FIDELITY: Fine-grained Interpretable Distillation for Effective Language Insights and Topic Yielding

Divyansh Singh
University of Florida
divyansh.singh@ufl.edu

Brodie Mather
IHMC
bmather@ihmc.org

Demi Zhang
University of Florida
zhang.yidan@ufl.edu

Patrick Lehman
University of Florida
patricklehman@ufl.edu

Justin Ho
University of Florida
justinho@ufl.edu

Bonnie J. Dorr
University of Florida
bonniejdorr@ufl.edu

Abstract

The rapid expansion of text data has increased the need for effective methods to distill meaningful information from large datasets. Traditional and state-of-the-art approaches have made significant strides in topic modeling, yet they fall short in generating contextually specific and semantically intuitive topics, particularly in dynamic environments and low-resource languages. Additionally, multi-document summarization systems often struggle with issues like redundancy, scalability, and maintaining readability. We introduce FIDELITY (Fine-grained Interpretable Distillation for Effective Language Insights and Topic Yielding), a hybrid method that combines topic modeling and text summarization to produce fine-grained, semantically rich, and contextually relevant output. FIDELITY enhances dataset accessibility and interpretability, outperforming traditional models in topic diversity, similarity, and in the ability to process new, unseen documents. Additionally, it demonstrates robust multilingual capabilities, effectively handling low-resource languages like Tagalog. This makes FIDELITY a powerful tool for distilling and understanding complex textual data, providing detailed insights while maintaining the necessary granularity for practical applications.

1 Introduction

Understanding and simplifying meaningful information from large text datasets (Zadgaonkar and Agrawal, 2024) is vital for NLP tasks, e.g., sentiment analysis (Sharma et al., 2024), which gauges public opinion and trend analysis (Sivanandham et al., 2021), which identifies emerging topics in fields like technology and healthcare. Topic modeling and summarization techniques have been leveraged to uncover and present themes through word patterns and clusters of frequently co-occurring words (Blei, 2012). Such techniques are useful in organizing and summarizing legal documents (Sargeant et al., 2024) and in academic research for grouping related research papers (Asmussen and Møller, 2019).

However, existing approaches often struggle to generate semantically intuitive and contextually specific granular topics, which are necessary for text distillation. Topic modeling methods like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and BERTopic (Grootendorst, 2022) tend to generate broad, generic topics (Abdelrazek et al., 2023) or disconnected keywords leading to potential misinterpretation (Gillings and Hardie, 2022). These models also face challenges in extracting meaningful topics from new, unseen documents without retraining, limiting their usefulness in dynamic environments (Hoffman et al., 2010).

Multi-document summarization models face significant difficulties, such as redundancy, where similar information from multiple documents is repeated in the summary, leading to a lack of conciseness (Godbole et al., 2024). Furthermore, scaling to larger datasets often overwhelms these systems, leading to summaries that either omit important content (Ihsan et al., 2023) or include irrelevant information (Xiao et al., 2021).

Additionally, many multi-document summarization techniques struggle to generate summaries that are both readable and natural, often producing outputs that are disjointed or lack the fluidity of human-written summaries (Ihsan et al., 2023; Godbole et al., 2024). These challenges are compounded by the difficulty of generating coherent summaries that effectively integrate information from diverse sources, often resulting in inconsistent or difficult-to-read outputs (Xiao et al., 2021).

Prior approaches also fall short when addressing the significant challenge of processing low-resource languages. Most methods are designed for English and other high-resource languages, requiring large datasets that are scarce for low-resource

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2460–2472
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

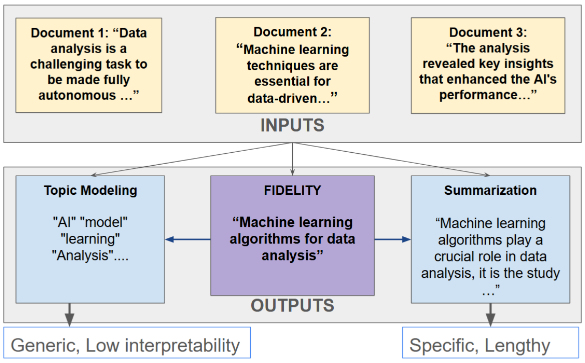
Document 1: "Data analysis is a challenging task to be made fully autonomous ..."
Document 2: "Machine learning techniques are essential for data-driven..."
Document 3: "The analysis revealed key insights that enhanced the AI's performance..."
INPUTS
Topic Modeling
"AI" "model" "learning" "Analysis"....
FIDELITY
"Machine learning algorithms for data analysis"
Summarization
"Machine learning algorithms play a crucial role in data analysis, it is the study ..."
OUTPUTS
Generic, Low interpretability
Specific, Lengthy

Figure 1: Positioning FIDELITY Between Topic Modeling and Summarization with Representative Examples

languages (e.g., Tagalog), thereby reducing performance (Medvecki et al., 2023).

We tackle these challenges with FIDELITY (Fine-grained Interpretable Distillation for Effective Language Insights and Topic Yielding), a text distillation approach that blends topic modeling and text summarization, aiming for optimal granularity. FIDELITY balances broad and specific techniques to reduce cognitive load while enhancing distillation effectiveness. Unlike traditional topic modeling or summarization, it creates semantically rich, simple, and interpretable phrasal outputs. Figure 1 illustrates how FIDELITY achieves this balance, producing text distillations that are both granular and concise. This hybrid approach ensures content is accessible yet detailed, balancing depth and clarity by combining the strengths of both techniques. FIDELITY also adapts to low resource languages, achieving Tagalog performance close to that of English.

FIDELITY's key contributions include: (1) Generating **fine-grained**, **contextually relevant** phrasal outputs with advanced topic modeling and summarization techniques; (2) outperforming state of the art in **diversity metrics** (Abdelrazek et al., 2023; Terragni et al., 2021); (3) applying LLAMA 2 (Touvron et al., 2023) to create relevant, **interpretable** phrases rather than standard keywords representations; (4) demonstrating a strong **multilingual capability** by effectively distilling phrases in Tagalog (a low-resource language), with coherence nearly on par with English; and (5) processing **new unseen documents**, leveraging refined topic clusters from previously processed datasets without retraining and identifying **multiple phrasal outputs** per document. The goal is to ensure comprehensive understanding, while maintaining the necessary granularity for practical applications.

2 Related Work

Since FIDELITY combines topic modeling and summarization, we begin by describing related work in these areas, briefly overviewing each to highlight their strengths and limitations. This provides context for our hybrid approach, which we then detail, emphasizing the rationale for combining these techniques to address existing challenges in text distillation.

2.1 Topic Modeling

Topic Modeling involves clustering documents based on shared words within a text corpus (Eklund and Forsman, 2022). LDA and Neural Topic Models (NTMs) are two popular approaches used within topic modeling (Jelodar et al., 2017; Grootendorst, 2022). LDA, a statistical topic model, is widely used for its simplicity and strong performance across domains. However, its bag-of-words foundation overlooks document semantics, limiting its performance, especially with short documents (Chang et al., 2009; Jelodar et al., 2017).

NTMs, developed to improve topic models (Miao et al., 2015), often employ pre-trained text embeddings to cluster documents and extract topics (Grootendorst, 2022; Eklund and Forsman, 2022; Sia et al., 2020). These techniques have been shown to outperform LDA (Wu et al., 2024).

While topic models excel at analyzing large document sets to extract shared topics and words, they often produce broad, generalized topics. For example, a collection of documents about dogs may be labeled as follows: [*pets, dogs, golden retriever, food*]. Topics at this level of granularity are often limited in their usefulness, highlighting the need for techniques that introduce a higher level of granularity, as adopted in FIDELITY. We note that summarization techniques, which typically yield a higher level of specificity than that of general topics, struggle in other ways that we describe next.

2.2 Summarization

In recent years, neural techniques have become the dominant approach for text summarization (Zhang et al., 2024). Architectures such as pre-trained language models (PLMs) and large language models (LLMs) have proven to be the most effective for addressing summarization challenges (He et al., 2023; Ziyu et al., 2023). Summarization is valuable for condensing the meaning of extensive texts into shorter versions, which is particularly useful

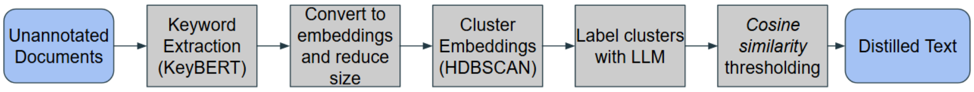
Unannotated Documents
Keyword Extraction (KeyBERT)
Convert to embeddings and reduce size
Cluster Embeddings (HDBSCAN)
Label clusters with LLM
Cosine similarity thresholding
Distilled Text

Figure 2: FIDELITY takes unannotated documents, uses KeyBERT to extract keywords, clusters with HDBSCAN to capture multiple granular themes, and employs an LLM to generate distilled text output for each cluster.

in fields with lengthy documents, such as legal and scientific domains (Polsley et al., 2016; Hayashi et al., 2023). Recent work has primarily focused on single-document summarization (Le Bronnec et al., 2024), which works well for shorter texts or when a small amount of documents can be merged.

However, single-document summarization struggles with large document volumes due to input size limitations and challenges in handling local themes (Xiao and Carenini, 2019). Multi-document summarization offers a solution that yields a single cohesive summary from multiple *related* documents (Fabbri et al., 2019), but it falls short when summaries are needed for topically *diverse* documents.

Earlier unsupervised multi-document summarization methods, like clustering and extraction techniques (Radev et al., 2004; Goldstein et al., 2000), often result in fragmented and incoherent summaries. These approaches rely heavily on extracting key sentences rather than generating new, synthesized content, making them less robust compared to modern abstractive methods that can produce more fluid and cohesive summaries (Mahajani et al., 2019). FIDELITY improves this by adopting an abstractive approach with BERT embeddings, building on the work of Reimers and Gurevych (2019), combined with a Large Language Model.

Most summarization models have a token limit; for instance, Xiao et al. (2021)'s *PRIMERA* is limited to 4,096 tokens, making it difficult to process larger datasets. By contrast, FIDELITY can handle datasets containing up to 5 million tokens, making it more effective for large datasets.

2.3 Distillation

Traditionally, the concept of distillation in NLP has been linked to tasks such as query-based distillation methods, where the focus is on synthesizing key information from multiple documents or sources based on specific queries (Florian et al., 2011). For instance, Babko-Malaya et al. (2012) define distillation as extracting atomic units of information, or "nuggets," from diverse multilingual sources in response to open-ended queries. This query-driven approach generates concise yet comprehensive outputs that are tailored to specific information needs.

Castelli et al. (2012) propose nugget-based system that extracts semantic units from a corpus for efficient information exploration. Their approach emphasizes the removal of redundancy and the generation of nuggets supported by multiple passages, showing how nugget extraction could serve various domains autonomously, without external queries. This aligns with the evolving need for more flexible and domain-agnostic distillation systems.

A method closely related to ours is the Topiary approach (Zajic et al., 2007), which augments the coverage of generic summaries by incorporating topics. While Topiary has been demonstrated to be top-performing in the Document Understanding Conference (DUC) Competition (Zajic et al., 2004), it primarily focuses on constructing headlines by combining topic terms with highly compressed sentences, by removing determiners.

Our approach builds on the distillation concepts highlighted above but avoids query-based techniques that require external inputs or topic-focused compression methods that rely on "fluency" tests. Instead, FIDELITY autonomously extracts and integrates thematic content, producing representative topic phrases that capture the core ideas of the documents while preserving their semantic richness and contextual relevance.

3 FIDELITY

Figure 2 illustrates the streamlined process for extracting and distilling key topics within documents, combining elements of topic modeling and summarization. Each step is outlined below.

3.1 Integrating Keyword Extraction with Topic Modeling and Summarization

Keyword extraction is widely used in topic modeling, and summarization. For example, BERTopic uses keyword extraction to identify terms that define topics, while TextRank employs it to highlight key parts of a document for concise summaries (Mihalcea and Tarau, 2004).

FIDELITY combines techniques from topic modeling and summarization, viewing documents as collections of n-gram key phrases that represent sub-topics, rather than isolated keywords. This approach offers richer context and captures relationships between words, resulting in a more accurate and nuanced representation of the document's content. Using a modified version of KeyBERT (Grootendorst, 2020), we extract up to 3-gram key phrases for our experiments, capturing more context and complexity than single keywords. The n-gram range is a hyperparameter adjustable for the desired granularity, since keywords with shorter ranges are suited for high-level summaries and longer ranges capturing finer sub-topics (Tahir et al., 2021).

Given a document D and a set of candidate keywords $K = \{k_1, k_2, \ldots, k_n\}$, the steps for calculating the relevance of each keyword are as follows:

1. Compute the document embedding:

$$\mathbf{E}_D = BERT(D) \tag{1}$$

2. Compute the keyword embeddings:

$$\mathbf{E}_{k_i} = BERT(k_i) \quad for\, i = 1, 2, \ldots, n \tag{2}$$

3. Calculate cosine similarity between the document embedding and each keyword embedding:

$$cosine\_similarity(\mathbf{E}_D, \mathbf{E}_{k_i}) = \frac{\mathbf{E}_D \cdot \mathbf{E}_{k_i}}{\|\mathbf{E}_D\| \|\mathbf{E}_{k_i}\|} \tag{3}$$

4. Select the top m keywords per document based on the cosine similarity scores. Initially, we choose $m = 10$ but apply a threshold function on these m keywords for each document. The overall process can be summarized by the equation:

$$\mathcal{K}_D = \{k_i \mid k_i \in K, \; score(\mathbf{E}_D, \mathbf{E}_{k_i}) \geq \tau\} \tag{4}$$

where $\mathcal{K}_D$ is the set of keywords, $score$ refers to the $cosine\_similarity$ and τ is a threshold value for the cosine similarity score.

Initially, The top 10 keywords are selected from each document, allowing each document to be associated with up to 10 distinct sub-topics. However, forcing KeyBERT to extract 10 keywords is not optimal, especially for shorter documents that may consist of a single sentence. To reduce duplication and avoid weak keyword selection we adjust the threshold (τ) for the similarity score to one-third of the maximum similarity score obtained for any keyword, i.e.,

$$\tau = \frac{\max(cosine\_similarity(\mathbf{E}_D, \mathbf{E}_{k_i}))}{3} \tag{5}$$

This additional threshold operation ensures that only keywords with significant similarity to the document are selected.

3.2 Converting Keywords to Embeddings and Dimensionality Reduction

To generate high-quality keyword embeddings, we use the `all-mpnet-base-v2`, a transformer-based model known for efficiently generating embeddings for various text lengths (Jayanthi et al., 2021) with minimal computational overhead. This model transforms text into 768-dimensional vector embeddings. We select MPNET for its computational efficiency and proven ability to capture nuanced relationships within texts, making it suitable for clustering. For more efficient clustering, we apply UMAP (McInnes et al., 2018) to reduce embedding dimensionality. UMAP preserves local and global structures, enhancing clustering cohesiveness and outperforming other dimensionality reduction methods (McInnes et al., 2018). The fitted UMAP model is stored for transforming new keyword embeddings during online processing. This conversion and reduction of embeddings is similar to the approach used in Grootendorst (2022) , where UMAP and all-mpnet-base-v2 is applied to document embeddings, whereas in our case, it is used for keyword embeddings.

3.3 Clustering using HDBSCAN

After dimensionality reduction, we apply HDBSCAN clustering (Malzer and Baum, 2020), which excels in handling varied densities and removing noise, making it ideal for keyword clustering. This is inspired by Grootendorst (2022)'s usage of HDBSCAN as their clustering algorithm.. By using a soft-clustering approach, it treats outliers as noise, uncovering implicit keyword groups that reflect underlying topics. This approach enhances document comprehension and enables more effective topic extraction by filtering out irrelevant keywords. Choosing HDBSCAN over other clustering algorithms enables us to ignore outlier keywords which do not aggregate towards a meaningful topic.

3.4 Generating Topic Representation Labels using LLAMA 2

Unlike BERTopic, which uses c-TF-IDF to generate top k topic-word distributions, we use an LLM (LLAMA 2) to name keyword clusters, as LLMs are known to produce interpretable summaries (Eppler et al., 2023), generating medical summaries with improved readability, clarity, and factual accuracy. BERTopic uses document clusters, while we use keyword clusters for extracting specific themes.

We use the LLAMA-2 70B Instruct model to generate short descriptions for each cluster, by prompting it with the cluster's top 50 keywords. The best result is obtained with this prompt, obviating the need for fine-tuning: ***Combine the following keywords into a general topic: {keyword list}. Only reply with a brief single phrase which is the topic and nothing else. Your reply should not be more than 6 words***. This approach allows us to efficiently generate meaningful labels without exceeding the model's token limit. Table 1 shows examples of keywords and their corresponding labels for the top 3 keyword clusters from the 20 News Dataset.

Prompt Keywords	**Topic Phrases**
gun hands law, know area gun, gun lobby, purpose let gun, private ownership guns...	**Gun Control and Firearm Regulations**
items cryptographic, encryption decryption based, encryption partner crime, use encryption sold, decrypt obviously won...	**Cryptography and Encryption: Uses and Regulations**
wiretapping actual wiretapping, help pay wiretap, law requires wiretaps, wiretap agency owes, current wiretap law...	**Wiretapping Laws, Usage, and Abuse**

Table 1: Examples of FIDELITY's Prompt Keywords and Topic-Phrase Output from 20 News Groups Dataset

3.5 Cosine Similarity Match and Collapsing of Similar Phrases

To evaluate the relevance of FIDELITY's distilled text to a document, we calculate cosine similarity between the text label vector $\mathbf{t}$ and the document vector $\mathbf{d}$, retaining only those pairs with a similarity of 0.4 or higher. We set this threshold at 0.4 to maintain homogeneity with the baselines, but it is a tunable hyperparameter that can be adjusted to filter stronger pairs or include weaker associations based on the desired document-topic phrase alignment (Gunawan et al., 2018). The threshold varies across embedding models due to differences in vector space properties and should be tuned to the use case for optimal performance. The filtered label-document set $\mathcal{T}'$ is defined as:

$$\mathcal{T}' = \{(\mathbf{t}, \mathbf{d}) \mid cosine\_similarity(\mathbf{t}, \mathbf{d}) \geq \tau\}$$

Redundancy is reduced by collapsing cluster labels with a cosine similarity of 0.7 or higher, selecting the most representative label. This process is repeated until no further cluster labels meet the merging criteria, resulting in a final set of distinct topic phrases. Table 2 presents some examples that are merged because they are too similar.

Collapsed Phrases
Debating the existence of God, Understanding the nature and existence of God
Gun Control and Firearm Regulations, Concealed Carry Laws and Regulations
Modems: Types, Usage, and Configuration, BPS Modem Technology and Usage

Table 2: Examples of FIDELITY's collapsed phrases for those considered too similar at a threshold of 0.7

Each final distilled output $\mathbf{T}_i$ is associated with a set of documents $\mathcal{D}_i$ and a set of representative keywords $\mathcal{K}_i$. The final output provides a comprehensive view of each distilled text phrase, including its associated documents and characteristic keywords. A document can be linked to multiple outputs due to different keywords. Mathematically, each output is represented as:

$$\mathbf{T}_i = \{\mathcal{D}_i, \mathcal{K}_i\}$$

where $\mathcal{D}_i$ is the set of documents associated with the distilled output $\mathbf{T}_i$ and $\mathcal{K}_i$ is the set of representative keywords for that output $\mathbf{T}_i$.

3.6 Distilling New Unseen Documents

Once trained, the model captures clusters associated with keywords using their embeddings, which can be used to extract relevant topic phrases from new, unseen documents without reprocessing of the corresponding dataset. The process involves preprocessing, keyword extraction, embedding, dimensionality reduction, and then mapping the reduced embeddings to pre-trained HDBSCAN clusters. Distilled text outputs are then assigned based on cosine similarity between the new document's embeddings and the predefined topics, enabling efficient and contextually relevant extraction without reprocessing the entire dataset.

4 Experimental Setup

This section outlines the experimental setup used to evaluate the performance of FIDELITY. Our computations are using 4 cores of an AMD 32-Core Processor and 2 NVIDIA A100 GPUs.

4.1 Dataset

To evaluate our pipeline, we use three datasets:

- 20 News Groups:[1] A widely used benchmark for topic classification/extraction, comprising 18,846 news articles across 20 categories.

[1] http://qwone.com/~jason/20Newsgroups/

- Trump Tweets:[2] A collection of 56,571 tweets from Donald Trump, spanning 2009 to 2021, offering a broad temporal range of topics.
- Philippine Social Media and News: A contemporary and region-specific dataset comprising 18 million records from Twitter, Reddit, Facebook, and Tumblr.[3]

The third dataset above includes broad cross-domain coverage and demonstrates a strong multilingual capability through comparisons between high-resource (English) and low-resource (Tagalog) conditions (see Section 4.3.2). This dataset contains 1,073,064 records relevant to the Philippines, focusing on themes like the South China Sea issue, Manila's relationship with the US, portraying China as a better "friend" to the Philippines than the US, corruption, and natural disasters (e.g., the 2022 earthquake, COVID-19). All the datasets are preprocessed by removing HTML tags, converting text to lowercase, and eliminating stop words. These clean, tokenized datasets serve as the basis for further analysis. A smaller holdout set of 230,244 records from late 2023 has been created for evaluating FIDELITY.

4.2 Baselines

FIDELITY is compared to two baseline topic modeling models: BERTopic and Latent Dirichlet Allocation (LDA), both well-known for topic extraction. For consistency, the `all-mpnet-base-v2` embedding model is used for both BERTopic and FIDELITY. For LDA, the number of topics is matched to those produced by BERTopic to assess the quality of topics generated by LDA. We also compare FIDELITY to the implementation by Mu et al. (2024), using their Inter-topic cosine similarity metric.

4.3 Evaluation Methods

FIDELITY operates at the intersection of summarization and topic modeling, which complicates the use of traditional metrics. Standard metrics typically applied in summarization and topic modeling are not directly applicable due to the unique nature of our approach, which combines elements of both fields. Instead, we employ specific topic modeling metrics to compare FIDELITY with baselines and demonstrate that FIDELITY generates more granular and interpretable topic-driven text distillations.

4.3.1 Evaluating Topic Phrase Granularity

Evaluating and comparing FIDELITY to models that produce topics of varying granularity presents unique challenges. Since our generated topic phrases use 3-gram keywords, metrics like traditional coherence score and perplexity (Aletras and Stevenson, 2013) do not apply in our study, as these metrics are designed for unigram or token-level evaluation. However, Section 4.3.2 provides a human evaluation of *relevance* and *coverage*, which together serve as a close proxy for automated coherence scores.

Traditional models such as BERTopic and Latent Dirichlet Allocation (LDA) tend to generate broader, more generic topics, whereas FIDELITY aims to produce a larger number of more specific topics. Due to the differing nature of these approaches, direct comparison using conventional metrics is not feasible. To effectively assess and demonstrate FIDELITY's ability to generate granular and informative topics, we apply these metrics:

Topic Diversity: We use topic diversity based on Word Embedding-based Pairwise Distance (WE-PD; Terragni 2021) to measure the diversity of words within keyword clusters by calculating pairwise distances between word embeddings. Mathematically, it is the complement of the Word Embedding-Based Pairwise Similarity (WEPS; Terragni et al. 2021). Higher WE-PD values indicate greater semantic divergence between clusters, suggesting that the keyword clusters are more distinct from each other. This is crucial for demonstrating the nuanced and detailed nature of the keyword clusters generated by FIDELITY. To accommodate 3-gram keyword clusters, we use sentence transformer embeddings instead of word embeddings.

Inter-Topic Cosine Similarity: We use inter-topic cosine similarlity (Mu et al., 2024) to assess similarity between topic phrases generated for keyword clusters. This is calculated by comparing their sentence embeddings. Lower values indicate more diverse and distinct topic phrases, showcasing FIDELITY's ability to produce a wide array of specific distilled texts with minimal overlap. Specifically, FIDELITY is evaluated on the topic phrases generated by LLAMA-2, while BERTopic and LDA are evaluated based on the top keyword in each corresponding cluster, this keyword serves as the primary representative of the cluster's theme.

[2] https://www.thetrumparchive.com/faq

[3] The Phillipine dataset distribution of records by platform is as follows: Twitter (888,130), Reddit (128,234), Facebook (55,844), Tumblr (856).

Number of Distinct Clusters Generated: The number of keyword clusters generated by a model is a complementary metric for assessing granularity. While it does not define granularity alone, when considered alongside Diversity and Similarity, it highlights the specificity of topic phrases. A higher number of clusters indicates more detailed categorization. Only BERTopic and FIDELITY generate clusters as an outcome, so we use the number of topics from BERTopic as input to LDA. Since these metrics rely on embeddings, and we perform topic collapse (see Section 3.5) based on embedding similarity, all metrics are evaluated prior to topic collapse to ensure an unbiased comparison.

4.3.2 Human and LLM-as-a-judge Evaluation: Multilingual Generation of Distilled Text

We use human judgment to evaluate FIDELITY's multilingual capability in generating topic phrases in both English (high resource) and Tagalog (low resource). A linguistics specialist conducts this evaluation to ensure linguistic and cultural accuracy. The evaluation focuses on the effectiveness of the generated phrases, which are aligned with keyword clusters (triples). Each cluster includes the top five keywords and associated context sentences, with separate assessments for Tagalog (338 distilled text outputs) and English (330).

The formal evaluation is preceded by a pilot annotation of 30 entries, conducted to familiarize the evaluator with the task and ensure reliability in the annotation process. Online dictionaries and limited Google translations are used by the evaluator to clarify meanings of words, jargon, and acronyms.

Relevance Score: This metric assesses whether each word in the distilled phrase is semantically related to at least one of the top five keywords. Relevance is scored as: Relevant (3) if four or more topic words are relevant to at least one keyword; Partially Relevant (2) if 2-3 words are related; and Irrelevant (1) if fewer than two words are related. This scoring is applied to both Tagalog and English to compare the models' multilingual performance.

Coverage Score: This metric assesses how effectively each keyword is represented by the generated phrase. Coverage is scored as: High (3) if the phrase covers four or more of the five triples; Medium (2) if it covers 2-3 triples; Low (1) if it covers fewer than two triples. This metric enables a direct comparison of the models' ability to generate comprehensive phrases in both languages.

Annotation Time: The average time taken by the annotator to evaluate each entry is recorded, providing insights into the efficiency of scaling up the study for larger multilingual datasets. To complement the human evaluation, we incorporate an LLM evaluator, specifically LLAMA 3.1 70B (Dubey et al., 2024), leveraging its state-of-the-art capabilities for natural language assessment. LLMs have demonstrated significant potential in evaluating natural language generation, often matching or surpassing human judgment in accuracy and reliability (Zheng et al., 2023).

In this study, LLAMA 3.1 evaluates semantic relevance and keyword coverage for distilled text, offering a scalable and unbiased complement to human judgment. Unlike human annotators, who assess based on the top five triples from each cluster due to task constraints, the LLM evaluates against all triples in a cluster, utilizing its higher context length to provide a more comprehensive assessment. This enhances the evaluation process for multilingual text generation in high-resource (English) and low-resource (Tagalog) settings.

4.4 Evaluating Unseen Document Distillation

To evaluate FIDELITY's ability to distill unseen documents, we conduct an experiment using 100 documents each from the 20 Newsgroups dataset and the Trump dataset. The goal is to determine the percentage of documents that FIDELITY could successfully distill. By analyzing the performance on these diverse datasets, we assess the model's robustness and versatility in handling different types of textual content, demonstrating its potential applicability across various domains.

5 Results and Analysis

Our evaluation (See Table 3) demonstrates that FIDELITY significantly outperforms traditional models like BERTopic and LDA in generating more granular topic phrases. The WE-PD metric indicates that FIDELITY achieves higher diversity values, producing more diverse and distinct topic clusters. The Inter-Topic Cosine Similarity metric (Mu et al., 2024) shows that FIDELITY achieves lower similarity values between topic phrases, suggesting minimal overlap and greater distinctiveness compared to BERTopic and LDA. Coupled with FIDELITY's ability to produce more clusters, these results shows its more granular distillation.

Enhanced granularity and depth provide a more

	20 News Groups			Trump		
	WE-PD	**Cosine Similarity**	**Topics**	**WE-PD**	**Cosine Similarity**	**Topics**
BERTopic	0.707	0.1981	193	0.698	0.2064	575
LDA	0.689	0.2213	193	0.628	0.3143	575
FIDELITY	**0.862**	**0.0686**	**1424**	**0.792**	**0.1464**	**2304**

Table 3: FIDELITY Comparison: Diversity, Inter-Topic Cosine Similarity, and Number of Topic Clusters

comprehensive understanding of the data, making large datasets more accessible and easier to interpret, without being generic, thus offering a clearer and more distilled analysis compared to traditional topic models. The quality of topic phrases generated by FIDELITY shows clear improvement over Mu et al. (2024), as indicated by inter-topic cosine similarity metrics. While Mu et al. (2024) report increasing similarities with more topics—Top 10 (0.155), Top 20 (0.197), and Top 30 (0.203)—FIDELITY consistently achieves lower scores: FIDELITY(0.1498, 0.1309, 0.1309). This suggests FIDELITY produces more distinct and less redundant topics, enhancing output quality overall.

Model	**20 News Groups**	**Trump**
BERTopic	0.350	0.399
LDA	0.197	0.261
FIDELITY	**0.460**	**0.482**

Table 4: FIDELITY's Average Topic Phrase-Document Similarity using the `all-minilm-l12-v2` (Liu et al., 2020) for computing embeddings

Table 4 further underscores FIDELITY's effectiveness in aligning distilled texts with document content. The average cosine similarity score between FIDELITY's topic phrases and their corresponding documents is significantly higher than that of LDA and BERTopic, indicating a closer phrase/document alignment. This likely stems from FIDELITY's ability to produce fine-grained, contextually specific topics that better capture the key themes of the documents.

A human evaluation of both English and Tagalog word triples reveals FIDELITY's promising multilingual capabilities (see Table 5). For the English data, the coverage score is near ceiling (average 2.97), indicating FIDELITY's strong ability to cover broad concepts. The relevance score is lower (averaged 2.2), suggesting that some words in the distilled texts are not relevant to the keywords.

The lower relevance score for Tagalog is primarily due to occasional "hallucinations" introduced by the underlying Generative AI machinery (described as an *innate limitation* in Xu et al. 2024b). For example, the system generates the text "Beards and their impact on relationships and society" for the keyword cluster ['beard', 'grow beard', 'join growing beard', 'neck beards', 'bearding']. This phrase is scored as irrelevant (1) because the "impact on relationships and society" part is hallucinated and unsupported by any of the keywords. Further discussion of these hallucinations is provided in the Limitations section.

Metric	**Human**		**LLM**	
	Tagalog	**English**	**Tagalog**	**English**
Relevance	1.92	2.20	2.86	2.96
Coverage	2.78	2.97	2.63	2.87

Table 5: Evaluation Scores for the Topic Phrases Generated by FIDELITY on English and Tagalog Triples

The average coverage score is relatively high in Tagalog (2.78), despite six of the outputs receiving a penalty of zero. This demonstrates FIDELITY's ability to distill highly comprehensive texts in low-resource languages like Tagalog. The zeros are assigned to abnormal topic phrases that do not appear in Tagalog (discussed further in the Limitations section). Occasional hallucinations also occur in the Tagalog data, further complicated by cross-language accuracy issues, resulting in slightly lower relevance scores (average 1.92).

LLM evaluations showed higher relevance and coverage scores compared to human evaluations in both English and Tagalog. For relevance, the scores were high for both English (2.96) and Tagalog (2.86). For coverage, the scores were also higher for LLM evaluations, with English (2.87) slightly outperforming Tagalog (2.63). However, the scores were not strongly correlated with human evaluations, with mean Pearson correlation scores of 0.24 for English and 0.19 for Tagalog. This is likely due to LLMs assessing all keywords in a cluster, whereas human evaluations considered only the top five.

The histogram (See Figure 3) shows the distribution of Topic Phrases per document for both the 20 News Group dataset and the Trump dataset. This reveals that a significant number of documents have more than two topic phrases, demonstrating FIDELITY's ability to identify multiple relevant themes within a single document.

Despite this, the average distilled outputs per

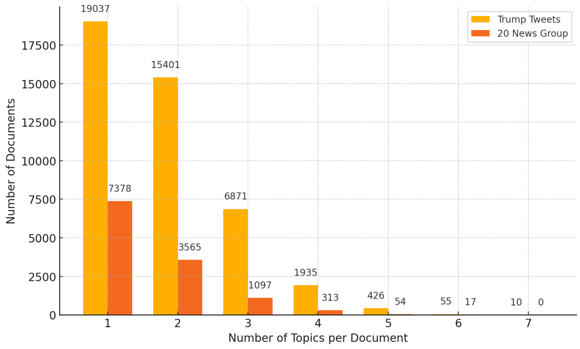

Figure 3: Histogram illustrating Number of Topics Identified Per Document for 20 News Group Dataset

document remain moderate at 1.56, demonstrating FIDELITY's effectiveness in extracting multiple themes without overwhelming users. It efficiently processes large datasets, such as the Philippine Social Media and News dataset (230,000 documents, 5.5 million tokens), generating around 6,000 distilled phrases. The top topic phrase is linked to 5,382 distinct documents.

6 Conclusion

FIDELITY offers a unique approach to distill and make large text datasets more accessible. Unlike traditional models that often produce broad, generic topics, FIDELITY generates fine-grained, contextually specific themes, balancing depth and clarity. This ensures that the content is both detailed and easy to interpret, making FIDELITY particularly valuable for applications requiring a nuanced understanding of large datasets.

The results show that FIDELITY generates more distinct and diverse themes with minimal overlap compared to models like LDA and BERTopic. This enhanced granularity enables a more comprehensive and accurate data analysis. FIDELITY's strong performance in the low-resource language Tagalog highlights its robustness and adaptability.

By extracting multiple relevant themes within documents and balancing diversity and simplicity, FIDELITY provides a more comprehensive and effective approach to analyzing large datasets. This capability not only makes it easier to identify key patterns and insights, but also ensures that the information remains accessible and manageable. This serves it as a powerful tool for distilling complex information, enabling users to efficiently read, understand, and apply the insights from their data.

Limitations

Despite its strengths, FIDELITY has a few limitations. Reliance on cosine similarity as the primary metric for keyword extraction, evaluation and topic collapse has its drawbacks. Cosine similarity, while effective for measuring vector similarity, does not consider vector magnitude and can struggle in high-dimensional spaces, potentially leading to the incorrect merging of distinct topics (Zhou et al., 2022; Steck et al., 2024). The process of collapsing similar topic labels based on cosine similarity may further oversimplify distinctions between closely related topics, causing a loss of nuance. Future work could explore leveraging large language models (LLMs) for merging similar topics, albeit with an increased computation cost (Xu et al., 2024a), as well as investigating alternative semantic similarity metrics to address these limitations.

Additionally, we use the `all-mpnet-base-v2` model, primarily trained on English data. This can result in inaccurate embeddings for multilingual data, particularly in low-resource languages like Tagalog. Its token length limit of 512 truncates longer documents could result in severe loss of information and context. Future work could explore models such as Wang et al. (2023), which support higher token limits, to address this issue and maintain computational efficiency. While a multilingual embedding model could address the language issue, and using models with higher token limits could better handle longer texts, these adjustments might come at the expense of reduced performance or increased computational complexity.

Another limitation of our approach is the occasional generation of hallucinated content, where the resulting topic phrases include words unrelated to the input keyword cluster. This is due to reliance on underlying LLM machinery, where hallucinations are always a possibility Xu et al. (2024b).

While FIDELITY demonstrates robust performance across multiple datasets and languages, we have not provided a case study to illustrate its performance in a real-world scenario. Our future work aims to assess FIDELITY's practical applicability in a complex, real-world context, such as Social Media Analysis. FIDELITY could enrich insights into perspectives and attitudes embedded within data when integrated with sentiment algorithms (Naskar et al., 2016).

A notable limitation is our reliance on a small number of baselines and an older LLM, LLAMA

2. Incorporating newer state-of-the-art models like GPT 4o (Achiam et al., 2023) with reduced hallucinations and better performance could enhance FIDELITY and improve text distillation. These long context LLMs, could serve as baselines and handle larger datasets efficiently. Integration and comparison with long-context LLMs will be a focus of our future work.

FIDELITY relies on predefined clusters established during training and cannot extract distillations for new topics outside the training data. This limits its adaptability in dynamic domains, as accommodating new topics would require reprocessing the model to include additional clusters, impacting efficiency. An area for future work is an exploration of dynamic clustering techniques, which adapt to evolving data and automatically adjust clusters in real-time, enabling the model to handle newer topics without reprocessing.

A further challenge lies in the low correlation between LLM and human evaluations (Mean Pearson correlation: 0.24 for English and 0.19 for Tagalog). This discrepancy stems from differing evaluation approaches: LLMs evaluate all keywords in a cluster, whereas humans focus only on the top five. The use of LLMs to evaluate outputs generated by another LLM introduces the risk of inflated scores, reflecting inherent biases in the LLM-as-a-judge approach (Wataoka et al., 2024). Future work should explore alternative evaluation frameworks to better align with human assessments.

Ethics Statement

Our work complies with the policies of the Twitter API and Meta LLaMA Terms and Conditions. No personal or sensitive information has been collected. We analyze user-generated Twitter content without interacting with users.The potential identifiers linked to the Twitter data include usernames or Twitter handles, which are publicly accessible to anyone with internet access.

The Tagalog Annotation assessment is carried out by a linguistics doctoral student who is not involved in algorithm development and is compensated at a standardized stipend rate. This study has been IRB-reviewed and has been determined to qualify as "nonhuman research."

All datasets, except the Philippine Social Media and News dataset, are publicly available, along with all three FIDELITY's outputs.[4]

Acknowledgement

This work is supported, in part, by DARPA Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):1–18.

Olga Babko-Malaya, Greg P Milette, Michael K Schneider, and Sarah Scogin. 2012. Identifying nuggets of information in gale distillation evaluation. In *LREC*, pages 2322–2327. Citeseer.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

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

Vittorio Castelli, Hema Raghavan, Radu Florian, Ding-Jung Han, Xiaoqiang Luo, and Salim Roukos. 2012. Distilling and exploring nuggets from a corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1006–1006.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[4] https://drive.google.com/drive/folders/12plW-Y1EtA4UdnTBKmOvOs6cS7rHB8iz?usp=sharing

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Michael B Eppler, Conner Ganjavi, J Everett Knudsen, Ryan J Davis, Oluwatobiloba Ayo-Ajibola, Aditya Desai, Lorenzo Storino Ramacciotti, Andrew Chen, Andre De Castro Abreu, Mihir M Desai, et al. 2023. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. *Urology practice*, 10(5):436–443.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Radu Hans Florian, Joseph Olive, Caitlin Christianson, John McCary, Radu Hans Florian, Connie Fournelle, Olga Babko-Malaya, Connie Fournelle, Olga Babko-Malaya, Radu Florian, et al. 2011. Distillation. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 617–743. Springer.

Mathew Gillings and Andrew Hardie. 2022. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities*, 38(2):530–543.

Aditi Godbole, Jabin Geevarghese George, and Smita Shandilya. 2024. Leveraging long-context large language models for multi-document understanding and summarization in enterprise applications. *arXiv preprint arXiv:2409.18454*.

Jade Goldstein, Vibhu O Mittal, Jaime G Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 workshop: automatic summarization*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Dani Gunawan, CA Sembiring, and Mohammad Andri Budiman. 2018. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series*, volume 978, page 012120. IOP Publishing.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's new? summarizing contributions in scientific literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-code++: A pre-trained language model optimized for abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5112, Toronto, Canada. Association for Computational Linguistics.

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 856–864, Red Hook, NY, USA. Curran Associates Inc.

Uswa Ihsan, Humaira Ashraf, and NZ Jhanjhi. 2023. Survey on multi-document summarization: Systematic literature review. *arXiv preprint arXiv:2312.12915*.

Sai Muralidhar Jayanthi, Varsha Embar, and Karthik Raghunathan. 2021. Evaluating pretrained transformer models for entity linking in task-oriented dialog. *arXiv preprint arXiv:2112.08327*.

Hamed Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. 2017. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *CoRR*, abs/1711.04305.

Florian Le Bronnec, Song Duong, Mathieu Ravaut, Alexandre Allauzen, Nancy Chen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2024. LOCOST: State-space models for long document abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1144–1159, St. Julian's, Malta. Association for Computational Linguistics.

W. Liu, X. Li, W. Yang, Y. Lin, X. Liu, S. Wang, C. Xie, L. Xu, W. Zhuang, X. Zhao, and L. Li. 2020. Minilm: Deep self-attention distillation for tiny transformers. *arXiv preprint arXiv:2002.10957*.

Abhishek Mahajani, Vinay Pandya, Isaac Maria, and Deepak Sharma. 2019. A comprehensive survey on extractive and abstractive techniques for text summarization. *Ambient Communications and Computer Systems: RACCCS-2018*, pages 339–351.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI)*, pages 223–228. IEEE.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Darija Medvecki, Bojana Bašaragin, Adela Ljajić, and Nikola Milošević. 2023. Multilingual transformer and bertopic for short text topic modeling: The case of serbian. In *Conference on Information Technology and its Applications*, pages 161–173. Springer.

Yishu Miao, Lei Yu, and Phil Blunsom. 2015. Neural variational inference for text processing. *CoRR*, abs/1511.06038.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.

Debashis Naskar, Sidahmed Mokaddem, Miguel Rebollo, and Eva Onaindia. 2016. Sentiment analysis in social networks through topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 46–53.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Holli Sargeant, Ahmed Izzidien, and Felix Steffek. 2024. Topic modelling case law using a large language model and a new taxonomy for uk law: Ai insights into summary judgment. *arXiv preprint arXiv:2405.12910*.

Neeraj Anand Sharma, ABM Shawkat Ali, and Muhammad Ashad Kabir. 2024. A review of sentiment analysis: tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, pages 1–38.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

S Sivanandham, A Sathish Kumar, R Pradeep, and Rajeswari Sridhar. 2021. Analysing research trends using topic modelling and trend prediction. In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1*, pages 157–166. Springer.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 887–890.

Noman Tahir, Muhammad Asif, Shahbaz Ahmad, Muhammad Sheraz Arshad Malik, Hanan Aljuaid, Muhammad Arif Butt, and Mobashar Rehman. 2021. Fng-ie: an improved graph-based method for keyword extraction from scholarly big-data. *PeerJ Computer Science*, 7:e389.

Silvia Terragni. 2021. Topic model diversity. Accessed: 2024-09-05.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–45. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2).

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Shaochen Xu, Zihao Wu, Huaqin Zhao, Peng Shu, Zhengliang Liu, Wenxiong Liao, Sheng Li, Andrea Sikora, Tianming Liu, and Xiang Li. 2024a. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Ashwini Zadgaonkar and Avinash J Agrawal. 2024. An approach for analyzing unstructured text data using topic modeling techniques for efficient information extraction. *New Generation Computing*, 42(1):109–134.

David Zajic, Bonnie J. Dorr, and Richard Schwartz. 2004. Bbn/umd at duc-2004: Topiary. In *Proceedings of the 2004 Document Understanding Conference (DUC 2004) at NLT/NAACL 2004, Boston, MA*, pages 112–119.

David M. Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard M. Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Inf. Process. Manag.*, 43(6):1549–1570.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *Preprint*, arXiv:2406.11289.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. *arXiv preprint arXiv:2205.05092*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.