace the misaligned word, using the topic coherence metric *Cv*, as referenced in [35]. We utilized Gensim's CoherenceModel to assess this coherence score. In this work, we focus solely on topic coherence, which encompasses a set of measures that describe the quality and interpretability of topics from a human perspective. The general formula for calculating coherence score *C* of a topic with a set of words $W = \{w_1, w_2, \ldots, w_N\}$ is:

$$C = \sum_{i=2}^{N} \sum_{j=1}^{i-1} PMI(wi, wj) \quad (2)$$

where PMI (*wi*, *wj*) is the Pointwise Mutual Information between words *wi* and *wj*, define as:

$$PMI\ (wi, wj) = \log \frac{P(wi, wj)}{P(wi)x\ P(wj)} \quad (3)$$

$P(wi, wj)$ is the probability of co-occurrence of words *wi* and *wj*, and $P(wi)$ and $P(wj)$ are the individual probabilities of the words. Calculating the PMI for all word pairs and aggregating the results yields the coherence scores, which indicates the overall semantic similarity of the words within the topic. Higher coherence score indicates more interpretable and meaningful topics.

This approach allows us to evaluate how well the identified topics align with human understanding and expectations, thereby improving the efficacy of topic modeling techniques. By focusing on topic coherence, we hope to provide results that are not only statistically sound but also meaningful and relevant to users.

3. Results and Discussion

This section evaluates the proposed TR-GPT-CF topic refinement method across six datasets—AGNews, TagMyNews, YahooAnswer, Newsgroup, SMS Spam, and Science Article—and four models: LDA, BERTopic, G-BAT, and NMF. We examine coherence score before and after applying the proposed method and benchmark these against baseline results for three shared datasets: AGNews, TagMyNews, YahooAnswer. The result demonstrates consistent improvements in topic coherence, highlighting the robustness of the proposed method across diverse datasets and models.

To validate the significance of the improvement in coherence scores, paired t-tests were conducted for each model across all datasets. The result indicate that the proposed method achieved statistically significant improvements ($p < 0.05$) for LDA, BERTopic, and G-BAT. For NMF, while the *p*-value was slightly above 0.05 ($p = 0.068$), the improvements exhibited a positive trend.

Table 4. Statistical Analysis of Coherence Score Improvement.

Model	t-statistic	*p*-value
LDA	-2.723	0.042
BERTopic	-3.491	0.017
G-BAT	-3.251	0.023
NMF	-2.318	0.068

3.1. Evaluating Topic Coherence Improvement Across Datasets

3.1.1. Improvement of Topic Coherence in AGNews Dataset

Table 5 shows the proposed refinement method significantly improved coherence scores for the AGNews datasets across all models. LDA demonstrated the highest percentage improvement at 4.4%, while BERTopic exhibit marginal improvement due to its already high baseline. From Table 4, we can see the *p*-value for LDA is 0.042, which is less than 0.05, indicating that the improvement in coherence scores after refinement is statistically significant for the LDA model.

 doi:10.20944/preprints202501.0825.v1

The modest improvement for BERTopic at 0.45% indicates that there is minimal opportunity for improvement as the model already produces highly coherent topics. Although this is in line with expectations, it may suggest that lower-performing models may benefit more from refining.

The AGNews dataset is known for having well-structured content with moderately long text and clear topics. Most model show high baseline coherence scores, which could explain why model like BERTopic and NMF show smaller gains. This underscores the significance of dataset structure in influencing refinement outcomes. In contrast, LDA, which traditionally struggles with capturing semantic coherence, benefits significantly from refinement. Similarly, G-BAT, initially one of the weaker models, also shows meaningful improvement, demonstrating the refinement method's capacity to strengthen underperforming models.

Table 5. Coherence Score for AGNews Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.591	0.617	4.40
BERTopic	0.897	0.901	0.45
G-BAT	0.453	0.471	3.97
NMF	0.771	0.790	2.45

3.1.2. Improvement of Topic Coherence in TagMyNews Dataset

Table 6 shows the refinement method demonstrated varied impacts across different model, with LDA showing the most significant improvement. Specifically, LDA's coherence score increased from 0.336 to 0.431, marking a 28.27% rise, the largest among all tested models. This substantial gain highlights the method's effectiveness in enhancing models that initially exhibit poor coherence. In contrast, BERTopic, an embedding-based model, also benefit from the refinement, though more moderately. Its score improved from 0.539 to 0.572, a 6.12% increase, suggesting that refinement techniques can effectively address challenges in datasets with shorter or noisier text like TagMyNews. Meanwhile, NMF displayed stable performance with its coherence score improving from 0.589 to 0.604, a 2.55% increase. Although modest, this improvement underscores the method's capability to boost coherence even in models that already have moderately strong baseline performance.

Table 6. Coherence Score for TagMyNews Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.336	0.431	28.27
BERTopic	0.539	0.572	6.12
G-BAT	0.646	0.650	0.62
NMF	0.589	0.604	2.55

G-BAT exhibit only a marginal improvement in its performance, with its coherence score increasing slightly from 0.646 to 0.650, a mere 0.62% increase. This minimal change suggest that G-BAT may already capture most of the coherence achievable for this dataset, indicating limited scope for further refinement. In a broader context, compared to the AGNews dataset, all models registered lower baseline scores when applied to the TagMyNews dataset. This disparity suggest that TagMyNews presents grate challenges for topic modeling, likely due to factors such as shorter text, noisier content, or overlapping topics.

3.1.3. Improvement of Topic Coherence in YahooAnswer Dataset

Table 7 shows that BERTopic demonstrated strong performance, showing a significant improvement from 0.706 to 0.745, a 5.52% increase. This substantial enhancement highlights the refinement method's ability to further improve an already high performing, embedding-based model on challenging dataset like YahooAnswer. We observed consistent improvement across all models,

 doi:10.20944/preprints202501.0825.v1

with percentage gains ranging from 3.01% for NMF to 5.52% for BERTopic, indicating the robustness and adaptability of the refinement method. Additionally, G-BAT also showcased notable gains, improving from 0.468 to 0.492, a 5.13% increase. This confirms that even models starting with lower baseline coherence can significantly benefit from the refinement process, emphasizing the method's broad applicability.

Table 7. Coherence Score for YahooAnswer Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.485	0.503	3.71
BERTopic	0.706	0.745	5.52
G-BAT	0.468	0.492	5.13
NMF	0.564	0.581	3.01

On the YahooAnswer dataset, LDA shows the smallest improvement, with a modest increase of 3.71% in coherence scores, which contrasts sharply with TagMyNews' 28.27% improvement. This difference could mean that the way YahooAnswers is structured makes it hard for probabilistic model like LDA to work, which means that efforts to improve them have less of an effect. Despite these challenges, the baseline coherence scores across all models are moderately high, suggesting that the YahooAnswer dataset generally provides a well-structured topic space. Therefore, there is some constraint on the potential for significant improvements. YahooAnswer's likely inclusion of moderately long and well-structured documents offers less of a challenge for topic models compared to more diverse datasets like TagMyNews. However, the presence of overlapping or diverse topic within YahooAnswers may still restrict further gains in coherence.

3.1.4. Improvement of Topic Coherence in Newsgroup Dataset

The refinement method has exhibited significantly but varied impacts across various models, as illustrated in Table 8. G-BAT exhibits the most significant improvement, with an extraordinary 40.19% increase in coherence score from 0.209 to 0.293.

Table 8. Coherence Score for Newsgroup Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.583	0.602	3.26
BERTopic	0.823	0.839	1.94
G-BAT	0.209	0.293	40.19
NMF	0.743	0.743	0.00

This significant achievement emphasizes the efficacy of the refinement procedure, particularly for model that initially encounter difficulties with coherence on this dataset. In contrast, BERTopic continues to demonstrate robust performance, maintaining high coherence scores that have marginally improved from 0.823 to 0.839, representing a 1.94% increase. In the interim, LDA exhibits a slight improvement, with its coherence score rising from 0.583 to 0.602, representing a 3.26% increase. This enhancement illustrates the refinement method's effectiveness in probabilistic models, such as LDA, and is particularly relevant in datasets with structured topics, such as Newsgroup.

Regarding the Newsgroup dataset, the refinement method's impact varied across models. NMF displayed no change in coherence scores, remaining at 0.743 both before and after refinement, suggesting that the refinement model method had no measurable effect on this model. This lack of improvement may indicate that NMF already captures the maximum coherence achievable for this dataset. Conversely, both LDA and BERTopic showed only small gains. The fact that LDA and BERTopic didn't make as much progress as they did on other datasets like TagMyNews suggests that the structured nature of Newsgroup, along with topics that overlap or are similar, may pose problems that make the refinement process less effective. These factors can make substantial coherence

 doi:10.20944/preprints202501.0825.v1

improvements more challenging, highlighting the complexity of enhancing topic model performance in dataset characterized by structured but similar content.

3.1.5. Improvement of Topic Coherence in SMS Spam Dataset

Table 10 indicates that the refining process has been effective across multiple models. G-BAT exhibited significant improvement, with its coherence score raising from 0.494 to 0.570, a 15.38% increase. This large enhancement demonstrates the method's capacity to improve coherence for embedding-based model in short-text dataset. Similarly, NMF showed a significant gain, with its score raising from 0.427 to 0.483, a 13.11% increase, indicating the method's effectiveness in even matrix factorization-based models within challenging datasets. On the other hand, LDA and BERTopic, had more moderate gains: LDA's coherence increase 5,64%, while BERTopic experienced a more robust improvement, a 9.09% increase. This improvement indicate that both models benefit from the refinement process, with BERTopic's stronger performance because to its embedding-based structure.

Table 9. Coherence Score for SMS Spam Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.461	0.487	5.64
BERTopic	0.506	0.552	9.09
G-BAT	0.494	0.570	15.38
NMF	0.427	0.483	13.11

The baseline scores for all models on the SMS Spam dataset are relatively low compared to those on datasets like AGNews or YahooAnswer, indicating that SMS Spam presents unique challenges for topic modeling. This difficulty is likely due to its short and informal text, which makes semantic coherence harder to achieve. Among the models, LDA shows only a modest improvement, achieving the smallest percentage gain, which underscores its limitations in handling short-text datasets. However, they also allow embedding-based models such as BERTopic and G-BAT to do well after being improved. This scenario illustrates how the characteristics of dataset can influence various topic modeling methods.

3.1.6. Improvement of Topic Coherence in Science Article Dataset

Table 10 indicates that G-BAT showed the most significant improvement among the models, with its coherence score rising from 0.265 to 0.341, representing a notable increase of 28.68%. This substantial gain underscores the refinement method's effectiveness for enhancing low-performing models on structured, domain-specific datasets such as Science Article. In contrast, BERTopic, which already had a high baseline coherence, exhibited a modest gain, improving from 0.731 to 0.740, representing a 1.23% increase. This slight improvement indicates the restricted potential for further enhancement given the model's robust initial performance. Similarly, LDA exhibited a modest improvement, with its coherence score increasing from 0.526 to 0.544, reflecting a 3.42% increase. This improvement illustrates the refinement method's ability to enhance coherence in probabilistic models, though to limited extent.

Table 10. Coherence Score for Science Article Dataset.

Model	Before Refinement	After Refinement	Improvement (%)
LDA	0.526	0.544	3.42
BERTopic	0.731	0.740	1.23
G-BAT	0.265	0.341	28.68
NMF	0.614	0.619	0.81

NMF showed moderate progress on the Science Article dataset, with its score marginally increasing from 0.614 to 0.619, a 0.81% increase. This minor change shows that the refinement

 doi:10.20944/preprints202501.0825.v1

procedure had little effect, probably because the model was already performing near its optimal level for this dataset. Similarly, aside from G-BAT, the other models only showed small improvement. This suggest that the Science Article dataset, with its well-structured, long, and technical content, probably provides a solid foundation that limits further improvement. This structure may explain why models with high baseline coherence scores, such as NMF and BERTopic, have minimal space for improvement, as the dataset's inherent characteristics already support a high level of topic coherence.

3.2. Evaluating Topic Coherence Improvements by Candidate Word Replacement: WordNet, GPT, and Combine Approaches

In this section, we explore the improvement in topic coherence across various datasets when employing different candidate word generation techniques: WordNet, GPT, and a combination of both. Each strategy has distinct benefits in aligning topic words more accurately, therefore improving the interpretability of the topics. Through the comparison of different approaches, we aim to underscore the effectiveness of each in facilitating more coherent and meaningful topics.

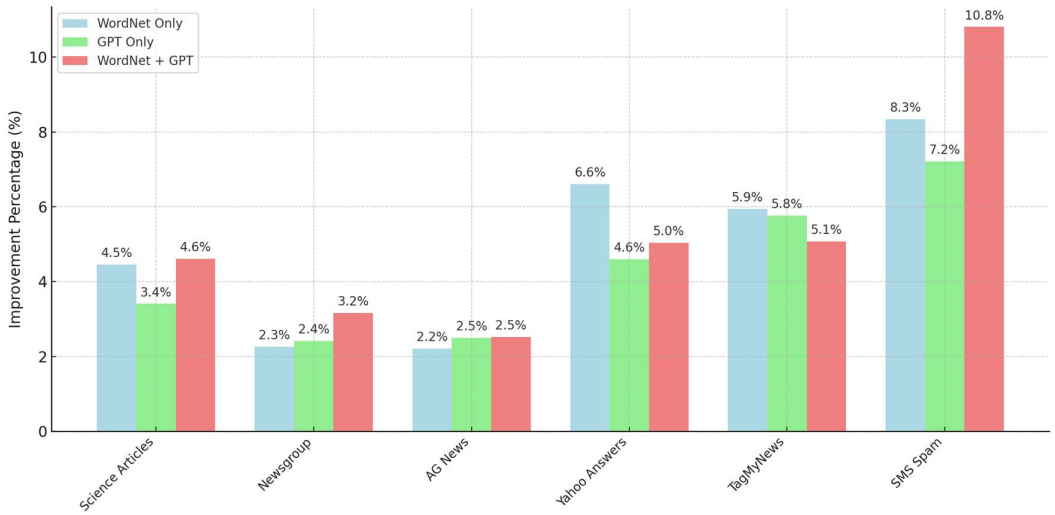

Figure 2. Topic coherence improvement by candidate word replacement: WordNet, GPT, and Combine approaches.

To improve topic coherence, WordNet and GPT were used to identify eligible words for refining. WordNet provided semantically similar candidates, such as synonyms and hypernyms, based on linguistic hierarchies, while GPT dynamically created contextually relevant alternatives using language model embeddings. This study investigates the use of WordNet and GPT in three different scenarios. First, we utilize only WordNet as generator to provide alternative words. Second, we use only GPT. Third, we employ a combination of both. This hybrid approach ensured a resilient selection of candidate words for topic refinement by combining contextual adaptability with semantic depth.

From Figure 2, we can observe that each dataset exhibits differences in the effectiveness of WordNet, GPT, and their combination in improving topic coherence. Below is the detailed analysis of these variations. Science Articles benefit from both semantic knowledge and contextual generation; therefore, the combine approaches are most effective. When WordNet and GPT are used together in this dataset, they perform slightly better than when used separately. This indicates that their combination is effective in handling more technical information. Due to their heterogeneity, Newsgroup datasets show minimal improvements across all approaches, making them difficult to enhance. The small gain is likely related to informal nature and diversity of Newsgroup dataset. AGNews demonstrates limited potential for enhancement, as the topics are already well-structured, providing limited opportunity for additional refinement. The hierarchical organization of WordNet aligns more closely with the structure of YahooAnswer dataset compared to GPT, leading to higher

 doi:10.20944/preprints202501.0825.v1

enhancement with WordNet. While both WordNet and GPT perform well individually for TagMyNews, but their combination does not yield additional benefits. WordNet demonstrates superiority in this context, highlighting its proficiency in datasets characterized by succinct and clearly delineated content. The combine approach is particularly effective for SMS Spam, as it utilizes both semantic understanding and contextual nuance to address brief, spam-related phrases.

Datasets with more structured content, such as Science Article and Yahoo Answer, tend to favor WordNet. In contrast, less structured or short-text datasets, such as SMS Spam and TagMyNews, benefit more from GPT and combined approaches. The combination of WordNet and GPT demonstrates the most significant improvements in datasets that present specific challenges, like SMS Spam. However, some datasets, such as AGNews and TagMyNews, show no significant advantage for the combine approach, as individual methods already achieve high coherence.

The comparative efficiency of WordNet, GPT, and their combined use vary by dataset. WordNet performs consistently, achieving its highest improvement of 8.3% in SMS Spam dataset, most likely due to its dependence on semantic similarity, which helps more organized context. In contrast, GPT alone perform comparably but slightly lower performance. For example, it achieves only a 3.4% improvement in the Science Articles dataset, versus WordNet's 4.5%. The combination of WordNet and GPT, however, yields the highest improvement percentages in most cases, such as 10.8% in SMS Spam and 4.6% in Science Articles. The result consistently shows that combining WordNet and GPT leads to greater improvements, implying that semantic relationships (WordNet) and contextual representations (GPT) complement each other effectively. This suggests that combining semantic clarity with contextual fluency creates a more effective mechanism for candidate's word replacement, thereby improving overall topic coherence.

3.3. Evaluating Topic Coherence by Qualitative Comparison

This section will provide a qualitative comparison of the results obtained from our experiment. In our scenario, each model is configured to extract 10 topics from each dataset, with each topic comprising 10 words. The results compare extracted topics from several topic models and demonstrates how our refinement method improves the topics by correcting misaligned words. Table 11 compares extracted topics from various topic models and highlights how refinement processes improve the topics by addressing misaligned words. It showcases how misaligned words are replaced with more appropriate terms using our refinement method.

The extracted topic column shows the initial topics generated by each model. While these topics represent the general themes, they often include misaligned words that do not fit well within topic context. After refinement, the topics are improved by replacing misaligned word with contextually appropriate replacement. The last two columns highlight the specific misaligned words identified by the misaligned word detection mechanism and their corresponding replacements. These replacements are derived using WordNet and GPT as discussed above, ensuring that the new words enhance topic coherence.

Table 11. Extracted, Refined Topic and Their Corresponding Replacement.

Dataset	Model	Extracted Topic	Refined Topic	Misaligned Word	Replacement Word
AGNews	LDA	**year**, u, sale, percent, share, cut, inc, profit, china, **report**	**sales_event**, u, sale, percent, share, cut, inc, profit, china, report	year, report	sales_event,
	NMF	president, bush, state, **afp**, election, united, Kerry, talk, john, nuclear	president, bush, state, **senator**, election, united, Kerry, talk, john, nuclear	afp	senator
	BERTopic	**tendulkar**, test, sachin, cricket, zealand,	**trial_run**, test, sachin, cricket, zealand,	tendulkar	trial_run

 doi:10.20944/preprints202501.0825.v1

	G-BAT	Australia, wicket, Nagpur, ponting, mcgrath bond, course, sale, poor, **chief**, charley, low, bay, coming, pick	Australia, wicket, nagpur, ponting, mcgrath bond, course, sale, poor, **quest** charley, low, bay, coming, pick	chief	quest
TagMyNews	LDA	world, **u**, year, job, court, star, coach, musical, john, **wednesday.**	world, **planet**, year, job, court, star, coach, musical, john, **earth.**	u, wednesday	planet, earth
	NMF	japan, nuclear, earthquake, **plant**, crisis, tsunami, radiation, stock, power, quake.	japan, nuclear, earthquake, **ionizing radiation**, crisis, tsunami, radiation, stock, power, quake.	plant	ionizing radiation
	BERTopic	trial, jury, insider, rajaratnam, guilty, former, **blagojevich**, prosecutor, lawyer, accused.	trial, jury, insider, rajaratnam, guilty, former, **prosecuting_officer**, prosecutor, lawyer, accused.	blagojevich	prosecuting_ officer
	G-BAT	yankee, south, focus, **abidjan**, shuttle, stake, Bahrain, wont, coach, nuclear	yankee, south, focus, **center**, shuttle, stake, Bahrain, wont, coach, nuclear	abidjan	center
YahooAnswer	LDA	range, x, water, **b**, weight, size, test, running, speed, force.	range, x, water, **mass**, weight, size, test, running, speed, force.	b	mass
	NMF	help, thanks, **plz**, problem, tried, yahoo, appreciated, site, computer	help, thanks, **lend a hand** problem, tried, yahoo, appreciated, site, computer.	plz	lend a hand
	BERTopic	guy, friend, love, girl, relationship, talk, boyfriend, together, he, **married**	guy, friend, love, girl, relationship, talk, boyfriend, together, he, **young_man**.	married	young_man
	G-BAT	ability, mac, common, test, time, **shes**, running, medicine, deal, maybe	ability, mac, common, test, time, **trade**, running, medicine, deal, maybe.	shes	trade
Newsgroup	LDA	line, subject, organization, writes, article, like, one, **dont**, would, get.	line, subject, organization, writes, article, like, one, **pay_back**, would, get.	dont	pay_back
	NMF	window, file, program, problem, use, application, using, **manager**, run, server.	window, file, program, problem, use, application, using, **software**, run, server.	manager	software
	BERTopic	printer, font, print, deskjet, hp, laser, ink, bubblejet, **bj**, atm	printer, font, print, deskjet, hp, laser, ink, bubblejet, **laser printer**, atm.	bj	laser printer

 doi:10.20944/preprints202501.0825.v1

	G-BAT	drive, matthew, file, dead, **clipper**, ride, pat, drug, tax, manager.	drive, matthew, file, dead, **repulse**, ride, pat, drug, tax, manager.	clipper	repulse
SMS Spam	LDA	number, urgent, show, **prize**, send, claim, u, message, contact, sent.	number, urgent, show, **correspondence**, send, claim, u, message, contact, sent.	prize	correspondence
	NMF	ill, later, sorry, meeting, yeah, **aight**, tonight, right, meet, thing.	ill, later, sorry, meeting, yeah, **match**, tonight, right, meet, thing.	aight	match
	BERTopic	lunch, dinner, eat, food, pizza, hungry, weight, eating, **lor**, menu.	lunch, dinner, eat, food, pizza, hungry, weight, eating, **selection**, menu.	lor	selection
	G-BAT	abiola, loving, **ltgt**, player, cool, later, big, waiting, regard, dude.	abiola, loving, **bed**, player, cool, later, big, waiting, regard, dude.	ltgt	bed
Science Article	LDA	state, system, phase, quantum, transition, **field**, magnetic, interaction, spin, energy.	state, system, phase, quantum, transition, **changeover**, magnetic, interaction, spin, energy.	field	changeover
	NMF	learning, deep, task, training, machine, model, feature, neural, **classification**, representation.	learning, deep, task, training, machine, model, feature, neural, **train**, representation.	classification	train
	BERTopic	logic, program, language, semantic, **automaton**, proof, calculus, verification.	logic, program, language, semantic, **reasoning**, proof, calculus, verification.	automaton	reasoning
	G-BAT	graph, space, constraint, site, integer, logic, frame, patient, diffusion, **clustering**.	graph, space, constraint, site, integer, logic, frame, patient, diffusion, **dispersal**.	clustering	dispersal

4. Conclusion

This study presents TR-GPT-CF, a novel approach for post-extracted topic refinement. It employs z-score centroid-based misaligned word detection and hybrid semantic-contextual approach for word replacement, which utilizes WordNet and GPT. To evaluate the effectiveness of our refinement method, we applied four topic modeling techniques—LDA, NMF, BERTopic, and Gaussian-BAT—across six datasets: AGNews, TagMyNews, YahooAnswer, Newsgroup, SMS Spam, and Kaggle's Science Articles. Using this four topic modeling techniques and six datasets, we evaluated the extracted topics by calculating the percentage improvement in coherence before and after applying the refinement method. In addition, we investigate the enhancement of topic coherence across six datasets through the implementation of various candidate word generation techniques, including WordNet, GPT, and a combination of both. Each strategy has its own benefits in aligning topic word more accurately, which makes the topics easier to understand. Through the

 doi:10.20944/preprints202501.0825.v1

comparison of different approaches, we hope to show how well each one assists in rendering topics more logical and significant.

TR-GPT-CF shown enhancements across all datasets. It is highly effective at refining coherence in simpler datasets with less linguistic complexity. Furthermore, it effectively improves coherence for moderately structured datasets, rendering it appropriate for semi-structured data. Building on this foundation, the combination approach of WordNet and GPT consistently provides the most significant improvements in topic coherence across diverse datasets. This is attributed to WordNet's semantic grounding and GPT's contextual adaptability. The synergy between the two addresses both semantic precision and contextual fluency, making it robust for both structured and informal datasets. The combined approach is highly recommended for challenging datasets that need both domain knowledge and contextual fluency. Individual approaches such as WordNet only or GPT only, may suffice in straightforward datasets where only one aspect—either semantics or context—is critical. This highlights the important of our work, as human evaluation of models is both costly and time-intensive, underscoring the value of our efficient, automated solutions.

Our future work will focus on further automating the review process and broadening our methodologies to encompass a wider range of datasets. We will also investigate the use of in-context learning paradigm to generate alternative words as a means to enhance the quality of GPT's responses.

Author Contributions: Conceptualization, H.-S.Y.; methodology, I.W.; software, I.W.; validation, H.-S.Y., and I.W.; formal analysis, I.W.; investigation, I.W.; data curation, I.W.; writing—original draft preparation, I.W.; writing—review and editing, I.W. and H.-S.Y.; visualization, I.W.; supervision, H.-S.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. K. Taghandiki and M. Mohammadi, "Topic Modeling: Exploring the Processes, Tools, Challenges and Applications," Authorea Preprints, Oct. 2023, doi: 10.36227/TECHRXIV.23528283.V1.
2. A. Meddeb and L. Ben Romdhane, "Using Topic Modeling and Word Embedding for Topic Extraction in Twitter," Procedia Comput Sci, vol. 207, pp. 790–799, Jan. 2022, doi: 10.1016/J.PROCS.2022.09.134.
3. H. Li, Y. Qian, Y. Jiang, Y. Liu, and F. Zhou, "A novel label-based multimodal topic model for social media analysis," Decis Support Syst, vol. 164, p. 113863, Jan. 2023, doi: 10.1016/J.DSS.2022.113863.
4. H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," Educ Inf Technol (Dordr), vol. 28, no. 5, pp. 5567–5584, May 2023, doi: 10.1007/S10639-022-11373-1/TABLES/8.
5. "Sentiment Analysis and Topic Modeling Regarding Online Classes on the Reddit Platform: Educators versus Learners | Enhanced Reader."
6. E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, and M. Spruit, "Topic Modeling for Interpretable Text Classification From EHRs," Front Big Data, vol. 5, p. 846930, May 2022, doi: 10.3389/FDATA.2022.846930/BIBTEX.
7. S. Somani, M. M. van Buchem, A. Sarraju, T. Hernandez-Boussard, and F. Rodriguez, "Artificial Intelligence-Enabled Analysis of Statin-Related Topics and Sentiments on Social Media," JAMA Netw Open, vol. 6, no. 4, p. e239747, Apr. 2023, doi: 10.1001/jamanetworkopen.2023.9747.
8. J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of Topic Models," Foundations and Trends® in Information Retrieval, vol. 11, no. 2–3, pp. 143–296, 2017, doi: 10.1561/1500000030.

9. D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, no. Jan, pp. 993–1022, 2003.
10. D. M. Blei, "Probabilistic topic models," Commun ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
11. L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," SpringerPlus 2016 5:1, vol. 5, no. 1, pp. 1–22, Sep. 2016, doi: 10.1186/S40064-016-3252-8.
12. D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," https://doi.org/10.1214/07-AOAS114, vol. 1, no. 1, pp. 17–35, Jun. 2007, doi: 10.1214/07-AOAS114.
13. Z. Fang, Y. He, and R. Procter, "BERTTM: Leveraging Contextualized Word Embeddings from Pre-trained Language Models for Neural Topic Modeling," May 2023, [Online]. Available: http://arxiv.org/abs/2305.09329
14. M. Bewong et al., "DATM: A Novel Data Agnostic Topic Modeling Technique With Improved Effectiveness for Both Short and Long Text," IEEE Access, vol. 11, pp. 32826–32841, 2023, doi: 10.1109/ACCESS.2023.3262653.
15. A. H. Marani and E. P. S. Baumer, "A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability," ACM Comput Surv, vol. 56, no. 5, Feb. 2023, doi: 10.1145/3623269/ASSET/A48A9FE3-A684-40CE-9D57-15DA3553F686/ASSETS/GRAPHIC/CSUR-2022-0908-F07.JPG.
16. J. Chang, S. Gerrish, C. Wang, J. Boyd-graber, and D. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," Adv Neural Inf Process Syst, vol. 22, 2009.
17. T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater, "The human touch: How non-expert users perceive, interpret, and fix topic models," Int J Hum Comput Stud, vol. 105, pp. 28–42, Sep. 2017, doi: 10.1016/J.IJHCS.2017.03.007.
18. M. El-Assady, R. Kehlbeck, C. Collins, D. Keim, and O. Deussen, "Semantic concept spaces: Guided topic model refinement using word-embedding projections," IEEE Trans Vis Comput Graph, vol. 26, no. 1, pp. 1001–1011, Jan. 2020, doi: 10.1109/TVCG.2019.2934654.
19. F. Sperrle, H. Schäfer, D. Keim, and M. El-Assady, "Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement," Computer Graphics Forum, vol. 40, no. 3, pp. 215–226, Jun. 2021, doi: 10.1111/CGF.14301.
20. K. M. H. Ur Rehman and K. Wakabayashi, "Keyphrase-based Refinement Functions for Efficient Improvement on Document-Topic Association in Human-in-the-Loop Topic Models," Journal of Information Processing, vol. 31, pp. 353–364, 2023, doi: 10.2197/IPSJJIP.31.353.
21. S. Chang, R. Wang, P. Ren, and H. Huang, "Enhanced Short Text Modeling: Leveraging Large Language Models for Topic Refinement," ArXiv, Mar. 2024, Accessed: Nov. 30, 2024. [Online]. Available: https://arxiv.org/abs/2403.17706v1
22. "News-Classification/train_data.csv at master · vijaynandwani/News-Classification · GitHub." Accessed: Dec. 05, 2024. [Online]. Available: https://github.com/vijaynandwani/News-Classification/blob/master/train_data.csv
23. "SMS Spam Collection Dataset." Accessed: Dec. 05, 2024. [Online]. Available: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
24. "Topic Modeling for Research Articles." Accessed: Dec. 05, 2024. [Online]. Available: https://www.kaggle.com/datasets/blessondensil294/topic-modeling-for-research-articles?select=train.csv
25. X. Wu, C. Li, Y. Zhu, and Y. Miao, "Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder," EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1772–1782, 2020, doi: 10.18653/V1/2020.EMNLP-MAIN.138.
26. W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," pp. 267–273, Jul. 2003, doi: 10.1145/860435.860485.
27. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, Accessed: Jun. 20, 2023. [Online]. Available: https://arxiv.org/abs/2203.05794v1

 doi:10.20944/preprints202501.0825.v1

28. R. Wang et al., "Neural Topic Modeling with Bidirectional Adversarial Training," Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 340–350, Apr. 2020, doi: 10.18653/v1/2020.acl-main.32.
29. G. A. Miller, "WordNet," Commun ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.
30. "API platform | OpenAI." Accessed: Dec. 17, 2024. [Online]. Available: https://openai.com/api/
31. S. Bhatia, J. H. Lau, and T. Baldwin, "Topic Intrusion for Automatic Topic Model Evaluation," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 844–849, 2018, doi: 10.18653/V1/D18-1098.
32. A. Thielmann, A. Reuter, Q. Seifert, E. Bergherr, and B. Säfken, "Topics in the Haystack: Enhancing Topic Quality through Corpus Expansion," Computational Linguistics, vol. 50, no. 2, pp. 619–655, Jun. 2024, doi: 10.1162/COLI_A_00506.
33. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Dec. 17, 2024. [Online]. Available: https://arxiv.org/abs/1810.04805v2
34. P. D. Domanski, "Statistical outlier labelling - A comparative study," 7th International Conference on Control, Decision and Information Technologies, CoDIT 2020, pp. 439–444, Jun. 2020, doi: 10.1109/CODIT49905.2020.9263920.
35. M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408, Feb. 2015, doi: 10.1145/2684822.2685324.