

Original Paper

Identifying Public Response Topics to the Centers for Disease Control and Prevention's COVID-19 Communications on Social Media: Infoveillance Study Using Large Language Model-Based Rephrasing

Wangjiaxuan Xin¹, MSc; Shuhua Yin², PhD; Albert Park³, PhD; Yaorong Ge⁴, PhD; Shi Chen^{2,4}, PhD

¹College of Computing and Informatics, The University of North Carolina at Charlotte, Charlotte, NC, United States

²Department of Public Health and Health Administration, The University of North Carolina at Charlotte, Charlotte, NC, United States

³School of Health Information Science, University of Victoria, Victoria, BC, Canada

⁴School of Data Science, The University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:

Wangjiaxuan Xin, MSc
College of Computing and Informatics
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223
United States
Email: wxin@charlotte.edu

Abstract

Background: Public health agencies increasingly use social media platforms such as Twitter (subsequently rebranded X, X Corp) to monitor public responses to official health communications and support timely decision-making during health crises. Public replies to communications from official agencies provide valuable insights into population-level discourse and engagement with health policies. However, analyzing large-scale short-text responses remains challenging because social media replies are often brief, informal, ambiguous, and linguistically fragmented. Topic modeling offers a scalable approach for identifying thematic patterns in such data, but its performance is often limited when applied to social media short texts. Recent advances in large language models (LLMs) provide new opportunities to improve text normalization before topic modeling, but their value and limitations for public health infoveillance remain underexplored.

Objective: This study develops and evaluates TM-Rephrase, a model-agnostic LLM-based rephrasing framework designed to improve the performance of topic models for short public health social media texts while ensuring rephrasing preserves the original semantic meaning, stance or intent, and tone.

Methods: We analyzed 25,027 public replies to official Centers for Disease Control and Prevention tweets on Twitter collected between May 2020 and November 2022. TM-Rephrase transformed informal and noisy short texts into more standardized expressions using 2 prompt-guided schemes: general rephrasing and colloquial-to-formal rephrasing. Original and rephrased texts were analyzed using multiple topic models, with rephrasing generated by Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct. Topic quality was evaluated using coherence, uniqueness, redundancy, and diversity. We also conducted an expert semantic-fidelity validation study in which 4 expert raters evaluated whether rephrased texts preserved the original meaning, stance or intent, and tone.

Results: TM-Rephrase generally improved topic quality across models, rephrasing schemes, and LLM backbones. For latent Dirichlet allocation, C_v coherence increased from 0.3094 without rephrasing to 0.5004 with colloquial-to-formal rephrasing. For BERTopic (BERT-based topic modeling), C_v coherence improved from 0.4078 to 0.4734. Diversity-related metrics also improved in most rephrasing settings. Additional checks using C_{NPMI} , C_{UCI} , and U_{MASS} generally supported the main improvement pattern. Sensitivity analysis across different topic numbers further suggested that the coherence gains were not limited to the main setting. Expert validation showed favorable semantic preservation overall, with a mean rating of 3.90 (SD 1.01) and median of 4.00 (IQR 3.00-5.00) of 5.00, and acceptable interrater reliability, intraclass correlation coefficient = 0.736. General rephrasing received stronger semantic-fidelity ratings than colloquial-to-formal rephrasing, suggesting that stronger

formalization may improve topic readability while introducing greater risk of changes to the original meaning, stance or intent, and tone.

Conclusions: TM-Rephrase can enhance the coherence, distinctiveness, and interpretability of topic modeling outputs for short, noisy public health social media data. However, it is a text-normalization preprocessing strategy for topic enhancement rather than a fully neutral replacement for original public discourse.

JMIR Infodemiology 2026;6:e98319; doi: [10.2196/98319](https://doi.org/10.2196/98319)

Keywords: text mining; topic modeling; public health infoveillance; COVID-19; social media analytics; large language models

Introduction.

Background

The wide adoption of social media has fundamentally transformed public discourse, particularly during public health crises (eg, the COVID-19 pandemic) [1]. Twitter, characterized by rapid information dissemination and a concise conversational format, has emerged as a critical yet methodologically challenging data source for understanding and analyzing public discussion topics and the dynamics of information exchange during the COVID-19 pandemic [2]. In addition to enabling real-time communication, these platforms support large-scale public engagement, where individuals actively express opinions, share experiences, and respond to official health communications. For public health agencies such as the US Centers for Disease Control and Prevention (CDC), analyzing short-text public replies that express opinions and concerns provides valuable insights for informing more effective communication strategies to the public [3]. Such innovative surveillance and analyses can help capture evolving public concerns, assess responses to current health policies, and improve the alignment between institutional messaging and public needs. However, from a public health perspective, effectively leveraging these data remains challenging because of the unstructured nature of short, user-generated social media texts, which complicates the systematic monitoring and timely interpretation of population-level responses. In particular, the inherent brevity, informality, and ambiguity of individual tweets, along with the 280-character limit [4], pose significant challenges for traditional social media data mining and analytics. These characteristics are particularly problematic for topic modeling, which aims to identify latent thematic structures in textual corpora, because limited textual context constrains reliable semantic interpretation.

The contextual sparsity of social media short texts often leads to topics of keywords that are less coherent and less diverse (ie, more overlapping topics with redundant or unrelated keywords), thereby reducing interpretability and limiting their utility for understanding various public opinions during the pandemic. This limitation is further compounded by the heterogeneous nature of online public discourse, where variations in writing and expression styles introduce additional complexity into textual data. This problem becomes particularly pronounced when attempting to capture emerging and latent thematic structures in communications during public health emergencies, where semantic ambiguity,

brevity, and informal linguistic styles (eg, abbreviations, slang, layman expressions, and misspellings) are prevalent [5]. Moreover, even when topic models identify clusters of topic keywords, they often lack meaningful interpretability without additional contextualization or external knowledge from experts, making it difficult to translate model outputs into coherent narratives for public health decision support. Consequently, these challenges limit the effectiveness of topic modeling in supporting systematic analysis of public health discourse. These issues suggest the need to standardize and formalize short social media texts into more contextually formal content before the application of topic modeling in health communications, thereby improving semantic clarity and enhancing the interpretability of downstream analytical results.

Topic modeling, as a widely used method for discovering latent thematic structures in textual corpora, is exemplified by the classic probabilistic model, latent Dirichlet allocation (LDA), proposed by Blei et al [6] in 2003. However, despite its broad adoption, applying LDA or its variants to short texts from social media platforms poses substantial challenges. The brevity, informality, and noise of tweets weaken statistical signals and exacerbate the context sparsity, leading to incoherent, redundant, or less interpretable topics generated from these probabilistic models [7].

To address these challenges, some enrichment strategies have been proposed. One line of research augments short texts with external knowledge or auxiliary contextual signals to enhance statistical robustness [8]. Another direction leverages distributed representations of words and documents, incorporating embeddings or neural architectures to map texts into richer semantic representations. Prominent examples include BERTopic (BERT-based topic modeling) [9] and Top2Vec [10], which use more recent transformer-based embeddings and clustering methods to generate more semantically meaningful topics. Recent work has explored more advanced neural network architectures, such as FASTopic (fast, adaptive, stable, and transferable topic model) [11] and topic-semantic contrastive learning [12], which demonstrate improved handling of context sparsity and have the potential to be applied in health communications.

However, all these approaches focus on algorithmic model redesign with parameterized inference strategies for short texts, which may adapt poorly to real-life large-scale data (eg, public health-related discourse on social media). In addition, such topic modeling algorithms typically output sets

of top-ranked keywords that can be difficult for analysts or readers without additional domain knowledge to interpret.

Regarding the public discourse during COVID-19, the pandemic fundamentally transformed social media into the primary platform for real-time public discourse, where Twitter has been extensively used to study public attitudes and evolving concerns toward the pandemic and various health policies [13,14]. Scholars have analyzed public reactions to official communications from health agencies such as the CDC, focusing on opinion and thematic feedback to health policies [3,15]. Additional research has explored geospatial and demographic variations in public concerns, demonstrating the heterogeneous impact of the pandemic on different communities [16]. Despite the active research and the value of social media for effective public health intelligence during emergencies, it often overlooks the interpretability issues arising from short and noisy texts in online discourse, especially on social media platforms.

Motivated by Maini et al [17], we leverage large language models (LLMs) as a text preprocessing approach for rephrasing short texts for topic modeling, aiming to standardize and formalize public replies to CDC tweets on Twitter. This framework, termed TM-Rephrase, builds on recent advances in LLMs, which have demonstrated powerful performance across diverse text-processing tasks, including text summarization [18], sentiment analysis [19], and information retrieval [20].

TM-Rephrase differs from conventional text preprocessing approaches. Standard text normalization and spelling correction primarily address surface-level noise, such as misspellings, inconsistent casing, punctuation, or informal abbreviations. Data augmentation usually creates additional training examples to improve model learning. In contrast, TM-Rephrase uses LLM-based, prompt-constrained rephrasing to transform each short social media reply into a more standardized and contextually explicit expression for topic modeling. Its distinguishing characteristic lies in treating rephrasing as a model-agnostic, data-centric preprocessing strategy for improving topic-model outputs rather than as a replacement for topic modeling algorithms or as generic text rewriting.

In this study, we examine the impact of TM-Rephrase on the quality of topics derived from 4 representative topic models, using a dataset of public replies to CDC Twitter tweets during the COVID-19 pandemic. We hypothesize that rephrasing original short-text public replies on social media through TM-Rephrase enhances topic quality in terms of both intratopic semantic coherence and inter-topic metrics of topic model outputs, which are represented as sets of keywords. Specifically, we investigate two rephrasing schemes: (1) general rephrasing and (2) colloquial-to-formal rephrasing, and evaluate their effectiveness across multiple

topic modeling algorithms, including probabilistic algorithms (eg, LDA [6]) and more advanced neural network-based topic models (eg, BERTopic [9] and FASTopic [11]).

Research Questions and Objectives

The research questions (RQs) of this study are as follows: (RQ1) whether and how effectively TM-Rephrase improves the quality and interpretability of topic model outputs for short texts in health communications, (RQ2) how different rephrasing schemes influence topic model performance, and (RQ3) how to interpret public discussions (ie, replies to CDC during the COVID-19 pandemic) based on the topics and inform public health practice.

This study aims to improve the analysis of public responses to official public health communications on social media to support more effective public health surveillance. To achieve this, we develop and evaluate TM-Rephrase, an LLM-based model-agnostic framework that enhances the quality and interpretability of topic modeling for short-text public replies to CDC communications during the COVID-19 pandemic.

Methods

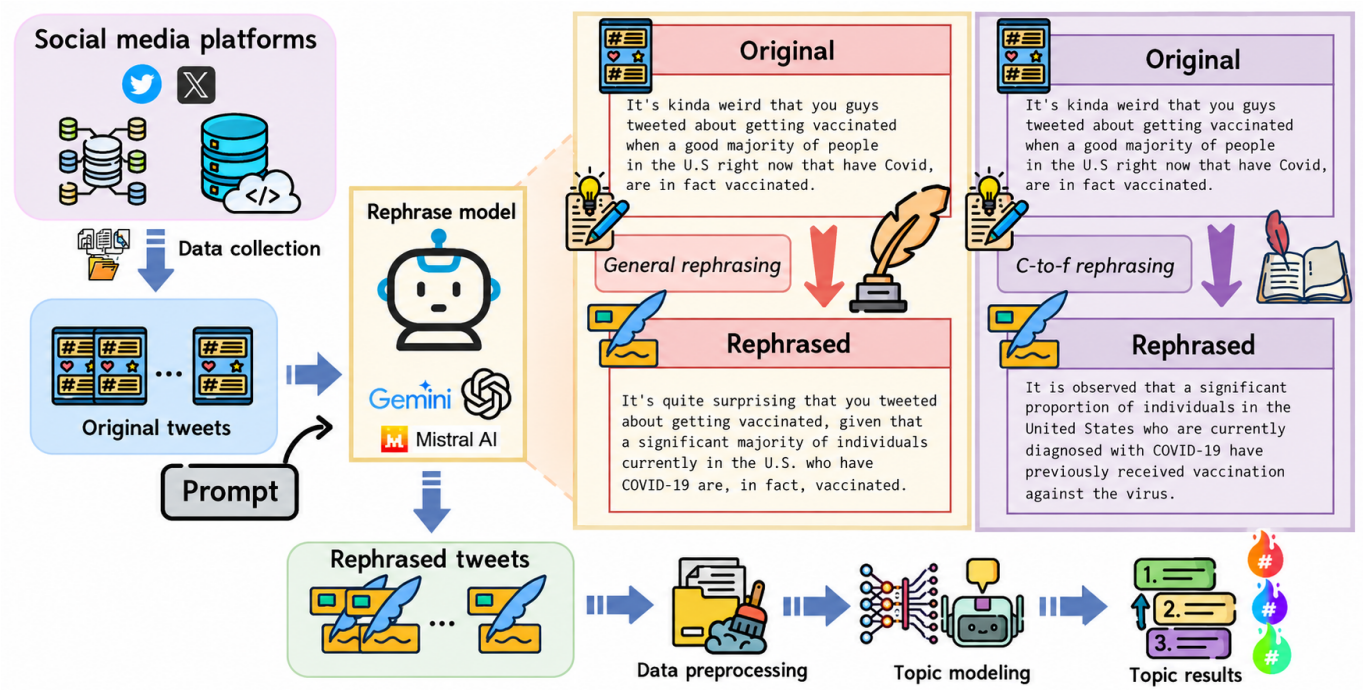
Study Design

TM-Rephrase is designed as a multistage pipeline to improve topic modeling performance in public health communications, as shown in Figure 1. The pipeline is composed of four major stages: (1) data collection, (2) LLM-based rephrasing, (3) data preprocessing, and (4) topic modeling with a selected algorithm and evaluation of the resulting topics, as detailed in the following subsections. We used this pipeline to study how rephrased text of public responses to CDC tweets on Twitter impacts topic modeling outcomes compared to nonrephrased ones.

Before introducing the model details, we begin by formulating the problem setting with preliminary notations that are used throughout this paper. Let $D = \{d_1, d_2, \dots, d_N\}$ denote a corpus of N short-text tweets (ie, replies to CDC tweets). Each document d_i is typically brief and context-limited because of the inherent sparsity of social media discourse.

The goal of topic modeling in this setting is to extract and model a set of K latent topics, denoted by $T = \{T_1, T_2, \dots, T_K\}$, where each topic T_k is represented by a set of its top- T representative keywords, extracted or generated by the corresponding topic modeling algorithm, that is, $T_k = \{w_1, w_2, \dots, w_T\}, k = 1, 2, \dots, K$. These representative keywords serve as the semantic expression of the topic and form the basis for downstream topic interpretation, evaluation, and further analysis.

Figure 1. Overview of the TM-Rephrase framework. The pipeline integrates data collection, LLM-based rephrasing, data preprocessing, and topic modeling algorithm application with evaluation.



Data Collection

We constructed a dataset of public responses to official CDC communications on Twitter during the COVID-19 pandemic, covering the period from May 2020 to November 2022. Data were collected via the Twitter Academic API and search queries (Section 1: data search queries in [Multimedia Appendix 1](#)), tracking 7 official CDC accounts: @CDCgov, @CDCDirector, @CDCGlobal, @CDCMMWR, @CDCtravel, @DrReddCDC, and @CDCEmergency. The initial retrieval yielded 180,090 public interactions associated with these CDC accounts, including replies and direct mentions. The retrieved metadata included tweet posting dates, tweet text, tweet account ID, account type, public engagement metrics, referenced tweet type, and referenced tweet IDs. Amazon Web Services was used as the computing or storage environment for managing the collected data. For the purpose of this study, only tweets that could be linked to an original CDC tweet were retained. Approximately 49,000

direct mentions that referenced a CDC account but were not replies to a specific CDC tweet were removed. For example, a tweet that mentions @CDCEmergency independently from a user’s own account was not considered a reply to the CDC communication. Another 104,000 reply-like interactions were further excluded as they could not be reliably linked to an accessible original CDC tweet, primarily because the original CDC tweet was deleted, unavailable, or otherwise inaccessible at the time of data processing. This step was necessary to preserve conversational context between each public reply and the corresponding CDC communication.

The final analytic dataset consisted of 25,027 public replies linked to 1,512 unique original CDC tweets. These replies capture public opinions, concerns, and discussions in response to official CDC communications during a major public health emergency. We retained only English-language tweets. Statistics of the final original and rephrased datasets are presented in [Table 1](#).

Table 1. Statistics of the original and rephrased datasets of public replies to CDC^a communications during the COVID-19 pandemic.

Quantity of original and rephrased datasets	Original	General rephrased	C-to-f ^b rephrased
Mean number of words	27.33	27.76	28.14
SD of word counts	14.23	15.12	14.11
Minimum number of words	1	1	1
25th percentile number of words	15	16	16
Number of words, median (IQR)	27 (15-40)	27 (16-41)	29 (16-42)

Quantity of original and rephrased datasets	Original	General rephrased	C-to-f ^b rephrased
75th percentile number of words	40	41	42
Maximum number of words	64	66	68

^aCDC: US Centers for Disease Control and Prevention.

^bC-to-f: colloquial to formal.

Formally, the dataset $D = \{d_1, d_2, \dots, d_N\}$ denotes the original corpus, where each short-text reply is denoted as:

$$d_i = (w_1^i, w_2^i, \dots, w_{L_i}^i), i = 1, 2, \dots, N \quad (1)$$

which contains a limited number of tokens L_i of tweet d_i as each tweet reply is limited to 280 characters [4]. The short and informal nature of these short texts results in sparse lexical co-occurrence signals, motivating the need for semantic enrichment before topic modeling algorithms are applied.

LLM-Based Rephrasing

The second stage introduces an LLM-based rephrasing operator that transforms each short-text document (ie, a public reply to a CDC tweet) into a semantically and contextually refined version that is similar to CDC's official communication. Let R_θ denote a rephrase model (ie, an LLM) parameterized by θ . For each short-text reply d_i , the rephrasing process is defined as:

$$\hat{d}_i = R_\theta(\pi, d_i), \quad i = 1, 2, \dots, N \quad (2)$$

where $\hat{d}_i$ refers to the rephrased short-text reply, and π denotes a carefully designed prompt that constrains the generation process to preserve semantic fidelity and ensure that rephrasing is executed under a certain scheme that will be introduced in this subsection. This rephrasing stage is crucial for the downstream topic modeling of the public replies and for better understanding of public responses to official health communications during emergencies.

For the rephrasing stage, we experimented with 3 different LLMs: Google Gemini (Gemini 2.5 Flash [21]), OpenAI generative pretrained transformer (GPT-4o mini [22]), and an open-source LLM (Mistral-7B-Instruct [23]), given their strong natural language generation capabilities and complementary trade-offs between performance and cost for large-scale applications. We used a low sampling temperature (0.2) to ensure semantic fidelity and minimize stochastic variation during automated rephrasing, consistent with findings from prior work [24].

To systematically investigate how different styles of linguistic refinement influence topic modeling performance, TM-Rephrase incorporates 2 distinct prompt-guided rephrasing schemes that we developed: general rephrasing and colloquial-to-formal rephrasing. Although both schemes are designed to preserve semantic fidelity, they differ in the

degree and style of linguistic transformation applied to the original short texts of public replies. Prompts for these 2 schemes are shown in Table S1 of [Multimedia Appendix 1](#) in Section 2: prompt information for rephrasing schemes.

The general rephrasing scheme performs lightweight linguistic refinement while strictly maintaining the meaning and informational content of the original text. Its primary objective is to improve grammatical correctness, syntactic clarity, and overall readability without altering domain-specific terminology, named entities, hashtags, usernames, or technical details. Concretely, the general rephrasing scheme fixes typographical errors, incomplete sentence fragments, inconsistent verb tenses, and ambiguous phrasing, all of which are common in social media discourse. It may restructure fragmented clauses into complete sentences and resolve minor ambiguities that arise from informal writing conventions. However, it avoids introducing new concepts, removing key terms, or substituting specialized vocabulary. As such, general rephrasing can be viewed as a minimal-intervention normalization process that enhances structural regularity while preserving the lexical distribution of domain-relevant tokens.

The colloquial-to-formal rephrasing scheme transforms informal, conversational, or slang-based expressions into formal, professional English suitable for public health communication contexts. The rephrased outputs are structured to resemble the tone and style of public health reports or professional summaries, using complete sentences and standardized grammar while avoiding slang, contractions, and casual phrasing. During this process, implicit references—such as omitted subjects, abbreviated expressions, or context-dependent meanings (eg, “this shot,” “they said,” or fragmentary statements lacking clear referents)—are made explicit by clarifying the intended subject, action, or relationship within the sentence. Fragmented or elliptical constructions are rewritten into fully articulated and logically coherent statements to improve readability and interpretability. Strict constraints are applied to preserve the original meaning, including retaining all named entities, hashtags, usernames, and domain-specific terminology. No additional claims, interpretations, or external information are introduced.

These two rephrasing schemes enable a controlled comparison between moderate linguistic normalization and more substantial stylistic formalization for health communications. This design allows us to evaluate whether incremental grammatical refinement alone suffices to improve topic modeling performance, or whether deeper formalization is required for optimal topic quality. In order to ensure semantic

alignment between the original and rephrased texts, the prompt explicitly instructs the LLMs to retain meaning and avoid modifying domain-specific content.

The resulting rephrased corpus is defined in [equation 3](#), where $\hat{d}_i$ represents each individual rephrased short-text reply:

$$\hat{D} = \{\hat{d}_1, \hat{d}_2, \dots, \hat{d}_N\} \quad (3)$$

Expert Validation of Semantic Fidelity

To evaluate whether LLM-based rephrasing preserved the meaning, stance or intent, and tone of the original public replies, we further conducted a human expert validation study. A sample of 30 original CDC-reply tweets was randomly selected for validation. For each original tweet, 2 rephrased versions were evaluated: one generated using the general rephrasing scheme and one generated using the colloquial-to-formal scheme, resulting in 60 original-rephrased text pairs across 2 schemes.

Four expert raters with complementary expertise in health communication, social media research, linguistics, and public health independently evaluated the rephrased texts. Each rater compared the original tweet with the corresponding rephrased version in a paired format and provided a single holistic semantic-fidelity rating using a 5-point Likert scale. Raters were instructed to consider whether the rephrased version preserved the original meaning, stance or intent, and tone, but these dimensions were not rated separately. As the task required direct comparison between original and rephrased texts, complete blinding to the rephrasing scheme was not feasible. Raters completed the evaluations independently. We summarized the validation results using the mean, median, and distribution of expert ratings. To assess interrater reliability, we calculated a 2-way random-effects, absolute-agreement intraclass correlation coefficient (ICC[2,4]) for the aggregated expert ratings.

Data Preprocessing

Before applying topic modeling, both the original and rephrased corpora undergo a consistent and standardized preprocessing procedure to ensure comparability and reduce extraneous noise. The preprocessing step includes the removal of URLs, punctuation marks, and stop words, which do not contribute meaningful semantic information to topic inference during topic modeling.

At the same time, semantically meaningful elements are carefully preserved. Emojis are retained through their textual descriptions to maintain their affective or contextual signals. For example, the emoji 😊 is converted to a textual representation “smiling face,” and 🏠 is mapped to “face with medical mask.” This conversion ensures that nontextual, yet informative elements remain encoded within the lexical space rather than be discarded as noise.

All texts are further normalized by converting all characters to lowercase, followed by tokenization and

lemmatization. Lowercasing eliminates case-based lexical duplication, tokenization segments text into analyzable units, and lemmatization reduces inflected forms of tokens to their base form, thereby consolidating semantically equivalent variants into unified lexical representations. All these operations promote lexical consistency and stabilize downstream word co-occurrence statistics in topic modeling.

Formally, let $P_\gamma(\bullet)$ denote the preprocessing operator parameterized by γ , which encapsulates the sequence of normalization, tokenization, and lemmatization in the data preprocessing step. After preprocessing, each original document d_i and rephrased document $\hat{d}_i$ are respectively converted into the preprocessed format:

$$p_i = P_\gamma(d_i), \quad \hat{p}_i = P_\gamma(\hat{d}_i), \quad i = 1, 2, \dots, N \quad (4)$$

Here $p_i \in R^{|V|}$ and $\hat{p}_i \in R^{|V|}$ denote the tokenized representations of the original and the rephrased documents (ie, short-text replies), respectively, under a shared vocabulary space V . Maintaining identical preprocessing across both corpora ensures that any observed differences in topic modeling performance can be credited to the rephrasing rather than discrepancies in preprocessing or representation learning.

Topic Models

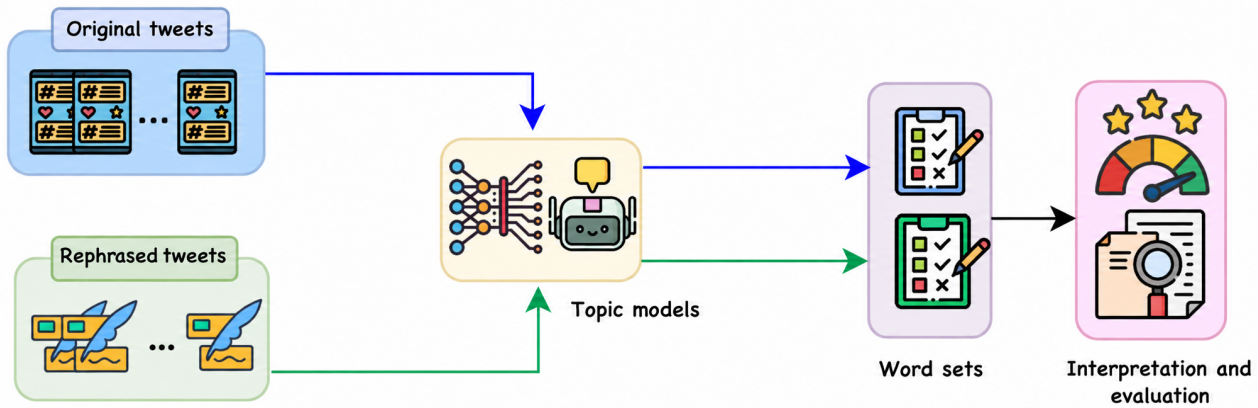
In the final stage of the pipeline, topic modeling algorithms are applied, and their performance is systematically evaluated. As illustrated in [Figure 2](#), both the original and rephrased replies are modeled by the same topic models, and the resulting topics are compared and evaluated through quantitative metrics and expert semantic fidelity evaluation. The original corpus D and rephrased corpus $\hat{D}$ are independently input into the same specific topic modeling algorithm M_{TM} to ensure methodological consistency:

$$T = M_{TM}(D), \quad \hat{T} = M_{TM}(\hat{D}) \quad (5)$$

This parallel modeling design yields 2 corresponding sets of topic outputs, that is, one derived from the original texts and the other from the rephrased texts. This enables a direct and controlled comparison. Due to the same topic modeling procedure, hyperparameter settings for topic models and preprocessing configurations are maintained across both corpora. Therefore, any observed differences in topic quality can be credited to the rephrasing scheme rather than the specific topic modeling algorithm.

To comprehensively evaluate the effects of TM-Rephrase, 4 representative topic models were applied, including conventional statistical and neural network-based topic models. This suite of different topic models enables a systematic analysis of how different topic modeling algorithms respond to TM-Rephrase.

Figure 2. Outline of the topic evaluation process. Both the original and rephrased tweets were modeled using the same topic models, and the resulting topics were compared and evaluated through quantitative metrics and expert semantic fidelity evaluation.



LDA [6] is a foundational probabilistic topic model serving as the benchmark in topic modeling studies. It is effective for classic topic discovery but is usually challenged by data sparsity in short social media texts.

BERTopic [9] is a neural topic model using contextual embeddings (eg, Sentence-BERT [25]) and density-based clustering. It is considered well-suited for noisy, informal social media data.

FASTopic [11] is based on a transformer backbone that directly captures semantic relationships between document embeddings and learnable topic-word embeddings. It uses an embedding transport plan as an optimization objective to enhance topic-word and document-topic associations.

Topic-semantic contrastive topic model (TSCTM) [12] addresses short-text sparsity via contrastive learning, generating dense semantic vectorized representations and topic distributions.

Together, these different topic models ensure that the evaluation of TM-Rephrase is not limited to a single topic modeling paradigm, but captures methodological robustness across multiple topic models. For all 4 topic models, we fixed the number of topics (K) at 8, based on an exploratory analysis to balance thematic detail and interpretability for the study dataset. For each topic, represented as a set of keywords, we retained the top 15 keywords, ranked according to the model's topic-word probability. Details of the experimental implementation are available in Section S3: implementation details in [Multimedia Appendix 1](#).

Topic Modeling Performance Evaluation

Overview

Four quantitative evaluation metrics were computed to quantify topic coherence, uniqueness, redundancy, and diversity, with details shown below.

Topic Coherence (C_v)

C_v [26] measures the semantic coherence among the topic keywords, which are determined by topic-word probabilities. This metric was chosen based on the mathematical

analysis and justifications in the study by Wu [27]. This metric fundamentally relies on normalized pointwise mutual information (NPMI) to quantify pairwise word semantic associations. For any 2 words w_i and w_j , their NPMI score is defined as:

$$NPMI(w_i, w_j) = \frac{\log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j) + \epsilon}\right)}{-\log(P(w_i, w_j) + \epsilon)} \quad (6)$$

Here, $P(w_i)$ and $P(w_j)$ are the probabilities of observing words w_i and w_j , respectively, and $P(w_i, w_j)$ is the co-occurrence probability of the word pair (w_i, w_j) within a defined external corpus (the Wikipedia [Wikimedia Foundation, Inc] corpus was used in this study [28]). A small constant ϵ (eg, 10^{-12}) is added to avoid division by 0 or taking the logarithm of 0. Building on the NPMI formulation, the C_v coherence score can be formally expressed as follows:

$$C_v = \frac{1}{T} \sum_{i=1}^T \cos(\mathbf{v}_{NPMI}(x_i), \mathbf{v}_{NPMI}(\{x_j\}_{j=1}^T)) \quad (7)$$

Here, $\mathbf{v}_{NPMI}(x_i)$ represents the vector of NPMI scores between a word x_i and all other words in the topic keyword list, defined as:

$$\mathbf{v}_{NPMI}(x_i) = \{NPMI(x_i, x_j)\}_{j=1, \dots, T} \quad (8)$$

Similarly, the aggregated NPMI vector across all words in the topic keyword list is defined as:

$$\mathbf{v}_{NPMI}(\{x_j\}_{j=1}^T) = \left\{ \sum_{i=1}^T NPMI(x_i, x_j) \right\}_{j=1, \dots, T} \quad (9)$$

where T is the number of top keywords in the topic ($T=15$ in this study). The cosine similarity is then computed between the word-specific and aggregated vectors, and the final C_v score is obtained by averaging across all words. A higher C_v score indicates stronger semantic coherence and better interpretability of the topics.

Topic Uniqueness

Nan et al [29] proposed topic uniqueness (TU), a metric that quantifies how distinct the top keywords within each topic are across the entire topic set T . This helps identify whether individual topics use distinctive vocabulary. Given K topics and the top T keywords of each topic, TU is computed as:

$$TU = \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{T} \sum_{x_i \in T_k} \frac{1}{\#(x_i)} \right) \quad (10)$$

where T_k denotes the top keyword set of the k -th topic, and $\#(x_i)$ denotes the occurrence count of word x_i in the top T words of all topics. TU ranges from $1/K$ to 1, and a higher TU score indicates more unique topics.

Topic Redundancy

Topic redundancy (TR) [30] was developed to measure the overlap of top keywords across different topics. The formula for TR is given as:

$$TR = \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{T} \sum_{x_i \in T_k} \frac{\#(x_i) - 1}{K - 1} \right) \quad (11)$$

A lower TR score suggests that the topics are more distinct.

Topic Diversity

The formula for the topic diversity (TD) [31], which measures the proportion of unique top keywords across all topics, is defined as:

$$TD = \frac{1}{K} \sum_{k=1}^K \frac{1}{T} \sum_{x_i \in T_k} I(\#(x_i)) \quad (12)$$

A higher TD score indicates more diverse topics with less word overlap. This metric works by identifying words that are unique to a single topic's top keyword list. The indicator function $I(\bullet)$ is defined by a stepwise function. This function returns 1 if and only if a word x_i appears in the top keyword list of exactly 1 topic, and 0 otherwise.

These 4 metrics, C_v , TU , TR , and TD , evaluate topic quality across complementary dimensions. In brief, C_v measures semantic coherence among a topic's top keywords, TU assesses topic uniqueness across topics, TR quantifies topic redundancy due to topical overlap, and TD evaluates whether the model produces a diverse set of topics. These metrics provide a comprehensive set of measures for evaluating and comparing topics generated from original nonrephrased tweets and those rephrased through 2 different TM-Rephrase schemes (general and colloquial-to-formal rephrasing).

Through systematic comparisons between generated topics based on original and rephrased texts respectively, we evaluated the extent to which the model-agnostic TM-Rephrase framework improves the intratopic coherence and intertopic quality.

Ethical Considerations

This study analyzed publicly accessible online tweets from Twitter that were replies to official CDC communications during the COVID-19 pandemic. To comply with the platform's privacy policy, data collection was limited to publicly accessible tweets. The social media data component did not involve direct interaction with Twitter users, recruitment of social media users, intervention, or collection of private information beyond publicly available tweet content. Informed consent from social media users was not obtained because the analysis was based solely on publicly accessible online tweets.

This study, including the human expert validation component, was reviewed by the Office of Research Protections and Integrity at the University of North Carolina at Charlotte and received a Notice of Determination of Exemption. This study was determined to meet the exempt category cited under 45 CFR 46.104(d), Exemption Category 2. The institutional review board study number is IRB-26-1225, titled "Human Evaluation of Topic Modeling Performance for Text Analysis of Public Health-Related Social Media Short Texts," with an approval date of July 8, 2026 (Figure S1 in [Multimedia Appendix 1](#) in Section S4: notice of determination of exemption from IRB Office in UNC Charlotte).

To minimize privacy risks, no usernames, profile information, or other direct user identifiers are reported in this paper. Example tweets, when presented, are used only for methodological illustration, and the main findings are reported in aggregate form. Expert validation results are also summarized in aggregate, without identifying individual raters.

Results

Overview

This section presents the results of quantitative evaluation of TM-Rephrase, followed by illustrative examples. We report results across four topic models, FASTopic, BERTopic, TSCTM, and LDA, under three conditions, as follows: (1) original tweets without any rephrasing, (2) general rephrasing, and (3) colloquial-to-formal rephrasing, with three different LLMs used for rephrasing (ie, Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct). Topic quality was quantitatively assessed using 4 metrics: C_v , TU , TR , and TD , as described in the Topic Modeling Performance Evaluation section.

Before introducing the quantitative results, we refer to [Table 1](#) to assess potential text expansion or compression caused by TM-Rephrase. Both general and colloquial-to-formal rephrasing schemes hardly change token lengths: the mean increases marginally from 27.33 (original, SD 14.23) to 27.76 (general, SD 15.12) and 28.14 (colloquial-to-formal, SD 14.11), and the median changes slightly from 27 (IQR 15-40) to 29 (colloquial-to-formal, IQR 16-42). Quartiles and maxima show similar minor differences, indicating that

rephrasing preserves the overall text length distribution, and any effects on topic modeling are unlikely to stem from token counts.

Quantitative Results

Overview

Table 2 reports the quantitative evaluation results under a 3-factor design, including 4 topic modeling algorithms, 3

LLM backends, and 2 rephrasing schemes plus the original baseline, resulting in a total of 28 configurations. This design enables a systematic assessment of the main effects of rephrasing, as well as its interaction with model architecture and LLM choice.

Table 2. Quantitative metric results for various topic models, both with and without TM-Rephrase by three LLMs^{a,b}.

	$C_t(0-1) \uparrow^c$	$TU^d(0.125-1) \uparrow$	$TR^e(0-1) \downarrow$	$TD^f(0-1) \uparrow$
FASTopic ^g w/o ^h rephr ^{i,j}	.3388	.9917	.0024	.975
Gemini 2.5 Flash				
FASTopic w/ ^k general rephr ^j	.3723	.9917	.0024	.9917
FASTopic w/ c-to-f rephr ^j	.3301	1 ^l	0 ^l	1 ^l
GPT-4o mini				
FASTopic w/ general rephr	.3746 ^l	.9917	.0024	.9917
FASTopic w/ c-to-f rephr	.3422	1 ^l	0 ^l	1 ^l
Mistral-7B-Instruct				
FASTopic w/ general rephr	.3728	.9917	.0024	.9917
FASTopic w/ c-to-f rephr	.3321	.9917	.0024	.9917
BERTopic ^m w/o rephr	.4078	.4667	.3976	.4667
Gemini 2.5 Flash				
BERTopic w/ general rephr	.4564	.525 ^l	.3357 ^l	.525 ^l
BERTopic w/ c-to-f rephr	.4734 ^l	.525 ^l	.3452	.525 ^l
GPT-4o mini				
BERTopic w/ general rephr	.4612	.4667	.3357 ^l	.4667
BERTopic w/ c-to-f rephr	.4687	.525 ^l	.3452	.525 ^l
Mistral-7B-Instruct				
BERTopic w/ general rephr	.4574	.4667	.3357 ^l	.4667
BERTopic w/ c-to-f rephr	.4713	.4667	.3976	.4667
TSCTM ⁿ w/o rephr	.3094	.9917	.0024	.975
Gemini 2.5 Flash				
TSCTM w/ general rephr	.3145	.9833	.0048	.9833
TSCTM w/ c-to-f rephr	.3394 ^l	1 ^l	0 ^l	1 ^l
GPT-4o mini				
TSCTM w/ general rephr	.3267	.9917	.0024	.9917
TSCTM w/ c-to-f rephr	.3331	1 ^l	0 ^l	1 ^l
Mistral-7B-Instruct				
TSCTM w/ general rephr	.3104	.9833	.0024	.9833
TSCTM w/ c-to-f rephr	.3213	.9917	.0024	.975
LDA ^o w/o rephr	.3094	.575 ^l	.3095	.575 ^l
Gemini 2.5 Flash				
LDA w/ general rephr	.4206	.5583	.3214	.5583
LDA w/ c-to-f rephr	.5004 ^l	.575 ^l	.3048 ^l	.575 ^l
GPT-4o mini				
LDA w/ general rephr	.4197	.5583	.3214	.5583
LDA w/ c-to-f rephr	.4917	.575 ^l	.3095	.575 ^l
Mistral-7B-Instruct				
LDA w/ general rephr	.4008	.5583	.3095	.5583
LDA w/ c-to-f rephr	.4762	.5583	.3095	.5583

^aLLM: large language model.

^bTable rows of TM-Rephrase are results based on Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct, respectively.

^cArrows in the header indicate the desired direction for each metric (higher ↑ or lower ↓ is better).

^dTU: topic uniqueness.

^eTR: topic redundancy.

^fTD: topic diversity.

^gFASTopic: fast, adaptive, stable, and transferable topic model.

^hw/o: without.

ⁱRephr stands for the rephrasing.

^jGeneral rephr stands for the general rephrasing scheme, while c-to-f rephr stands for the colloquial-to-formal rephrasing scheme.

^kw: with.

^lThe best results within each model group.

^mBERTopic: BERT-based topic modeling.

ⁿTSCTM: topic-semantic contrastive topic model.

^oLDA: latent Dirichlet allocation.

Effect of Rephrasing Schemes

Across most configurations, TM-Rephrase improves topic quality, with the most pronounced and stable gains observed in topic coherence (C_v). However, the magnitude and consistency of improvement vary across topic models, evaluation metrics, and rephrasing schemes. In particular, the colloquial-to-formal rephrasing scheme generally yields the highest coherence scores. For example, with LDA, coherence increases substantially from 0.3094 (without rephrasing) to 0.5004 under Gemini 2.5 Flash, with similarly large improvements observed for GPT-4o mini (0.4917) and Mistral-7B-Instruct (0.4762). These results indicate that rephrasing effectively mitigates contextual sparsity in short texts and enhances semantic consistency across topics.

As C_v coherence relies on word co-occurrence statistics derived from an external reference corpus, its values may be influenced by how closely the topic keywords align with the lexical patterns of that corpus. In this study, the reference corpus was Wikipedia, which may better represent standardized or encyclopedic expressions than informal social media language. Therefore, colloquial-to-formal rephrasing may partly improve C_v by shifting topic keywords toward lexical patterns more characteristic of Wikipedia, rather than solely by improving the substantive representation of public meaning.

Effect Across Topic Models

The magnitude of improvement varies across topic models. Probabilistic models such as LDA exhibit the largest gains, reflecting their strong dependence on lexical co-occurrence patterns. BERTopic shows consistent but more moderate improvements, with coherence increasing across all LLMs and rephrasing schemes, and reaching its highest value under Gemini 2.5 Flash ($C_v=0.4734$). In contrast, FASTopic demonstrates relatively smaller and less consistent gains, with slight decreases in coherence under the colloquial-to-formal scheme in some cases (eg, $C_v=0.3301$

under Gemini 2.5 Flash). This suggests that embedding-based models may already partially capture semantic relationships despite informal language, reducing the effectiveness of input standardization. TSCTM, meanwhile, benefits substantially from rephrasing in terms of diversity-related metrics, achieving perfect topic separation ($TU=1$, $TD=1$, and $TR=0$) under colloquial-to-formal rephrasing with both Gemini 2.5 Flash and GPT-4o mini.

Effect of LLM Backbones

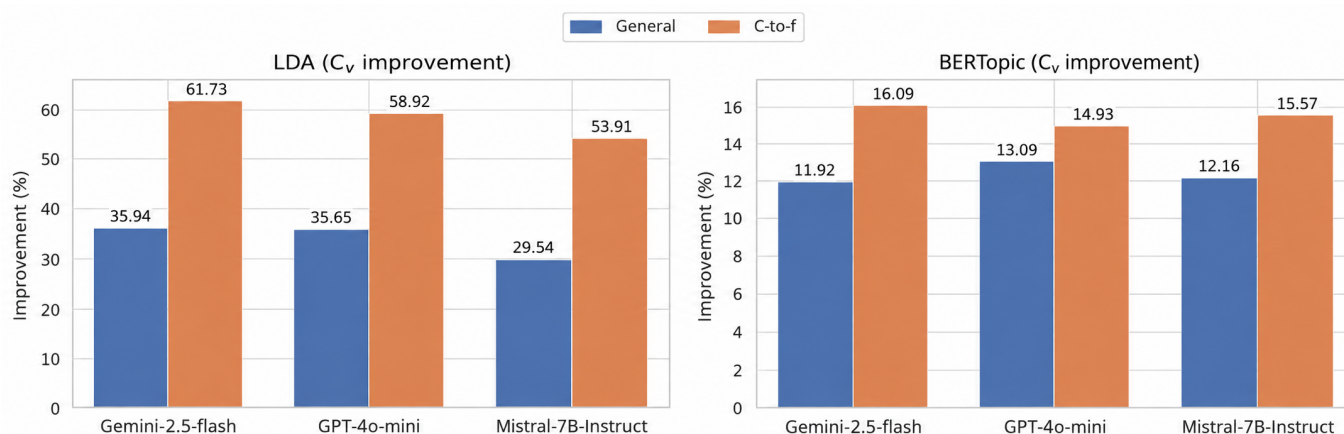
The improvements from TM-Rephrase are broadly similar across all 3 LLM backbones. Although minor variations in absolute performance are observed, the patterns remain clear and consistent: rephrasing improves coherence and diversity metrics across topic models regardless of the underlying LLM. For example, LDA coherence gains and TSCTM diversity improvements are observed consistently across Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct. This stability suggests that the effectiveness of TM-Rephrase is not dependent on a specific LLM, but rather reflects a generalizable rephrasing strategy.

Diversity and Redundancy Metrics

In addition to improving intratopic coherence, rephrasing also improves intertopic quality. Across FASTopic, BERTopic, and TSCTM, TU and TD generally increase, while TR decreases or remains similar. The most notable improvement is observed in TSCTM under colloquial-to-formal rephrasing, where perfect diversity scores are achieved, indicating fully distinct topic clusters. These findings suggest that rephrasing enhances not only intratopic coherence but also intertopic separation.

Figure 3 shows that colloquial-to-formal rephrasing often achieves larger coherence gains than general rephrasing for LDA and BERTopic, while the overall improvement patterns are observed across Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct. These results demonstrate the robustness of the TM-Rephrase framework as well.

Figure 3. C_v improvement (%) across large language models (LLMs) and rephrasing schemes for LDA and BERTopic. LDA: latent Dirichlet allocation.



Additionally, a nuanced trade-off between intratopic coherence and inter- TD emerges, especially for the colloquial-to-formal rephrasing scheme. While this scheme often delivers the highest coherence and perfect diversity scores, it can occasionally reduce TU in certain topic models. For example, colloquial-to-formal rephrasing maximizes coherence across all LLMs but slightly lowers TU compared to the nonrephrased baseline in LDA. This suggests that aggressive formalization might homogenize lexical choices across topics, highlighting the importance of matching rephrasing strength with model characteristics. In addition, the comparison across LLM backbones reveals a high degree of robustness in TM-Rephrase. While Gemini 2.5 Flash often achieves the strongest absolute results, GPT-4o mini and Mistral-7B-Instruct closely track its performance across

all metrics and models. The relative ranking of rephrasing schemes and topic models remains largely unchanged, indicating that TM-Rephrase is not tightly tied to a specific LLM but instead constitutes a generalizable topic-model-agnostic paradigm.

To demonstrate whether the observed improvements were dependent on the use of C_v coherence alone, we further evaluated topic coherence using 3 additional metrics: C_{NPMI} [26], C_{UCI} [32], and U_{MASS} [33]. These metrics provide complementary views of topic coherence based on word co-occurrence patterns, with higher values indicating better coherence for all 3 metrics. The results are reported in Table 3.

Table 3. Quantitative coherence metric results for various topic models, both with and without TM-Rephrase by 3 LLMs^{a,b}.

	$C_{NPMI} (-1, 1) \uparrow^c$	$C_{UCI} (-\infty, +\infty) \uparrow$	$U_{MASS} (-\infty, 0] \uparrow$
FASTopic ^d w/o ^e rephr ^f	-0.033	-1.477	-5.035
Gemini 2.5 Flash			
FASTopic w/ ^g general rephr ^h	-0.021 ⁱ	-1.461	-4.671 ⁱ
FASTopic w/ c-to-f rephr ^h	-0.036	-1.456 ⁱ	-4.818
GPT-4o mini			
FASTopic w/ general rephr	-0.027	-1.471	-4.833
FASTopic w/ c-to-f rephr	-0.031	-1.466	-4.874
Mistral-7B-Instruct			
FASTopic w/ general rephr	-0.024	-1.469	-4.726
FASTopic w/ c-to-f rephr	-0.035	-1.465	-4.819
BERTopic ^j w/o rephr	.028	.38	-2.358
Gemini 2.5 Flash			
BERTopic w/ general rephr	.049 ⁱ	.401	-2.18
BERTopic w/ c-to-f rephr	.045	.386	-2.163 ⁱ
GPT-4o mini			
BERTopic w/ general rephr	.031	.411 ⁱ	-2.212
BERTopic w/ c-to-f rephr	.042	.387	-2.319
Mistral-7B-Instruct			
BERTopic w/ general rephr	.044	.385	-2.338

	$C_{NPMI} (-1, 1) \uparrow^c$	$C_{UCI} (-\infty, +\infty) \uparrow$	$U_{MASS} (-\infty, 0] \uparrow$
BERTopic w/ c-to-f rephr	.03	.381	-2.314
TSCTM ^k w/o rephr	-0.021	-1.065	-3.512
Gemini 2.5 Flash			
TSCTM w/ general rephr	-0.017	-0.955	-3.141
TSCTM w/ c-to-f rephr	-0.013 ⁱ	-0.876 ⁱ	-2.949
GPT-4o mini			
TSCTM w/ general rephr	-0.015	-1.032	-2.941 ⁱ
TSCTM w/ c-to-f rephr	-0.016	-0.944	-2.997
Mistral-7B-Instruct			
TSCTM w/ general rephr	-0.02	-0.912	-3.217
TSCTM w/ c-to-f rephr	-0.018	-0.998	-3.111
LDA ^l w/o rephr	.017	-0.084	-2.457
Gemini 2.5 Flash			
LDA w/ general rephr	.038	.282	-2.401
LDA w/ c-to-f rephr	.071 ⁱ	.848 ⁱ	-2.146
GPT-4o mini			
LDA w/ general rephr	.069	.193	-2.39
LDA w/ c-to-f rephr	.058	.764	-2.418
Mistral-7B-Instruct			
LDA w/ general rephr	.047	.242	-2.06 ⁱ
LDA w/ c-to-f rephr	.066	.656	-2.279

^aLLM: large language model.

^bTable rows of TM-Rephrase are results based on Gemini 2.5 Flash, GPT-4o mini, and Mistral-7B-Instruct, respectively.

^cArrows in the header indicate the desired direction for each metric (higher $\uparrow$ or lower $\downarrow$ is better).

^dFASTopic: fast, adaptive, stable, and transferable topic model.

^ew/o: without.

^fRephr stands for rephrasing.

^gw/: with.

^hGeneral rephr stands for the general rephrasing scheme, while c-to-f rephr stands for the colloquial-to-formal rephrasing scheme.

ⁱThe best results within each model group.

^jBERTopic: BERT-based topic modeling.

^kTSCTM: topic-semantic contrastive topic model.

^lLDA: latent Dirichlet allocation.

The results support the main finding that TM-Rephrase generally improves topic coherence, although the magnitude of improvement varies across topic models, LLM backbones, and rephrasing schemes. The greatest and most consistent improvements are observed for LDA. For example, under Gemini 2.5 Flash, LDA improves from 0.017 to 0.071 on C_{NPMI} , from -0.084 to 0.848 on C_{UCI} , and from -2.457 to -2.146 on U_{MASS} after colloquial-to-formal rephrasing. Similar improvements are also observed for GPT-4o mini and Mistral-7B-Instruct, suggesting that the coherence gains for LDA are not specific to a single LLM backbone.

For BERTopic and TSCTM, the results also show improved coherence after rephrasing, although the best-performing rephrasing scheme differs by metric and LLM. For BERTopic, rephrased texts improve C_{NPMI} and C_{UCI} in most settings, with Gemini 2.5 Flash general rephrasing achieving the highest C_{NPMI} score and GPT-4o mini general rephrasing achieving the highest C_{UCI} score. For TSCTM, both general and colloquial-to-formal rephrasing improve coherence relative to the baseline without rephrasing across

most metrics, with colloquial-to-formal rephrasing under Gemini 2.5 Flash performing best on C_{NPMI} and C_{UCI} .

The effects are less consistent for FASTopic. General rephrasing improves C_{NPMI} and U_{MASS} in several settings, while colloquial-to-formal rephrasing does not consistently improve C_{NPMI} relative to the nonrephrased baseline. This suggests that embedding-based or neural topic models may already capture some semantic relationships in noisy short texts, making them less sensitive to input-level formalization than LDA. Therefore, these results support a more cautious interpretation: TM-Rephrase generally improves topic coherence, especially for LDA, BERTopic, and TSCTM, but the magnitude of effects is topic model-dependent.

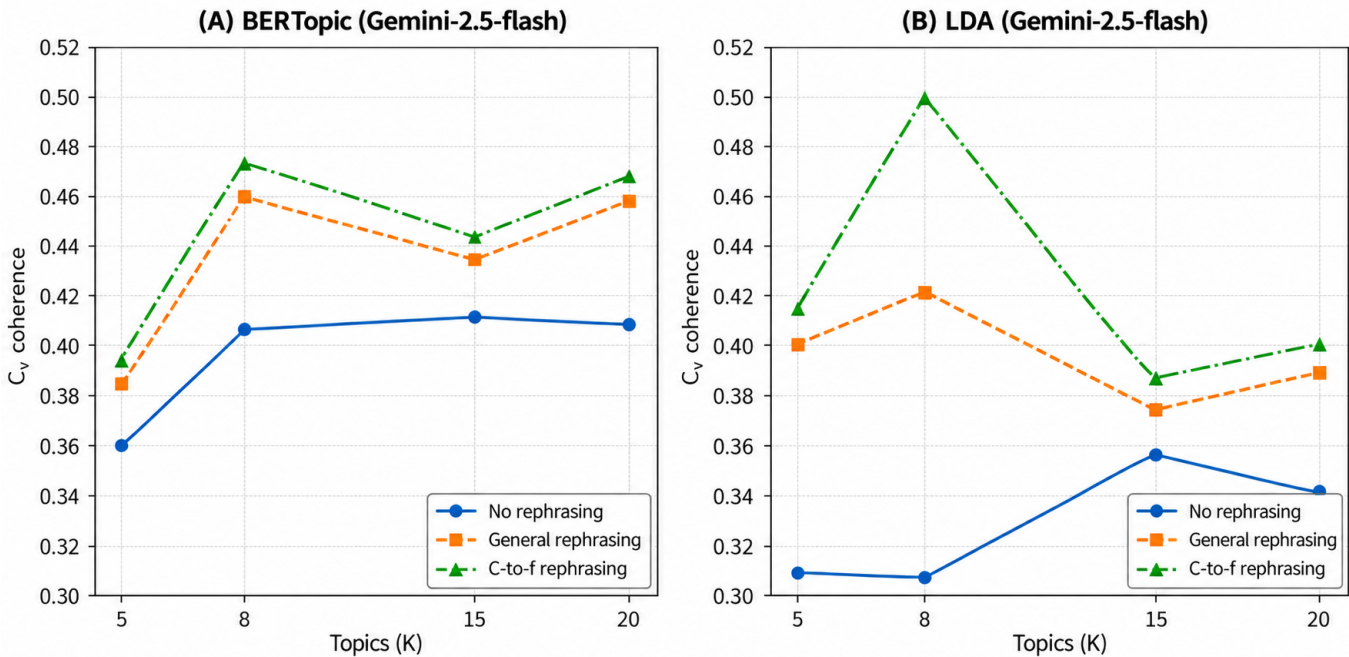
Sensitivity Analysis Across Different Topic Numbers

As shown in Figure 4, both general and colloquial-to-formal rephrasing improved C_v coherence across all tested K values for both models. For BERTopic, colloquial-to-formal rephrasing improved C_v from 0.3445 to 0.3741 at 5 and from 0.4096 to 0.4682 at 20. For LDA, colloquial-to-formal

rephrasing improved C_v from 0.3102 to 0.4093 at 5 and from 0.3411 to 0.4002 at 20. These results suggest that the coherence gains from TM-Rephrase are not limited to the

original 8 setting, although the magnitude of improvement varies across topic numbers and topic models.

Figure 4. Sensitivity of topic coherence to number of topics and rephrasing schemes for (A) BERTopic and (B) LDA, based on Gemini 2.5 Flash rephrasing backbone. BERTopic: BERT-based topic modeling; LDA: latent Dirichlet allocation.



These quantitative results suggest that TM-Rephrase generally enhances intratopic coherence and can improve intertopic quality metrics for short, noisy texts in health communications, although the effects are model- and metric-dependent. The 2 different rephrasing schemes that we developed have similar overall performance. The general rephrasing scheme offers balanced and reliable improvements, while the colloquial-to-formal scheme has a slight edge, especially for lexically sensitive topic models such as LDA and BERTopic. These findings demonstrate the effectiveness and robustness of TM-Rephrase as an LLM-based model-agnostic framework for topic modeling in health communications.

Results of Expert Ratings of Rephrasing Semantic Fidelity

The expert validation results (Table 4) provided additional evidence regarding the semantic fidelity of the 2 LLM-based rephrasing schemes. Across all 240 expert ratings, the mean semantic fidelity score was 3.90 (SD 1.01), and the median score was 4.00 (IQR 3.00-5.00) out of 5.0, suggesting that rephrasing generally preserved the original meaning.

Table 4. Results of expert validation on rephrasing semantic fidelity.

Validation scope	Ratings	Mean (SD)	Median (IQR)	Ratings ≥4 (%)	ICC(2,4) ^a , average-measure
Overall	240	3.90 (1.01)	4.00 (3.00-5.00)	65.42	0.736
General rephrasing	120	4.38 (0.78)	5.00 (4.00-5.00)	87.50	0.613
C-to-f ^b rephrasing	120	3.42 (0.99)	3.00 (3.00-4.00)	43.33	0.553

^aICC: intraclass correlation coefficient.

^bC-to-f: colloquial-to-formal.

However, the 2 rephrasing schemes showed different fidelity patterns. General rephrasing received higher expert ratings, with a mean score of 4.38 (SD 0.78), a median score of 5.00 (IQR 4.00-5.00), and 87.50% (105/120) of ratings of 4 or above. In contrast, the colloquial-to-formal scheme received a lower mean score of 3.42 (SD 0.99), a median score of 3.00 (IQR 3.00-4.00), and 43.33% (52/120) of ratings of 4

or above. This result suggests that general rephrasing better preserved the original meaning, stance, or intent, and tone, whereas colloquial-to-formal rephrasing introduced greater changes because of its stronger formalization.

As meaning, stance, or intent, and tone were evaluated through a single holistic rating rather than separate dimension-specific scores, the validation results cannot determine

quantitatively which aspect of fidelity contributed most to the lower ratings for colloquial-to-formal rephrasing. Therefore, the lower colloquial-to-formal ratings should be interpreted as indicating a greater risk of fidelity change overall, rather than as definitive evidence that tone or stance or intent alone was the primary source of reduced fidelity. However, 1 expert noted that meaning and stance or intent were often similar while tone differed, suggesting that professionalization of tone may be one important contributor.

The overall average interrater reliability was ICC(2,4)=0.736, indicating acceptable agreement among expert raters for the aggregated semantic fidelity ratings. Scheme-specific intraclass correlation coefficient values were lower, with ICC(2,4)=0.613 for general rephrasing and ICC(2,4)=0.553 for colloquial-to-formal rephrasing. This

pattern suggests that expert judgments were more variable when evaluating the degree to which formalized rephrasing preserved tone and stance, which is consistent with the greater linguistic transformation involved in the colloquial-to-formal scheme. Generally, these findings support a more nuanced interpretation: general rephrasing provides stronger preservation of the original public voice, whereas colloquial-to-formal rephrasing offers clearer formalization but involves a possible trade-off in tone and rhetorical fidelity.

Illustrative Topic-Level Examples

To contextualize the quantitative findings, Tables 5 and 6 provide illustrative examples of topic keywords and topic assignments before and after rephrasing.

Table 5. Top 15 keywords for LDA^a topics based on Gemini 2.5 Flash ($K=8$)^b.

Topic ID	Keywords
w/o ^c rephrasing	
1	vaccine, covid, get, people, dont ^d , mask, kid, cdc, child, need, school, vaccinated, like ^d , work, trump
2	mask, covid, wear, virus, vaccine, wearing, people, stop, dont ^d , flu, need, face, work, social, spread
3	covid, death, vaccine, cdc, people, case, day ^d , many ^d , child, one ^d , dont ^d , symptom, week ^d , died, year ^d
4	vaccine, effect, dose, people, covid, know, covaxin, dont ^d , pfizer, shot, like ^d , mrna, child, long, one ^d
5	covid, death, rate, case, cdc, data, infection, people, immunity, vaccination, study, show ^d , variant, state, number
6	vaccine, cdc, covid, pfizer, stop, fda, people, transmission, please ^d , health, year ^d , child, approved, public, prevent
7	covid, ivermectin, dont ^d , mask, cdc, people, work, vaccine, stop, positive, tested, like ^d , getting ^d , dose, doctor
8	vaccine, covid, vaccinated, get ^d , flu, still ^d , shot, people, child, year ^d , got ^d , even ^d , booster, fully ^d , getting ^d
w/ ^e general rephrasing scheme	
1	covid, cdc, vaccine, death, please ^d , mask, pandemic, trump, virus, could ^d , american, health, public, regarding, trust
2	mask, vaccinate, wear, covid, please ^d , personal_protective_equipment, get ^d , vaccine, school, individual, society, social_distance, virus, wearing, still ^d
3	covid, vaccine, individual, flu, vaccinated, people, case, virus, immunity, child, stop, positive, tested, get ^d , one ^d
4	vaccine, covid, ivermectin, mrna, pfizer, people, lie, individual, death, effective, health, covaxin, received, moderna, treatment
5	vaccine, covid, child, shot, booster, effective, mrna, prevent, flu, please, receive, cdc, people, covaxin, variant
6	mask, covid, day ^d , cases, test, people, virus, individual, cdc, one ^d , work, vaccine, hospital, positive, wear
7	covid, child, vaccine, death, risk, vaccination, data, year ^d , people, cdc, rate, individual, adverse, virus, cause
8	vaccine, risk, side_effect, testing, tweet, pfizer, would ^d , covid, child, vaccination, kid, also ^d , hospital, concern, original
w/ c-to-f ^f rephrasing scheme	
1	mask, covid, individual, vaccine, public, operation, vaccinated, state, virus, health, wear, public_measure, may, work, mandate
2	covid, child, vaccine, individual, vaccination, regarding, death, concern, risk, year ^d , virus, age, variant, may, significant
3	covid, individual, vaccine, positive, test, statement, author, mask, vaccination, testing, result, current, case, information, president
4	disease, cdc, prevention, control, center, covid, tweet, ivermectin, individual, health, professional, public, treatment, travel, coronavirus
5	vaccine, covid, effect, associated, data, symptom, adverse, cdc, child, efficiency, vaccination, side_effect, potential, disease, death
6	individual, vaccine, vaccination, covid, vaccinated, regarding, may, immunity, infection, health, efficacy, influenza, public, risk, concern
7	vaccine, vaccination, covid, individual, booster, pfizer, received, pharmaceutical, dose, trust, cdc, company, fda, financial, administration
8	public, health, regarding, covid, concern, pandemic, vaccine, social_distance, significant, information, current, mask, trump, cdc, individual

^aLDA: latent Dirichlet allocation.

^bResults with rephrasing are based on the input texts rephrased through Gemini 2.5 Flash.

^cw/o: without.

^dIrrelevant keywords (off-topic words identified through author interpretation).

^ew/: with.

^fC-to-f: colloquial-to-formal.

Table 6. Illustrative examples of original, general, and c-to-f^a rephrased tweet replies with corresponding LDA^b and BERTopic^c topic assignments^d.

Text type	Text content	Topic assigned by LDA	Topic assigned by BERTopic
Example 1			
Original	Are you kidding me with this? Don't be stupid enough to let your perfect baby be a guinea pig for these people. People are getting bells palsy after getting the vaccine.	vaccine ^e , covid, get, people, dont, mask, kid ^e , cdc, child ^e , need, school, vaccinated ^e , like, work, trump	covaxin, novavax, vaccine ^e , approve, mrna, booster, ichoosecovaxin, need, want, fda, safe ^e , people, child ^e , wait, ocugen
General rephrased	Are you serious? Don't be foolish and allow your healthy child to be a test subject for these individuals. Reports indicate people are developing Bell's palsy after receiving the vaccine.	vaccine ^e , risk ^e , side_effect ^e , testing ^e , tweet, pfizer, would, covid, child ^e , vaccination ^e , kid ^e , also, hospital, concern ^e , original	vaccine ^e , covaxin, omicron, novavax, variant, mrna, delta, booster, approve, child ^e , vaccinate, side_effect ^e , people, risk ^e , concern ^e
C-to-f rephrased	Concerns have been raised regarding potential adverse effects, specifically Bell's palsy, following vaccination. It is imperative to approach public health recommendations and medical interventions with informed decision-making.	vaccine ^e , covid, effect, associated, data, symptom, adverse ^e , cdc, child ^e , efficiency, vaccination ^e , side_effect ^e , potential, disease, death	vaccine ^e , covaxin, novavax, mrna, approval, subject ^e , receive, adverse ^e , booster, vaccination, risk ^e , efficacy, child ^e , side_effect ^e , concern ^e
Example 2.			
Original	Yes - let's trust Pfizer - same company that entered a \$2.3 billion criminal plea deal for lying! Same company that is exempt from liability and CANT be sued for any injuries - yup sign me right up *eye roll*	vaccine ^e , effect, dose, people, covid, know, covaxin, dont, pfizer ^e , shot, like, mrna, child, long, one	pfizer ^e , vaccine ^e , effect, mrna, covid, cdc, moderna, people, shot, child, test, get, know, stop, receive
General rephrased	Should we trust Pfizer? The same company that previously entered a \$2.3 billion criminal plea deal for lying is now exempt from liability and cannot be sued for any injuries. So, yes, sign me right up. *eye roll*	vaccine ^e , covid, ivermectin, mrna, pfizer ^e , people, lie ^e , individual, death, effective, health, covaxin, received, moderna, treatment	vaccine ^e , pfizer ^e , individual, vaccination, trial, mrna, concern ^e , lie ^e , report, utility, public, moderna, cdc, fda, information
C-to-f rephrased	Concerns exist regarding public trust in Pfizer, given its history, including a \$2.3 billion criminal plea deal for deceptive practices. Furthermore, the company's exemption from liability and inability to be sued for injuries raises significant questions regarding accountability.	vaccine ^e , vaccination ^e , covid, individual, booster, pfizer ^e , received, pharmaceutical, dose, trust ^e , cdc, company ^e , fda, financial, administration	vaccine ^e , pfizer ^e , individual, vaccination ^e , responsibility ^e , mrna, concern ^e , profit ^e , business, trust ^e , public, moderna, cdc, fda, information

^aC-to-f: colloquial-to-formal.

^bLDA: latent Dirichlet allocation.

^cBERTopic: BERT-based topic modeling.

^dResults with rephrasing are based on the input texts rephrased through Gemini 2.5 Flash.

^eWords that better describe topics.

These examples are intended to demonstrate how TM-Rephrase may affect interpretability. In interpreting these examples, keywords were considered more informative when they were specific, domain-relevant, and clearly aligned with public health themes, whereas generic terms, colloquial artifacts, and weakly related words were treated as less informative.

In Table 5, LDA-derived topics without rephrasing exhibit limited semantic coherence, particularly for topic 2 (public health measures), which is dominated by generic and less informative terms (eg, “mask,” “wear,” and “stop”) and

includes off-topic noise such as “dont.” General rephrasing improves topical specificity by introducing more contextually meaningful health-related terms (eg, “personal_protective_equipment” and “social_distance”), although some colloquial artifacts persist (eg, “please” and “could”).

In contrast, colloquial-to-formal rephrasing produces more coherent and structured topics, characterized by standardized and policy-relevant terminology (eg, “mandate” and “public_measure”). A similar pattern is observed in vaccine-related topics (eg, topic 5), where baseline outputs contain vague terms (eg, “like” and “get”), while rephrasing,

particularly for colloquial-to-formal, yields more precise and analytically meaningful keywords (eg, “adverse,” “associated,” and “efficiency”). These examples suggest that increasing linguistic formality and contextual explicitness can reduce less informative terms and make selected topic outputs easier to interpret in relation to public health discourse.

Table 6 further illustrates how rephrasing may affect the apparent semantic alignment between selected replies and their assigned topics. In **Table 6**, rephrasing improves the semantic alignment between tweet content and assigned topics across both LDA and BERTopic. In example 1, the original tweet expresses concerns about vaccine side effects using informal and fragmented language, resulting in generic or weakly focused topic keywords (eg, “vaccine” and “child”). General rephrasing introduces explicit health-related terms such as “side effect,” “risk,” and “concern,” improving thematic alignment. Colloquial-to-formal rephrasing further strengthens this alignment by incorporating more formal and domain-specific terminology (eg, “adverse”), enabling both models to produce more coherent and semantically grounded topics. A similar pattern is observed in example 2, where the original tweet contains colloquial and emotionally charged expressions regarding Pfizer.

Topic assignments from the original text include noisy and less informative terms (eg, “people,” “dont,” and “get”). General rephrasing improves relevance by introducing clearer semantic cues (eg, “concern”), while colloquial-to-formal rephrasing further refines the discourse with more structured and policy-relevant vocabulary (eg, “trust,” “responsibility,” and “profit”), allowing topic models, particularly BERTopic, to capture higher-level institutional themes.

Discussion

Principal Findings.

This study developed and evaluated TM-Rephrase, a model-agnostic framework that uses LLM-based rephrasing to standardize short, informal public health texts while aiming to preserve their original meaning. Using 25,027 public replies to CDC tweets on Twitter during the COVID-19 pandemic, we found that rephrasing improved topic coherence and topic distinctiveness while generally preserving semantic fidelity across multiple topic modeling approaches.

For RQ1 (whether and how effectively TM-Rephrase improves the quality and interpretability of topic model outputs for short texts in health communications), the results show that TM-Rephrase improves topic quality and interpretability for short, noisy social media texts. The strongest coherence gains were observed for LDA, where C_v increased from 0.3094 to 0.5004, suggesting that rephrasing helps mitigate lexical sparsity and fragmented wording. Improvements were also observed for embedding-based approaches such as BERTopic, indicating that the benefits of rephrasing are not limited to one type of topic model. Across multiple evaluation metrics, including C_v , C_{NPMI} , C_{UCI} , U_{MASS} , TU , TR , and TD , the results suggest that TM-Rephrase produ-

ces topic-word distributions that are more coherent, less redundant, and easier to interpret.

However, these improvements should be interpreted cautiously. Automated coherence metrics may partly reward the lexical regularity introduced by rephrasing, and higher topic coherence does not necessarily mean that all meaningful public-response signals are preserved. In infoveillance contexts, sarcasm, distrust, anger, uncertainty, and oppositional framing are not merely noise; they may represent important public health signals. Therefore, we interpret TM-Rephrase as improving the readability and interpretability of topic-model outputs, rather than as fully validating the substantive meaning of public responses.

For RQ2 (how different rephrasing schemes influence topic model performance), the 2 rephrasing schemes showed different effects. General rephrasing produced balanced and consistent improvements while better preserving the original public voice. In contrast, the colloquial-to-formal scheme often produced stronger gains in coherence and topic distinctiveness, especially for lexically sensitive models such as LDA and TSCTM. At the same time, colloquial-to-formal rephrasing involves greater linguistic transformation and may alter tone, rhetorical force, or framing. This suggests that prompt design should be aligned with the analytic goal: general rephrasing may be preferable when preserving public expression is important, while colloquial-to-formal rephrasing may be useful when the goal is to maximize topical clarity and model interpretability.

For RQ3 (how to interpret public discussions, that is, replies to CDC during the COVID-19 pandemic, based on the topics and inform public health practice), the improved topic representations helped clarify public responses to CDC communications. Rephrased topics more clearly captured recurring concerns related to vaccination attitudes, perceived risks, preventive behaviors, mistrust of official guidance, and uncertainty about public health recommendations. These findings suggest that TM-Rephrase can support public health social listening by helping analysts organize large volumes of noisy public replies into more interpretable thematic structures.

Practically, TM-Rephrase may serve as a human-in-the-loop preprocessing strategy for public health social listening rather than a replacement for expert interpretation. By reducing lexical fragmentation and improving the readability of topic-model outputs, it may help analysts organize large volumes of noisy public replies and identify themes that warrant closer review, such as vaccine safety concerns, mistrust of official guidance, confusion about preventive measures, or perceived gaps between public messaging and lived experience. However, this study does not directly evaluate whether TM-Rephrase improves public health decisions, communication outcomes, or detection of otherwise missed concerns. Therefore, its public health utility should be interpreted as a potential application that requires further validation in real-world infoveillance and communication workflows.

Generally, this study provides evidence that LLM-based rephrasing can enhance topic modeling of short public health texts, while also highlighting the need for careful validation. Future work should further evaluate semantic fidelity through larger-scale human expert review, topic intrusion tasks, posttopic alignment assessments, and cross-platform testing across different public health contexts.

Limitations and Future Work

This study has several limitations that should be addressed in future research. First, during dataset curation, we did not perform additional bot or spam classification beyond the filtering described in Data Collection, because the primary objective was to evaluate the effect of rephrasing on topic modeling outputs rather than to infer public opinion. The filtering process may also introduce selection bias. In particular, replies linked to deleted or inaccessible CDC tweets were excluded, and these excluded interactions may differ systematically from the retained replies. Therefore, the final corpus should be interpreted as a dataset of public replies with recoverable CDC conversational context, rather than as a complete representation of all public interactions with CDC accounts or all COVID-19 discourse on Twitter. In addition, because the analysis focuses on replies to official CDC communications on Twitter during the COVID-19 pandemic, future research should test TM-Rephrase across other platforms, health topics, population groups, and public health communication contexts before applying it in operational surveillance or social listening systems.

Second, although LLM-based rephrasing is designed to preserve semantic meaning while improving linguistic clarity, it may introduce an important trade-off for public health infoveillance. Rephrasing can reduce lexical fragmentation and improve topic coherence, but it may also soften or alter meaningful public-discourse signals, including sarcasm, anger, distrust, misinformation cues, uncertainty, oppositional framing, and colloquial expressions. These features should not be treated merely as noise, because they may reflect public resistance, confusion, risk perception, or distrust toward health institutions.

The expert semantic-fidelity validation provides partial evidence regarding this trade-off. Overall, expert ratings suggested generally favorable preservation of meaning, stance or intent, and tone; however, general rephrasing received stronger semantic-fidelity ratings than colloquial-to-formal rephrasing. This suggests that stronger formalization may be more likely to alter tone, rhetorical force, or stance expression. Therefore, automated improvements in coherence should not be interpreted as definitive evidence of improved substantive interpretability. In infoveillance applications, original and rephrased texts should be examined together, especially when sarcasm, distrust, anger, uncertainty, misinformation cues, or oppositional framing are

central to the RQ. In addition, the expert validation used a single holistic semantic-fidelity rating that asked raters to consider meaning, stance or intent, and tone together; therefore, this study cannot quantitatively determine which specific dimension of fidelity was most affected by rephrasing. Future validation should rate meaning, stance or intent, tone, uncertainty, sarcasm, and distrust as separate dimensions to better identify which aspects of public discourse are most vulnerable to meaning drift or tone normalization.

Recent AI alignment research suggests that differences across LLM backbones may extend beyond linguistic performance. Lau et al [34] found that LLMs can vary in value-priority profiles and alignment with human judgments. This is relevant to TM-Rephrase because rephrasing public discourse may involve implicit choices about tone, framing, emphasis, and socially sensitive meanings. Thus, although our results show broadly similar topic-modeling improvements across the LLMs tested, future work should examine whether different LLMs preserve meaning, stance, tone, uncertainty, and distrust in comparable ways, especially for sensitive public-health replies.

Moreover, this study used a fixed number of topics and a selected set of representative topic modeling approaches to enable consistent comparison between original and rephrased texts. Although this design supports methodological evaluation, public health communication is dynamic and context-dependent. Future work could adapt TM-Rephrase to better capture evolving public concerns across different stages of health emergencies.

Conclusions

This study highlights the value of improving input text quality for online public health infoveillance and analytics, particularly those based on social media. By leveraging LLM-based rephrasing, TM-Rephrase enables more interpretable, consistent, and reliable detection of thematic patterns in short, noisy public responses to official health communications during health emergencies.

The data-centric and model-agnostic TM-Rephrase framework also offers a practical and scalable solution to enhance the usability of large-scale social media data for infoveillance. The findings suggest that linguistic standardization plays an important role in strengthening the interpretability of analytical outputs and supporting a clearer understanding of public concerns from large volumes of raw online data.

As social media continues to serve as a key channel for public engagement on various health topics, novel analytical frameworks, such as TM-Rephrase, enable more effective monitoring of public discourse and better-informed public health communication strategies, especially during health emergencies.

Acknowledgments

The authors gratefully acknowledge Dr Sijia Qian and Dr Ron Lunsford for participating as expert validators in the semantic fidelity assessment of original and large language model–rephrased social media tweets. Their feedback helped evaluate whether the rephrased versions preserved the meaning, stance or intent, and tone of the original tweets. The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy [35]; 2025), the following tasks were delegated to GenAI tools under full human supervision: selection of research methods, proofreading and editing, and reformatting. The GenAI tool used was ChatGPT-5.5 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the outcomes.

Funding

This study was supported by the National Science Foundation (DMS-2436227). The funding organization had no role in this study's design, data collection and analysis, decision to publish, or preparation of this paper. The opinions, findings, conclusions, and recommendations presented in this paper are those of the authors alone and do not necessarily represent the views of the funding organization.

Data Availability

The datasets analyzed during this study consist of public replies to Centers for Disease Control and Prevention communications on Twitter (subsequently rebranded as X). The raw tweet-level social media datasets are not publicly shared because redistribution of user-generated social media content on Brandwatch may be restricted by platform terms and because public tweets may still contain content that could raise privacy or reidentification concerns. Aggregated results and derived materials supporting the findings are available from the corresponding author upon reasonable request, subject to applicable platform policies and ethical considerations.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Queries, prompts, implementation details, and notice of the institutional review board exemption determination.

[PDF File (Adobe File), 364 KB-Multimedia Appendix 1]

References

1. Rauchfleisch A, Vogler D, Eisenegger M. Public sphere in crisis mode: how the COVID-19 pandemic influenced public discourse and user behaviour in the Swiss Twitter-sphere. *Javnost*. 2021;28(2):129-148. [doi: [10.1080/13183222.2021.1923622](https://doi.org/10.1080/13183222.2021.1923622)] [Medline: [34393592](https://pubmed.ncbi.nlm.nih.gov/34393592/)]
2. Zhang C, Xu S, Li Z, Hu S. Understanding concerns, sentiments, and disparities among population groups during the COVID-19 pandemic via Twitter data mining: large-scale cross-sectional study. *J Med Internet Res*. Mar 5, 2021;23(3):e26482. [doi: [10.2196/26482](https://doi.org/10.2196/26482)] [Medline: [33617460](https://pubmed.ncbi.nlm.nih.gov/33617460/)]
3. Yin S, Chen S, Ge Y. Dynamic associations between Centers for Disease Control and Prevention social media contents and epidemic measures during COVID-19: in-foveillance study. *JMIR Infodemiology*. Jan 23, 2024;4(1):e49756. [doi: [10.2196/49756](https://doi.org/10.2196/49756)] [Medline: [38261367](https://pubmed.ncbi.nlm.nih.gov/38261367/)]
4. Timeline of X. Wikipedia. URL: https://en.wikipedia.org/wiki/Timeline_of_Twitter [Accessed 2026-09-10]
5. Laureate CDP, Buntine W, Linger H. A systematic review of the use of topic models for short text social media analysis. *Artif Intell Rev*. May 1, 2023;56:1-33. [doi: [10.1007/s10462-023-10471-x](https://doi.org/10.1007/s10462-023-10471-x)] [Medline: [37362887](https://pubmed.ncbi.nlm.nih.gov/37362887/)]
6. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022. URL: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> [Accessed 2026-09-10]
7. Hong L, Davison BD. Empirical study of topic modeling in Twitter. Presented at: Proceedings of the First Workshop on Social Media Analytics; Jul 25-28, 2010:80-88; Washington, DC. [doi: [10.1145/1964858.1964870](https://doi.org/10.1145/1964858.1964870)]
8. Rajagopal D, Olsher D, Cambria E, Kwok K. Commonsense-based topic modeling. Presented at: Proceedings of the second international workshop on issues of sentiment discovery and opinion mining; Aug 11-12, 2013:1-8; Chicago, IL. [doi: [10.1145/2502069.2502075](https://doi.org/10.1145/2502069.2502075)]
9. BERTopic. URL: <https://maartengr.github.io/BERTopic/> [Accessed 2026-09-10]
10. Angelov D. Top2Vec: distributed representations of topics. *arXiv*. Preprint posted online on Aug 19, 2020. [doi: [10.48550/arXiv.2008.09470](https://doi.org/10.48550/arXiv.2008.09470)]
11. Wu X, Nguyen T, Zhang DC, Wang WY, Luu AT. FASTopic: pretrained transformer is a fast, adaptive, stable, and transferable topic model. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024:84447-84481; Vancouver, BC, Canada. [doi: [10.52202/079017-2683](https://doi.org/10.52202/079017-2683)]

12. Wu X, Luu AT, Dong X. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Dec 7-11, 2022:2748-2760; Abu Dhabi, United Arab Emirates. [doi: [10.18653/v1/2022.emnlp-main.176](https://doi.org/10.18653/v1/2022.emnlp-main.176)]
13. Yang JZ, Liu Z, Wong JC. Information seeking and information sharing during the COVID-19 pandemic. *Commun Q*. Jan 1, 2022;70(1):1-21. [doi: [10.1080/01463373.2021.1995772](https://doi.org/10.1080/01463373.2021.1995772)]
14. AbuRaed AGT, Prikryl EA, Carenini G, Janjua NZ. Long COVID Discourse in Canada, the United States, and Europe: topic modeling and sentiment analysis of Twitter data. *J Med Internet Res*. Dec 9, 2024;26:e59425. [doi: [10.2196/59425](https://doi.org/10.2196/59425)] [Medline: [39652387](https://pubmed.ncbi.nlm.nih.gov/39652387/)]
15. James L, McPhail H, Foisey L, Donelle L, Bauer M, Kothari A. Exploring communication by public health leaders and organizations during the pandemic: a content analysis of COVID-related tweets. *Can J Public Health*. Aug 2023;114(4):563-583. [doi: [10.17269/s41997-023-00783-4](https://doi.org/10.17269/s41997-023-00783-4)] [Medline: [37349662](https://pubmed.ncbi.nlm.nih.gov/37349662/)]
16. Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill*. May 29, 2020;6(2):e19273. [doi: [10.2196/19273](https://doi.org/10.2196/19273)] [Medline: [32427106](https://pubmed.ncbi.nlm.nih.gov/32427106/)]
17. Maini P, Seto S, Bai R, Grangier D, Zhang Y, Jaitly N. Rephrasing the web: a recipe for compute and data-efficient language modeling. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1; 14044-14072; Bangkok, Thailand. [doi: [10.18653/v1/2024.acl-long.757](https://doi.org/10.18653/v1/2024.acl-long.757)]
18. Zhang Y, Jin H, Meng D, Wang J, Tan J. A comprehensive survey on automatic text summarization with exploration of LLM-based methods. *Neurocomputing*. Jan 2026;663:131928. [doi: [10.1016/j.neucom.2025.131928](https://doi.org/10.1016/j.neucom.2025.131928)]
19. Miah MSU, Kabir MM, Sarwar TB, Safran M, Alfarhood S, Mridha MF. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci Rep*. Apr 26, 2024;14(1):9603. [doi: [10.1038/s41598-024-60210-7](https://doi.org/10.1038/s41598-024-60210-7)] [Medline: [38671064](https://pubmed.ncbi.nlm.nih.gov/38671064/)]
20. Liu Z, Zhou Y, Zhu Y, et al. Information retrieval meets large language models. Presented at: Companion Proceedings of the ACM Web Conference 2024; May 13-17, 2024:1586-1589; Singapore, Singapore. [doi: [10.1145/3589335.3641299](https://doi.org/10.1145/3589335.3641299)]
21. Gemini 2.5 flash model. Google Cloud. URL: <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash> [Accessed 2026-09-10]
22. GPT-4o mini: advancing cost-efficient intelligence. OpenAI. URL: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> [Accessed 2026-09-10]
23. Mistral-7B-instruct-v0.2. Hugging Face. URL: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2> [Accessed 2026-09-10]
24. Renze M. The effect of sampling temperature on problem solving in large language models. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2024; Nov 12-16, 2024:7346-7356; Miami, FL. [doi: [10.18653/v1/2024.findings-emnlp.432](https://doi.org/10.18653/v1/2024.findings-emnlp.432)]
25. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 3982-3992; Hong Kong, China. [doi: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410)]
26. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015; Shanghai, China. [doi: [10.1145/2684822.2685324](https://doi.org/10.1145/2684822.2685324)]
27. Wu X. Towards effective neural topic modeling. Nanyang Technological University; 2024. URL: <https://dr.ntu.edu.sg/bitstreams/2b7465e1-7444-4cda-9078-9e7163cd4006/download> [Accessed 2026-09-10]
28. English Wikipedia database dump. Wikimedia Foundation. URL: <https://dumps.wikimedia.org/enwiki/latest/> [Accessed 2026-09-10]
29. Nan F, Ding R, Nallapati R, Xiang B. Topic modeling with Wasserstein autoencoders. Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019:6345-6381; Florence, Italy. [doi: [10.18653/v1/P19-1640](https://doi.org/10.18653/v1/P19-1640)]
30. Burkhardt S, Kramer S. Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *J Mach Learn Res*. 2019;20(131):1-27. URL: <https://jmlr.org/papers/v20/18-569.html> [Accessed 2026-09-10]
31. Dieng AB, Ruiz FJR, Blei DM. Topic modeling in embedding spaces. *Trans Assoc Comput Linguist*. Dec 2020;8:439-453. [doi: [10.1162/tacl_a_00325](https://doi.org/10.1162/tacl_a_00325)]
32. Newman D, Lau JH, Grieser K. Automatic evaluation of topic coherence. Presented at: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics; Jun 2-4, 2010; Los Angeles, CA. URL: <https://aclanthology.org/N10-1012.pdf> [Accessed 2026-09-10]

33. Mimno D, Wallach H, Talley E. Optimizing semantic coherence in topic models. Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Jul 27-31, 2011; Edinburgh, Scotland, UK. [doi: [10.5555/2145432.2145462](https://doi.org/10.5555/2145432.2145462)]
34. Lau GR, Low WY, Koh SM, Nah FFH, Hartanto A. Evaluating AI alignment in LLMs: output analysis of value priorities across 75 models with human benchmarking. arXiv. Preprint posted online on May 16, 2026. [doi: [10.48550/arXiv.2506.12617](https://doi.org/10.48550/arXiv.2506.12617)]
35. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res. Apr 2026;33(3):2544331. [doi: [10.1080/08989621.2025.2544331](https://doi.org/10.1080/08989621.2025.2544331)] [Medline: [40781729](https://pubmed.ncbi.nlm.nih.gov/40781729/)]

Abbreviations

BERTopic: BERT-based topic modeling
CDC: Centers for Disease Control and Prevention
FASTopic: fast, adaptive, stable, and transferable topic model
ICC: intraclass correlation coefficient
LDA: latent Dirichlet allocation
LLM: large language model
NPMI: normalized pointwise mutual information
RQ: research question
TD: topic diversity
TR: topic redundancy
TSCTM: topic-semantic contrastive topic model
TU: topic uniqueness

Edited by Tim Mackey; peer-reviewed by Gabriel Rongyang Lau, Hyunsang Son, Natalya Gevorgyan; submitted 15.Apr.2026; final revised version received 25.Aug.2026; accepted 31.Aug.2026; published 08.Oct.2026

Please cite as:

Xin W, Yin S, Park A, Ge Y, Chen S

Identifying Public Response Topics to the Centers for Disease Control and Prevention's COVID-19 Communications on Social Media: Infection Study Using Large Language Model-Based Rephrasing
JMIR Infodemiology 2026;6:e98319

URL: <https://infodemiology.jmir.org/2026/1/e98319>

doi: [10.2196/98319](https://doi.org/10.2196/98319)

© Wangjiaxuan Xin, Shuhua Yin, Albert Park, Yaorong Ge, Shi Chen. Originally published in JMIR Infodemiology (<https://infodemiology.jmir.org>), 08.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Infodemiology, is properly cited. The complete bibliographic information, a link to the original publication on <https://infodemiology.jmir.org/>, as well as this copyright and license information must be included.