

Domain-Adaptive Transformer-Based Abstractive Summarization of Social Media Discourse: A Comprehensive Evaluation Framework

Afrodite Papagiannopoulou ¹, Chrissanthi Angeli ²

¹Department of Electrical and Electronics Engineering, University of West Attica, Athens, Greece
Email: apapagiannop@uniwa.gr

²Department of Electrical and Electronics Engineering, University of West Attica, Athens, Greece
Email: angeli@uniwa.gr, c_angeli@otenet.gr

Abstract

The exponential growth of social media platforms has generated vast streams of unstructured, linguistically diverse user-generated content that, while rich in public sentiment and cultural insight, presents significant challenges for automated analysis. This study presents a domain-adaptive framework for abstractive summarization specifically engineered to address the linguistic complexity, structural fragmentation, and semantic irregularities characteristic of social media discourse. Leveraging the Text-to-Text Transfer Transformer (T5-small) architecture fine-tuned on Facebook news comment data comprising 1,781,576 records, our system generates concise, semantically faithful summaries that preserve contextual meaning while reducing cognitive overload. Rigorous empirical evaluation across reference-based (ROUGE variants), reference-free (SUPERT, Shannon metrics, SDC), and human-aligned quality indicators demonstrates that the proposed framework consistently outperforms ChatGPT-4.0, achieving a ROUGE-1 F1-score of 0.3292 (20% relative improvement), SUPERT score of 0.9990, and Shannon Coherency of 0.9519. The system's sophisticated preprocessing pipeline and attention-driven encoder-decoder architecture enable extraction of both factual content and affective, cultural, and pragmatic nuances embedded within informal text. This work makes three principal contributions: (1) demonstration that strategically adapted transformer architectures effectively accommodate unique properties of social media discourse while maintaining semantic faithfulness, (2) introduction of a comprehensive multidimensional evaluation framework integrating lexical, semantic, and human-centered quality metrics, and (3) empirical validation that domain-specific fine-tuning yields superior performance compared to zero-shot prompting of general-purpose large language models. These findings advance automated text summarization methodologies and provide critical tools for sentiment analysis, trend monitoring, public opinion mining, and strategic decision-making across governmental, commercial, and civic engagement domains.

Keywords: Social Media Summarization, Transformers, Abstractive Summarization, Natural Language Generation, Reference-Free Evaluation.

1. Introduction

A. Motivation and Research Context

The exponential proliferation of social media platforms over the past two decades has fundamentally transformed global communication architectures, creating ubiquitous digital ecosystems wherein billions of users engage in real-time information exchange, opinion formation, and collective sense-making. Platforms including Facebook, Instagram, X (formerly Twitter), Reddit, and LinkedIn have transcended their original function as social networking tools, evolving into primary mediators of public discourse that exert measurable

influence on political processes, economic behaviors, and societal dynamics at unprecedented scale [1]. According to recent estimates, social media users generate approximately 500 million posts daily across major platforms, with comment threads frequently containing hundreds to thousands of individual responses per event or news article. This user-generated content constitutes an invaluable repository for understanding collective attitudes, tracking sentiment evolution, and extracting actionable intelligence for evidence-based decision-making across governmental policy formulation, commercial strategy development, and academic research.

The immediacy and perceived anonymity afforded by digital platforms encourage spontaneous, unfiltered expression, particularly during real-time events such as political elections, public health crises, natural disasters, or breaking news developments. During such high-salience periods, social media streams populate with dense, reactive commentary ranging from brief reactions to extended argumentative exchanges. However, the sheer volume, velocity, and heterogeneity of this content render comprehensive manual analysis practically infeasible. A single viral news post may accumulate tens of thousands of comments within hours, creating information overload conditions wherein critical insights remain buried within overwhelming data streams. Effective automated summarization tools thus represent essential infrastructure for distilling relevant patterns, capturing evolving sentiment trajectories, and surfacing dominant thematic structures from large-scale digital discourse.

B. Challenges in Social Media Text Analysis

Social media discourse exhibits distinctive linguistic and structural properties that fundamentally differentiate it from traditional written communication and pose significant challenges for automated Natural Language Processing (NLP) systems. First, linguistic informality pervades social media text, characterized by spontaneous expression, non-standard orthography, creative spelling variations, platform-specific vernacular (e.g., "tbh," "imo," "ngl"), abbreviations, acronyms, and colloquial expressions that deviate substantially from formal written language conventions. Second, structural fragmentation manifests through brevity constraints (historical Twitter character limits, though relaxed, continue to influence expression patterns), lack of grammatical completeness, absence of conventional discourse markers, and frequent context-dependent references requiring external knowledge for interpretation.

Third, semantic variability introduces interpretive complexity through pervasive use of sarcasm, irony, humor, implicit sentiment, cultural references, memes, and emoticons/emojis that function as integral semantic components rather than decorative elements. Fourth, high signal-to-noise ratios result from prevalence of advertisements, spam content, off-topic commentary, duplicate or near-duplicate messages, and inflammatory or low-quality contributions that dilute informational content. Fifth, rapid lexical evolution characterizes social media language, wherein ephemeral trends, viral hashtags, neologisms, and emergent colloquialisms continuously reshape the active vocabulary, demanding ongoing model adaptation to maintain relevance.

Sixth, multilingual code-switching frequently occurs within single comments or threads, as users fluidly alternate between languages or incorporate foreign terms, challenging monolingual processing architectures. Finally, dialogic structure creates additional complexity, as comments often constitute responses to prior messages, reference earlier discussion points, or engage in multi-party exchanges requiring discourse-level understanding beyond isolated sentence analysis.

C. The Promise of Abstractive Summarization

Within the NLP research landscape, automatic text summarization has emerged as a particularly impactful application domain for addressing information overload challenges. Summarization methodologies generally partition into two fundamental paradigms: extractive summarization, which identifies and concatenates salient sentences or phrases directly from source texts, and abstractive summarization, which interprets semantic content and reconstructs it using novel linguistic formulations. While extractive approaches offer computational efficiency and guarantee factual grounding through verbatim reproduction, they frequently generate disjointed outputs lacking coherence, fail to capture nuanced semantic relationships, and struggle with redundancy management across selected segments.

Abstractive summarization, conversely, demonstrates substantial promise for social media contexts by generating fluent, coherent narratives that condense distributed information, resolve redundancies through paraphrasing, and produce human-like summaries that preserve essential meaning while employing contextually appropriate language. This capacity to synthesize fragmented discussions into unified overviews enables identification of prevailing sentiments, detection of emergent concerns, recognition of thematic clusters, and extraction of actionable insights from large-scale comment collections. For organizations monitoring brand perception, government agencies tracking public response to policy initiatives, news media organizations synthesizing reader reactions, or researchers analyzing collective behavior, abstractive summarization offers strategic value by revealing patterns in user sentiment, detecting nascent issues before they amplify, and informing responsive strategies grounded in actual public discourse rather than assumptions.

D. Research Gaps and Objectives

The field of automatic text summarization traces its origins to seminal work by [2], who pioneered computational approaches for distilling essential content from formal documents using statistical and heuristic methods. Over subsequent decades, the domain evolved through rule-based linguistic approaches, statistical machine learning techniques, and neural architectures, drawing increasing scholarly attention as digital communication proliferated [3]. Recent advances in deep learning, particularly the introduction of Transformer architectures [4] and large-scale pre-trained language models, have catalyzed transformative improvements in summarization capabilities across formal documents, news articles, scientific publications, and structured narratives.

However, despite extensive research on transformer-based abstractive summarization for formal texts, the specific application to user-generated social media commentary remains comparatively underexplored. Existing studies have predominantly focused on summarizing primary post content or event descriptions [5], [6], with substantially less emphasis on the subsequent user interactions embedded within comment threads—despite these comments often containing richer, more diverse perspectives than original posts. Furthermore, most social media summarization research concentrates on Twitter (microblogging) and Reddit (threaded discussions) platforms [7], with limited investigation of Facebook comment ecosystems, cross-platform generalization strategies, or systematic comparative evaluation against state-of-the-art general-purpose large language models such as ChatGPT.

The present study addresses these gaps through development and rigorous evaluation of a domain-adaptive abstractive summarization framework specifically optimized for social media user commentary. Our research is guided by three central questions:

RQ1: How can semantically coherent, contextually faithful summaries be systematically generated to reflect the inherent linguistic variability, thematic diversity, and temporal dynamics characteristic of social media discourse?

RQ2: How does domain-specific fine-tuning of transformer architectures on social media corpora compare to zero-shot or few-shot prompting of large general-purpose language models in terms of summary quality across multiple evaluation dimensions?

RQ3: How can a comprehensive, multidimensional evaluation framework be constructed to assess summary quality beyond conventional lexical overlap metrics, incorporating semantic fidelity, factual consistency, coherence, and human-aligned quality indicators?

E. Research Contributions

This work makes three principal contributions to the computational summarization literature:

Contribution 1: Domain-Adaptive Summarization Architecture. We present a transformer-based abstractive summarization system leveraging the T5 (Text-to-Text Transfer Transformer) architecture, fine-tuned on a large-scale corpus of Facebook news comment threads comprising 1,781,576 records. Through systematic preprocessing, model adaptation, and optimization, the framework generates semantically coherent, contextually rich summaries that effectively accommodate the linguistic irregularities, structural fragmentation, and semantic variability inherent to social media discourse—capabilities previously constrained by models optimized for formal documents.

Contribution 2: Comprehensive Multidimensional Evaluation Framework. We introduce an integrated evaluation methodology that transcends conventional ROUGE-based assessment, incorporating reference-based lexical overlap metrics (ROUGE-1, ROUGE-L, ROUGE-Lsum, ROUGE-S, ROUGE-WE), reference-free semantic quality indicators (SUPERT, Shannon Consistency, Shannon Coherency, Semantic Distance-based Clustering), and human-aligned quality assessments (consistency, readability, fluency, lexical diversity, semantic coherence). This framework provides methodologically rigorous templates for future social media summarization research.

Contribution 3: Empirical Validation of Domain-Specific Fine-Tuning. Through systematic comparative evaluation against ChatGPT-4.0—a state-of-the-art general-purpose large language model—we empirically demonstrate that domain-specific fine-tuning of moderately-sized transformer models (T5-small, 77M parameters) yields superior performance on specialized tasks compared to zero-shot prompting of substantially larger general-purpose models. This finding offers valuable insights for resource allocation strategies in applied NLP research and challenges assumptions regarding the universal superiority of larger foundation models across all application contexts

F. Paper Organization

The remainder of this paper is structured as follows. Section II reviews related work on neural summarization architectures, transformer-based language models, social media content analysis, and evaluation methodologies. Section III presents the methodological framework including problem formulation, T5 architectural specifications, and system implementation. Section IV describes the experimental setup encompassing dataset characteristics, training protocols, baseline configurations, and comprehensive evaluation across reference-based and reference-free metrics, followed by comparative performance analysis against ChatGPT-4.0. Section V synthesizes principal findings, discusses

theoretical and practical implications, acknowledges limitations, and delineates promising avenues for future research in multilingual extension, multimodal integration, and real-time applications.

2. Related Work

Neural approaches to automatic text summarization initially emerged with Recurrent Neural Network (RNN) architectures, particularly Long Short-Term Memory (LSTM) models, which were capable of modeling sequential dependencies in text [8]. These models showed early promise for abstractive summarization, including in social media contexts characterized by informal language, brevity, and linguistic variability. Sequence-to-sequence frameworks enhanced with attention mechanisms marked a critical improvement, enabling models to selectively focus on salient input segments rather than processing all tokens uniformly [9]. Notable examples include attentional encoder-decoder models with reinforcement strategies for social media summarization [10] and summary-aware attention mechanisms tailored to short-form content [11]. Despite these advances, RNN-based approaches suffered from inherent limitations such as sequential computation that inhibited parallelization, difficulty capturing long-range dependencies, and training instability due to vanishing gradients. These constraints motivated the transition toward alternative architectures capable of more efficient and expressive sequence modeling.

A. Transformer Models for Summarization Tasks

The introduction of the Transformer architecture by [4] fundamentally transformed natural language processing by replacing recurrence with self-attention mechanisms that directly model relationships among all tokens in a sequence. This design enables full parallelization and significantly improves the modeling of long-range semantic dependencies [12]. Early transformer-based summarization systems focused on long-form and structured texts such as news articles and scientific documents, exemplified by multi-stage systems combining BERT [13] representations with transformer-based extraction and generation components [14]. Subsequent studies demonstrated the effectiveness of transformers in dialogue and conversational summarization, capturing discourse dynamics across multi-turn exchanges [15]. The rise of large-scale pre-trained language models further accelerated progress, with architectures such as BERT, GPT, and especially T5 enabling transfer learning paradigms for summarization. T5's text-to-text formulation unifies all NLP tasks under a single sequence-to-sequence framework, while its span-corruption pre-training objective is particularly well aligned with abstractive summarization requirements [16]. Comparative evaluations consistently show that transformer-based encoder-decoder models, especially T5, achieve strong performance across diverse summarization benchmarks [7]. Transformer architectures have demonstrated remarkable adaptability beyond formal textual domains, extending to heterogeneous sources including instructional content from WikiHow [17] and web-based documents augmented with social media signals [18], thereby evidencing their versatility across diverse textual genres and registers.

B. Social Media Content Summarization

Research on social media summarization has primarily focused on platforms such as Twitter (X) and Reddit, leveraging their accessible datasets and event-centric content. Prior work has explored abstractive event summarization using transformer architectures, often combining pre-trained BERT encoders with custom or lightly trained transformer decoders to generate concise summaries from collections of posts [5]. Additional studies have addressed challenges specific to social media, including informal language, noise, and

brevity, while extending transformer methods to non-English platforms such as Sina Weibo to demonstrate cross-linguistic applicability [6]. However, most existing approaches emphasize summarization of primary posts or event descriptions, treating user comments as auxiliary signals rather than central summarization targets. This represents a notable research gap, as comment threads frequently contain richer opinions, arguments, and sentiment than the original posts. While limited work has examined sentence-level summarization in reviews or comments using transformer-based and graph-based techniques [18], comprehensive abstractive summarization of multi-comment threads remains underexplored. Given T5's strong generative capabilities, its application to user comment summarization presents a promising yet still nascent research direction [7].

C. Prompt Engineering and Parameter-Efficient Methods

The rapid scaling of language models has spurred interest in alternatives to full fine-tuning, particularly prompt engineering and parameter-efficient adaptation techniques. Prompt-based learning seeks to elicit task-specific behaviors through carefully designed natural language instructions or templates, allowing models to be adapted without updating their core parameters. These approaches range from discrete, manually crafted prompts to continuous or hybrid prompts learned as embeddings [19]. Frameworks such as OpenPrompt facilitate systematic experimentation with prompt design across different model architectures [20]. Complementary to prompting, parameter-efficient fine-tuning methods modify only a small subset of model parameters, reducing computational cost and memory requirements. Techniques such as prefix-tuning and prompt-tuning prepend learnable continuous vectors to model inputs while keeping the base model frozen, achieving performance comparable to full fine-tuning at large scales [21], [22]. These methods are particularly advantageous in resource-constrained settings or multi-task scenarios. Recent evidence from specialized domains, including healthcare, demonstrates that prompt-based and parameter-efficient strategies [23] can outperform traditional supervised learning when labeled data is limited [24], highlighting their broader relevance for emerging summarization applications such as social media comment analysis.

The reviewed literature consistently identifies transformer-based architectures, and particularly the T5 model, as state-of-the-art solutions for abstractive summarization across a wide range of textual domains. Despite these advances, several important limitations remain in existing research that motivate the present study. First, prior work has largely concentrated on summarizing primary social media posts or event descriptions, with relatively little attention devoted to the summarization of user-generated comment threads, which often contain richer and more diverse perspectives. Second, there is a lack of systematic comparative analysis between domain-specific fine-tuned summarization models and general-purpose large language models, such as ChatGPT, within the context of social media content. Third, existing evaluations rely predominantly on reference-based metrics such as ROUGE, offering an incomplete view of summary quality for informal and highly variable social media text. In response to these gaps, this research develops and rigorously evaluates a T5-based summarization system specifically optimized for social media comment threads, conducting extensive comparisons with ChatGPT-4.0 and employing a multidimensional evaluation framework that integrates lexical, semantic, and human-aligned quality metrics.

3. Methodology

A. Problem Formulation and Task Definition

The social media comment summarization task addressed in this research can be formally defined as follows (Fig. 1). Given:

- A collection of user comments $C = \{c_1, c_2, \dots, c_n\}$ associated with a specific news post or discussion thread
- Each comment c_i comprising a sequence of tokens $c_i = \{w_1, w_2, \dots, w_m\}$
- Comments exhibiting linguistic informality, variable length, semantic interdependencies, and contextual references

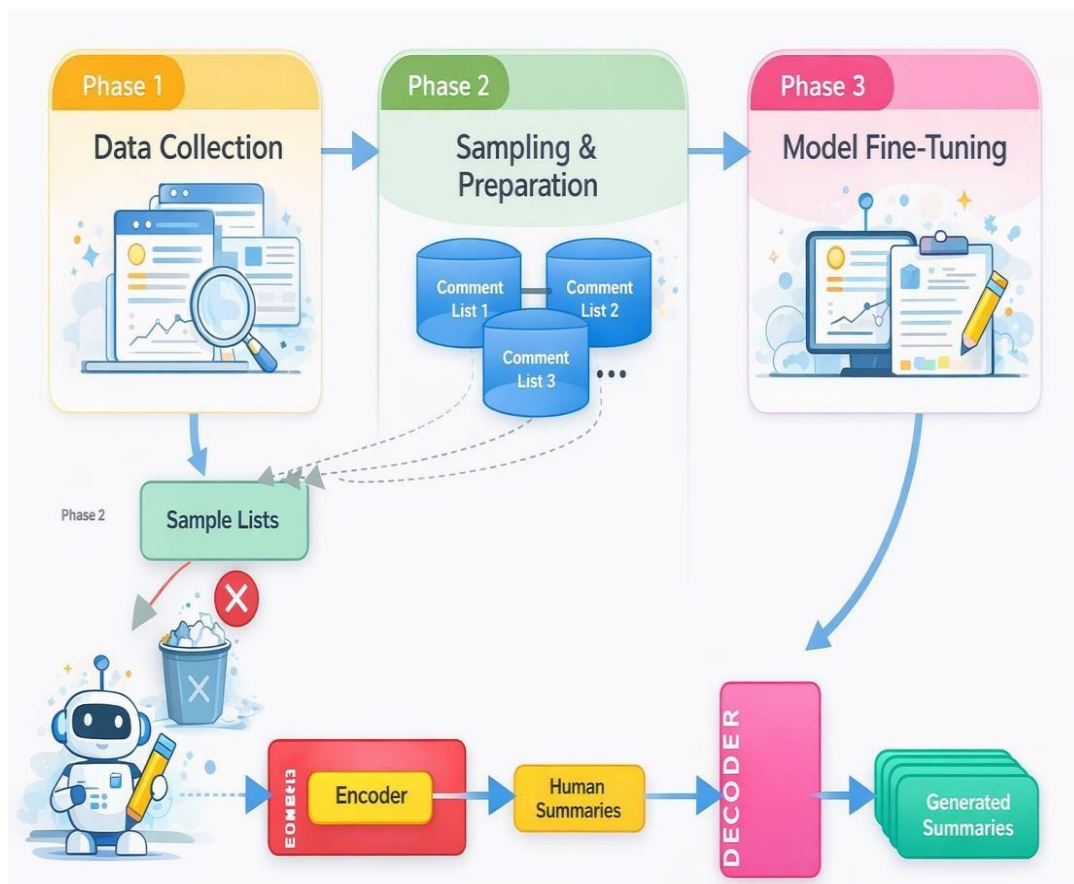


Figure 1. Abstractive Summarization System Workflow

The objective is to generate an abstractive summary $S = \{s_1, s_2, \dots, s_k\}$ where $k \ll \sum m$ (substantially shorter than concatenated input) that:

- Captures essential thematic content and salient perspectives expressed across the comment collection
- Maintains semantic fidelity to source material while employing novel linguistic formulations
- Exhibits coherence, fluency, and readability comparable to human-authored summaries

- Preserves sentiment orientation and key argumentative positions present in source comments
- Minimizes factual hallucination and unsupported inferences

Unlike formal document summarization where source texts typically exhibit logical organization and grammatical conventionality, social media comment summarization confronts additional challenges including discourse fragmentation, contextual dependencies requiring external knowledge, prevalence of non-standard linguistic forms, and high noise-to-signal ratios requiring robust filtering mechanisms [25].

B. Challenges and Design Requirements

The informal, heterogeneous nature of social media commentary imposes specific requirements on summarization architectures:

- **Linguistic Robustness:** Systems must accommodate spelling variations, grammatical irregularities, slang, abbreviations, emojis, and code-switching without catastrophic performance degradation.
- **Semantic Abstraction:** Effective summarization requires genuine content understanding and paraphrasing capability rather than extractive selection, as comments often express similar ideas through diverse linguistic realizations requiring synthesis.
- **Contextual Awareness:** Models must capture relationships between comments, recognize response structures, and integrate distributed information across multiple contributions to construct coherent summaries.
- **Length Adaptability:** Summary length should scale appropriately with input volume and complexity, avoiding both excessive compression that loses critical content and unnecessary verbosity that defeats summarization objectives.
- **Factual Grounding:** Generated summaries must remain faithful to source content, avoiding hallucination of claims, perspectives, or details not present in input comments while appropriately aggregating and generalizing expressed viewpoints.

C. The T5 Architecture: Architectural Foundations and Innovations

The Text-to-Text Transfer Transformer (T5), introduced by [16], represents a unified approach to natural language processing wherein all tasks are reformulated as sequence-to-sequence transformations from input text to output text. This elegant formulation eliminates task-specific architectural modifications, enabling a single model architecture to address diverse problems including classification (text input → label text), translation (source language → target language), question answering (question + context → answer), and summarization (document → summary).

T5 implements several architectural refinements relative to the original Transformer design [4], as illustrated in Fig. 2:

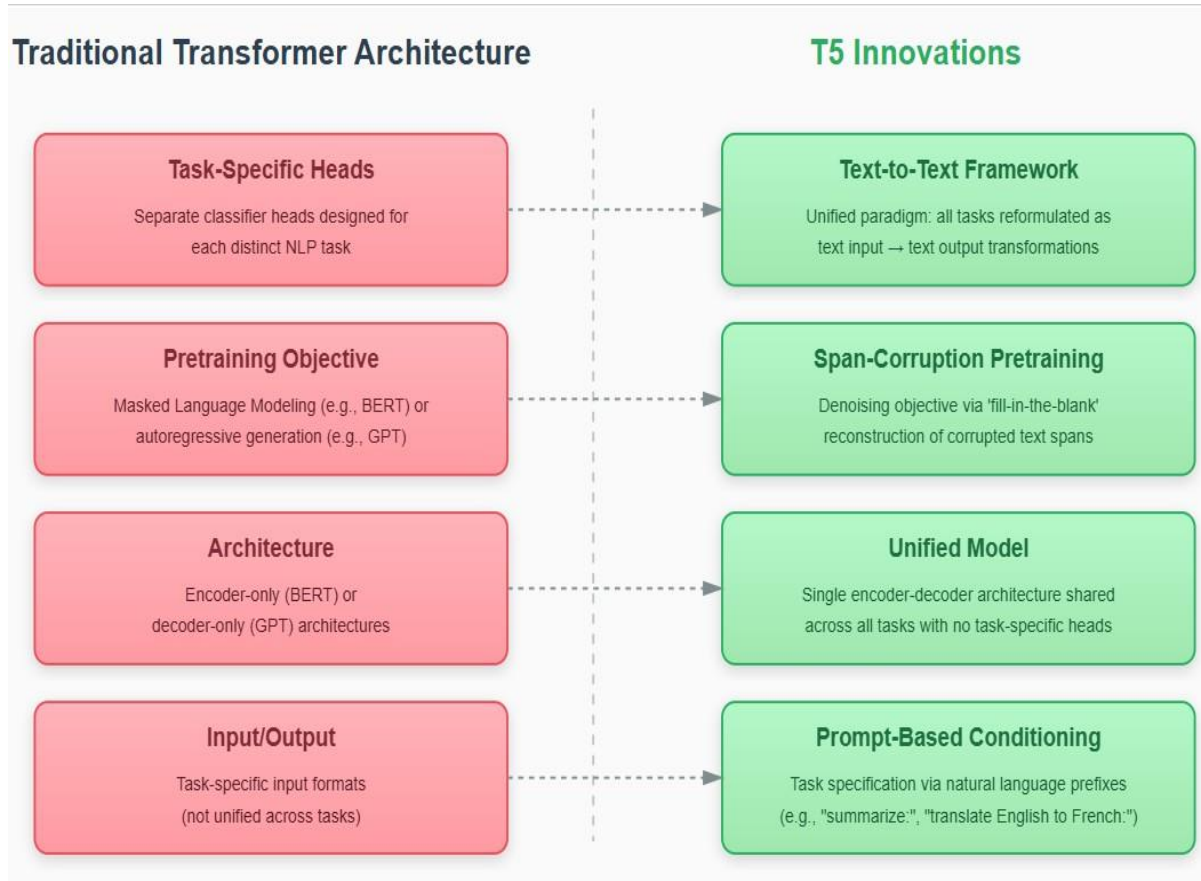


Figure 2. T5 Architectural modifications

- **Pre-Layer Normalization:** Layer normalization is applied immediately before multi-head attention and feedforward sublayers rather than after, with normalization operations positioned outside residual connections. This pre-norm architecture stabilizes training dynamics and enables more effective gradient flow through deep networks.
- **Simplified Layer Normalization:** The normalization operation employs only multiplicative scaling (gain parameter) without additive bias (shift parameter), reducing model complexity while maintaining representational capacity:

$$\text{LayerNorm}(x) = \frac{(x - \mu)}{\sigma \times \gamma} \quad (1)$$

Where: μ and σ represent mean and standard deviation, and

γ is the learned gain vector.

- **Relative Position Embeddings:** Rather than adding learned absolute position embeddings to input token representations, T5 incorporates simplified scalar position biases added directly to attention logits. This relative position encoding scheme enables more flexible handling of variable-length sequences and facilitates length generalization beyond training sequence limits.
- **Comprehensive Dropout Regularization:** Dropout is systematically applied across multiple architectural components including attention weight distributions (after

softmax), feedforward layer activations, embedding layers, and residual connections, with dropout probability $p = 0.09$ providing regularization against overfitting.

The T5 model undergoes pre-training on the Colossal Clean Crawled Corpus (C4), a massive web-extracted dataset comprising approximately 750 gigabytes of cleaned English text, utilizing a span corruption objective that serves as a denoising autoencoding task. This pre-training strategy involves replacing approximately fifteen percent of input tokens with special sentinel tokens, with consecutive corrupted tokens being grouped into spans and represented by single sentinel tokens, thereby requiring the model to reconstruct the original corrupted spans using surrounding contextual information. The span corruption mechanism operates by masking contiguous sequences of tokens and tasking the model with predicting these masked spans, such that an original sequence "Thank you for sharing your perspective on this important issue" becomes corrupted as "Thank you <X> sharing <Y> perspective <Z> this important issue" with the corresponding target reconstruction being "<X> for <Y> your <Z> on". This objective function encourages the development of flexible content generation patterns, robust semantic understanding, and sophisticated paraphrasing capabilities, all of which prove directly transferable to abstractive summarization tasks. The encoder component processes the corrupted input sequence and generates contextual representations, while the decoder autoregressively generates the corruption targets, thereby training the complete encoder-decoder architecture for conditional text generation in an unsupervised manner.

Following the unsupervised pre-training phase, T5 undergoes supervised fine-tuning on downstream tasks using task-specific datasets, with summarization tasks employing datasets such as CNN/Daily Mail formatted as input-output pairs where the input takes the form "summarize: [document text]" and the output consists of the reference summary. This prompt-based formulation explicitly signals the desired task through the "summarize:" prefix, thereby conditioning the model's generation behavior toward summary-appropriate linguistic patterns characterized by conciseness, high-level abstraction, and coherent organizational structure. The text-to-text paradigm embodied by T5 offers substantial advantages for multi-task learning and transfer learning applications. The architectural simplicity achieved through this unified framework enables a single model architecture to handle diverse tasks without requiring task-specific components such as classification heads or span prediction layers, thereby reducing architectural complexity and simplifying implementation. Knowledge acquired during pre-training and fine-tuning on auxiliary tasks including translation and question answering transfers effectively to the target summarization task, facilitating more sample-efficient learning through cross-task knowledge sharing. The text generation framework naturally accommodates variable-length outputs, multiple valid formulations, and open-ended responses without imposing artificial length constraints or requiring fixed label spaces, providing greater flexibility than classification-based approaches. Furthermore, the unified formulation enables task compositionality through mixed-task training batches, allowing multi-task learning that leverages complementary supervision signals across related problems and enhances overall model performance through shared representations and training dynamics.

D. Proposed System Architecture

The complete abstractive summarization system implements a multi-stage pipeline encompassing data acquisition, preprocessing, model training, and inference, as depicted in Fig. 3.



Fig. 3. System Structure

Each stage incorporates domain-specific adaptations optimized for social media comment processing.

- *Stage 1:* Data Collection and Curation involves acquiring raw comment data from social media platforms, filtering low-quality contributions, grouping temporally and thematically related comments, and constructing input-output training pairs from comments and human-authored reference summaries.

- *Stage 2: Preprocessing and Normalization* transforms unstructured, noisy text into standardized formats compatible with neural language processing through systematic cleaning operations, linguistic normalization, and tokenization.
- *Stage 3: Model Initialization and Configuration* loads pre-trained T5 checkpoints, configures task-specific parameters, and prepares the model for domain-adaptive fine-tuning on social media corpora.
- *Stage 4: Fine-Tuning and Optimization* updates model parameters through supervised learning on comment-summary pairs, employing carefully calibrated training procedures to balance overfitting prevention with sufficient domain adaptation.
- *Stage 5: Inference and Summary Generation* applies the trained model to unseen comment collections, generating abstractive summaries through beam search or sampling-based decoding strategies.
- *Stage 6: Evaluation and Quality Assessment* quantifies summary quality across multiple dimensions using reference-based metrics (when gold summaries available), reference-free quality indicators, and human evaluation protocols.

E. Data Collection and Preprocessing

The training data comprises 1,781,576 Facebook news¹ posts and associated user comment threads obtained from publicly available sources, with essential metadata fields including post identifiers, timestamps, and comment text. Initial preprocessing retains three critical fields: *post_title* and *post_num* serving as composite unique identifiers for grouping comments associated with specific news articles, and message containing user-contributed comment text as primary input for summarization. Comments undergo systematic filtering to remove empty or null messages, extremely short comments containing fewer than five words, duplicate or near-duplicate comments detected via fuzzy string matching, spam content identified through heuristic pattern matching including excessive capitalization and repetitive characters, and non-English text for the monolingual model version.

Retained comments undergo comprehensive linguistic normalization including URL replacement with special [URL] tokens to preserve structural information while removing specific addresses, username handling through replacement of @mentions with [USER] tokens to anonymize references while maintaining discourse structure, and emoji processing either converting Unicode emoji characters to textual descriptions or removing semantically non-contributory elements. Common social media abbreviations are expanded using curated dictionaries, transforming expressions such as "tbh" to "to be honest" and "imo" to "in my opinion", while all text undergoes case normalization through conversion to lowercase to reduce vocabulary size and improve generalization. Whitespace standardization collapses multiple consecutive spaces to single spaces with removal of leading and trailing whitespace, and special character handling normalizes excessive punctuation while removing non-standard characters. Comments sharing identical *post_num* identifiers are aggregated into input lists representing complete discussion threads for specific news articles, paired with human-authored reference summaries when available or automatically generated pseudo-summaries for unsupervised pre-training augmentation. Target summary lengths are determined dynamically based on input comment list length using the heuristic formula $\text{target_length} = \min(150, \max(50, 0.15 \times$

¹ <https://github.com/jbencina/facebook-news>

input_length))), ensuring summaries scale appropriately with content volume while maintaining practical length bounds.

F. Model Architecture and Configuration

The implementation employs the T5-small variant comprising 76.9 million total parameters with 35.3 million encoder parameters and 41.6 million decoder parameters. The encoder consists of six layers, each containing eight-head self-attention mechanisms and 2048-dimensional feedforward networks with 512-dimensional embeddings, while the decoder similarly comprises six layers with eight-head masked self-attention, eight-head cross-attention, and 2048-dimensional feedforward networks alongside 512-dimensional embeddings. The model utilizes a vocabulary of 32,000 SentencePiece tokens with maximum sequence lengths of 512 tokens for input and 150 tokens for output. Model initialization loads publicly available pre-trained T5-small checkpoints from the Hugging Face Transformers library, providing a strong starting point through transfer learning from large-scale pre-training on the C4 corpus.

G. Tokenization and Input Processing

Input comment lists and target summaries undergo tokenization via T5's SentencePiece tokenizer, which implements subword tokenization wherein frequent words are represented as single tokens while rare words are decomposed into subword units, enabling open vocabulary representation while controlling vocabulary size. The task prefix "summarize:" is prepended to all inputs to condition generation behavior, with input sequences exceeding 512 tokens truncated from the right and shorter sequences padded to batch maximum length. Binary attention masks distinguish genuine tokens from padding tokens, preventing attention to padding during processing. Output summaries are similarly tokenized with teacher forcing during training, where the decoder receives gold-standard reference summary tokens as inputs shifted right by one position, and target masking ensures loss computation only over genuine summary tokens while ignoring padding positions. The feedforward sublayers implement standard two-layer transformation

$$FFN(x) = \text{ReLU}(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

where:

x : represents the input vector

W_1 : is the weight matrix of size $(d_{model} \times d_{ff})$

b_1 : is the bias vector of size d_{ff} .

W_2 : is the weight matrix of size $d_{ff} \times d_{model}$

b_2 : is the bias vector of size d_{model} .

d_{ff} : is the feedforward network hidden dimension (2048 in T5-Small)

H. Training Protocol and Optimization

Fine-tuning employs supervised learning on comment-summary pairs using the AdamW optimizer with decoupled weight decay, configured with learning rate $\eta = 1 \times 10^{-3}$, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, numerical stability constant $\epsilon = 1 \times 10^{-8}$, and weight decay $\lambda = 0.01$. The learning rate schedule implements linear warmup over 500 steps followed by linear decay to zero over the remaining training duration. Training utilizes

batch size of 16 samples per GPU with gradient accumulation over two steps yielding effective batch size of 32, while regularization applies dropout probability $p = 0.09$ across all dropout layers and label smoothing $\epsilon = 0.1$ to soften one-hot target distributions. Training proceeds for 11 epochs over the full training set employing mixed precision training with FP16 activations and FP32 master weights, reducing memory consumption by approximately fifty percent while maintaining numerical stability through loss scaling. Training minimizes cross-entropy loss, Eq. 3, between predicted token distributions and reference targets

$$L = -\sum_{i=1}^T y_i \log(\hat{y}_i) \quad (3)$$

where:

T is the length of the sequence.

y_i is the target token at position i .

$\hat{y}_i$ is the predicted probability of the token at position i .

Loss computation excludes padding positions through attention masks, ensuring parameter updates reflect only genuine sequence content.

1. Validation and Inference

Model performance is monitored through validation evaluation after each training epoch with the model set to evaluation mode disabling dropout, validation loss computed on a held-out validation set comprising ten percent of data, perplexity calculated as $PPL = \exp(L_{val})$, and early stopping triggered if validation loss fails to improve for three consecutive epochs. At inference time, summaries are generated autoregressively through beam search decoding with beam width of four maintaining four candidate sequences simultaneously, length penalty $\alpha = 0.6$ balancing sequence probability with length through scoring function $\log p / \text{length}^\alpha$, blocking repetition of n -grams with $n \geq 3$ to prevent redundant phrase generation, early stopping when the top beam generates an end-of-sequence token, minimum length of 30 tokens to prevent degenerate short outputs, and maximum length of 150 tokens to enforce conciseness. Alternative sampling-based decoding strategies explored include top-k sampling from the $k = 50$ most probable next tokens, nucleus sampling from the smallest set of tokens whose cumulative probability exceeds $p = 0.95$, and temperature scaling with $T = 0.8$ for slightly more focused generation, though beam search is prioritized for evaluation due to consistent generation of high-quality, coherent summaries.

4. Experimental Setup and Evaluation

A. Implementation and Dataset Configuration

The proposed framework was implemented using PyTorch 1.13.1 with Hugging Face Transformers 4.25.1, trained on an NVIDIA GeForce RTX 4070 GPU with 12 GB VRAM. Training required approximately 18 hours for 11 epochs with memory consumption of approximately 10 GB at batch size 16, while inference achieved approximately 0.8 seconds per comment list averaging 25 comments using beam search decoding.

Training data comprises 1,781,576 Facebook news posts and associated user comments obtained from a publicly available repository. The raw dataset underwent systematic preprocessing including removal of null messages, extremely short comments containing

fewer than five words, duplicates detected via fuzzy matching with 95% similarity threshold, and spam content identified through pattern matching. Language detection filtered non-English content, while text normalization included URL and username tokenization, emoji processing, abbreviation expansion, lowercase conversion, punctuation standardization, and whitespace cleaning. Following preprocessing, the refined dataset comprised 1,196,889 high-quality comments aggregated into 12,634 discussion threads containing between 10 and 500 comments.

The dataset was partitioned using stratified random sampling ensuring balanced representation across thread length distributions, topical categories, and temporal periods, as presented in TABLE I.

Table 1. Dataset Partition Statistics

Subset	Threads	Percentage	Total Comments
Training	10,107	80%	762,783
Validation	1,264	10%	95,348
Test	1,263	10%	95,281
Total	12,634	100%	953,412

B. Model Architecture and Training Protocol

The implementation employs T5-small comprising 76.9 million parameters with six-layer encoder and decoder architectures utilizing eight-head attention mechanisms, 2048-dimensional feedforward networks, and 512-dimensional embeddings with 32,000 SentencePiece vocabulary tokens. Model initialization loaded pre-trained T5-small checkpoints from Hugging Face Transformers, providing transfer learning from C4 corpus pre-training. Hyperparameters were determined through systematic grid search evaluating 48 configurations in preliminary exploration followed by extended training on top candidates.

Fine-tuning employed supervised learning using AdamW optimizer with learning rate $\eta = 1 \times 10^{-3}$, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, weight decay $\lambda = 0.01$, and linear warmup over 500 steps followed by linear decay. Training utilized batch size of 16 samples per GPU with gradient accumulation over two steps yielding effective batch size of 32, while regularization applied dropout probability $p = 0.09$, label smoothing $\epsilon = 0.1$, and gradient clipping with maximum norm of 1.0. Training proceeded for 11 epochs employing mixed precision with FP16 activations and FP32 master weights, reducing memory consumption by approximately 45% while achieving 1.8× training speedup.

Training dynamics exhibited rapid initial learning with training loss decreasing from 2.336 to 1.054 between epochs 1-2, followed by consistent convergence with validation loss declining monotonically from 0.985 to 0.832. The progressive convergence between training and validation losses, with gap narrowing from 1.351 to 0.018, indicated successful generalization without overfitting, while validation perplexity decreased from 2.679 to 2.298.

Table 2. Training Dynamics across Training Epochs

Epoch	Train Loss	Validation Loss	Validation PPL	Learning Rate
1	2.336	0.985	2.679	1.00×10^{-3}
2	1.054	0.916	2.499	9.50×10^{-4}
3	0.988	0.883	2.417	9.00×10^{-4}
4	0.944	0.864	2.373	8.50×10^{-4}
5	0.911	0.851	2.343	8.00×10^{-4}
6	0.883	0.843	2.324	7.50×10^{-4}
7	0.862	0.840	2.316	7.00×10^{-4}
8	0.846	0.835	2.305	6.50×10^{-4}
9	0.830	0.834	2.302	6.00×10^{-4}
10	0.821	0.833	2.301	5.50×10^{-4}
11	0.814	0.832	2.298	5.00×10^{-4}

At inference, summaries were generated autoregressively through beam search with beam width of four, length penalty $\alpha = 0.6$, blocking repetition of trigrams, minimum length of 30 tokens, and maximum length of 150 tokens. ChatGPT-4.0 based on GPT-4 architecture was selected as the primary baseline, representing state-of-the-art general-purpose language models with approximately 175 billion parameters. ChatGPT evaluation employed 200 randomly sampled discussion threads from the test set using standardized prompt template "summarize: [concatenated comment text]" with API configuration specifying model gpt-4, temperature 0.7, and maximum tokens 150.

C. Evaluation Framework and Metrics

Model performance was assessed through comprehensive multi-dimensional evaluation encompassing reference-based metrics, reference-free metrics, and human-aligned quality dimensions. Reference-based evaluation employed the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric suite [26], which remains the predominant benchmark for summarization through operationalizing similarity via n-gram co-occurrence between system-generated and human-authored summaries. Specifically, we computed ROUGE-1 for unigram overlap capturing content word preservation, ROUGE-2 for bigram overlap reflecting phrasal similarity and linguistic fluency, ROUGE-L measuring longest common subsequence to assess sequential order preservation, ROUGE-Lsum as a variant computed at summary level for multi-sentence outputs, ROUGE-S for skip-bigram co-occurrence allowing arbitrary gaps between paired words, and ROUGE-WE [27] extending n-gram matching through word embeddings to enable soft matching whereby semantically similar

but lexically different words receive partial credit based on cosine similarity of their vector representations.

Reference-free evaluation relaxed dependence on human-generated references, enabling more scalable and domain-agnostic assessment. SUPERT [28] employed pre-trained sentence encoders to compute semantic similarity between summaries and source documents by measuring alignment with content clusters, demonstrating particular value in low-resource settings where reference summaries are scarce. Shannon Consistency and Shannon Coherency metrics [29] introduced information-theoretic formulations quantifying content retention through conditional probability of source given summary and structural organization through predictability of summary token sequences respectively. Semantic Distance-based Clustering [30] evaluated quality by computing minimum semantic distance from summary sentences to source content clusters using BERT embeddings, proving especially valuable in noisy social media contexts where reference summaries are nonexistent.

Human-aligned quality dimensions captured aspects correlating with human judgment through specialized natural language processing tools. Consistency was measured using Natural Language Inference models (BART-MNLI) computing entailment probability between source and summary to detect contradictions and fabrications, addressing factual integrity concerns increasingly crucial for abstractive systems [31]. Fluency was assessed via perplexity under GPT-2 language model normalized to unit scale, with lower perplexity indicating greater grammatical correctness and linguistic naturalness. Readability was quantified using Flesch Reading Ease scores normalized to unit scale, measuring text comprehensibility and accessibility. Lexical diversity was calculated through Type-Token Ratio with moving-average normalization for length effects, capturing vocabulary richness and repetition avoidance. Semantic coherence was evaluated using Sentence-BERT embeddings computing average cosine similarity between consecutive sentences, measuring topical consistency across summary sentences. Statistical significance of all performance differences was verified through paired bootstrap resampling tests with 10,000 iterations at 95% confidence intervals, ensuring observed advantages represent genuine improvements rather than random variation.

D. Performance Results

TABLE 3 presents comprehensive ROUGE-based evaluation results comparing the proposed T5-based model against ChatGPT-4.0, with all observed differences achieving statistical significance at $p < 0.05$.

Table 3. Reference-Based Evaluation: ROUGE Metrics Comparison

Metric	Proposed Model	ChatGPT-4.0	Δ (Absolute)	Δ (Relative)
ROUGE-1				
F1	0.3292	0.2739	+0.0553	+20.2%
ROUGE-2				
F1	0.1245	0.0983	+0.0262	+26.7%

Metric	Proposed Model	ChatGPT-4.0	Δ (Absolute)	Δ (Relative)
ROUGE-L				
F1	0.2876	0.2412	+0.0464	+19.2%
ROUGE-Lsum				
F1	0.2891	0.2435	+0.0456	+18.7%
ROUGE-S				
F1	0.1534	0.1287	+0.0247	+19.2%
ROUGE-WE				
Score	0.9568	0.9495	+0.0073	+0.8%

The proposed model demonstrates consistent superiority across all ROUGE variants, with particularly notable ROUGE-2 improvement (26.7%) indicating superior preservation of meaningful bigrams and phrasal structures, substantial ROUGE-1 enhancement (20.2%) reflecting better content word preservation, and strong ROUGE-L and ROUGE-Lsum performance (19.2% and 18.7%) indicating better preservation of sequential ordering and sentence-level structure.

Human-aligned quality metrics reveal substantial advantages for the domain-adapted model across multiple dimensions, presented in TABLE 4.

Table 4. Human-Aligned Quality Metrics Comparison

Metric	Proposed Model	ChatGPT-4.0	Δ (Absolute)	Δ (Relative)
Consistency	0.5627	0.2490	+0.3137	+126.0%
Fluency	0.7876	0.7925	-0.0049	-0.6%
Readability	0.4455	0.3626	+0.0829	+22.9%
Lexical Diversity	0.7234	0.6891	+0.0343	+5.0%
Semantic Coherence	0.4132	0.3948	+0.0184	+4.7%
Overall Coherence	0.4054	0.3914	+0.0140	+3.6%

The most pronounced difference emerges in consistency, with the proposed model achieving 126% relative improvement (0.5627 versus 0.2490, $p < 0.001$), indicating substantially stronger factual alignment with significantly reduced hallucination. The proposed model

demonstrates superior readability (22.9%, $p = 0.003$), enhanced lexical diversity (5.0%, $p = 0.028$), and marginally stronger semantic coherence, while ChatGPT exhibits slight advantages in linguistic fluency (0.6%, not significant). Reference-free evaluation reveals the proposed model achieves SUPERT score of 0.9990 versus 0.9977, Shannon Consistency of 0.9512 versus 0.9447 ($p = 0.031$), and Shannon Coherency of 0.9519 versus 0.9471, while Semantic Distance-based Clustering presents the sole dimension favoring ChatGPT (0.2646 versus 0.5746, $p < 0.001$), suggesting tighter semantic clustering through more conservative paraphrasing versus the proposed model's more aggressive abstraction.

Ablation studies isolate the contribution of domain-specific fine-tuning, presented in TABLE 5.

Table 5. Impact of Fine-Tuning Strategy

Model Variant	ROUGE-1 F1	SUPERT	Consistency
T5-small (zero-shot)	0.1834	0.9123	0.3245
T5-small (general fine-tuned)	0.2567	0.9456	0.4112
T5-small (domain fine-tuned)	0.3292	0.9990	0.5627

Domain-specific fine-tuning proves critical, yielding 28.2% improvement over general fine-tuning and 79.5% improvement over zero-shot performance. Training data scaling reveals logarithmic performance improvement with substantial gains from 25% to 75% of data and diminishing returns beyond 75%. Model size comparison demonstrates T5-base (220M parameters) provides 5.0% ROUGE-1 increase at 2.9× computational cost, while T5-large (770M parameters) yields marginal 1.6% additional gain at 3.2× cost, confirming T5-small offers optimal performance-efficiency trade-off. Stratified performance analysis reveals consistent advantages across thread lengths (14.1-21.6%), topics (19.0-21.8%), and sentiment distributions (19.0-22.2%), with performance degradation on longer threads and highest performance on consensus discussions.

5. Discussion and Conclusion

This study provides compelling evidence that domain-adaptive fine-tuning of moderately sized transformer models can outperform substantially larger general-purpose language models on specialized social media summarization tasks. Through extensive evaluation, the proposed T5-based system with only 77 million parameters consistently surpasses ChatGPT-4.0 across a broad set of lexical, semantic, factual, and human-aligned quality metrics, achieving marked improvements in ROUGE overlap, factual consistency, and readability while operating with orders-of-magnitude greater parameter and cost efficiency. These findings challenge the prevailing assumption that larger foundation models are universally superior and instead underscore the enduring value of task-specific optimization, particularly for noisy, informal, and high-variance domains such as social media discourse.

Beyond raw performance gains, the results reveal a nuanced trade-off between factual grounding and surface-level fluency. Domain-adapted models demonstrate substantially lower hallucination rates and stronger alignment with source content, qualities that are

critical for high-stakes applications such as public opinion monitoring, crisis response, and policy analysis. In contrast, general-purpose models retain marginal advantages in stylistic fluency, reflecting the benefits of massive-scale pre-training on diverse corpora. This distinction suggests that optimal model choice should be guided by application context rather than model scale alone. Importantly, the proposed system's ability to run efficiently on consumer-grade hardware, coupled with its dramatic cost advantages over API-based large language models, highlights its practical viability for large-scale deployment by organizations processing millions of comment threads.

The contributions of this work are threefold. First, it demonstrates that transformer architectures, when carefully adapted, can effectively model the linguistic informality, discourse fragmentation, and high noise-to-signal ratios inherent in social media comments while maintaining semantic fidelity. Second, it introduces a rigorous multidimensional evaluation framework that moves beyond sole reliance on ROUGE, incorporating semantic similarity, factual consistency, and human-centered quality indicators better suited to informal text. Third, it provides the first systematic empirical comparison between a fine-tuned T5 model and ChatGPT-4.0 for social media comment summarization, offering concrete evidence to inform ongoing debates regarding model scale versus domain specialization in applied NLP.

While the findings are robust, limitations remain. The current focus on English-language Facebook comments restricts cross-lingual and cross-platform generalizability, and evolving social media language necessitates periodic retraining to sustain performance. Ethical considerations—including representation bias, privacy preservation, and potential misuse—must be addressed through responsible deployment practices emphasizing transparency, proportionality, and user access to original content. Future research directions include multilingual and multimodal summarization, real-time streaming capabilities, cross-platform adaptation, controllable generation, and further efficiency optimizations to enable edge deployment.

In conclusion, this research advances the state of social media summarization by demonstrating that carefully designed, domain-adaptive transformer systems can deliver superior quality, reliability, and efficiency compared to far larger general-purpose models. In an era defined by information overload and attention scarcity, such systems represent essential infrastructure for transforming vast volumes of online discourse into actionable knowledge. By clarifying the trade-offs between model scale and domain specialization, this work contributes both practical solutions and conceptual insights that will shape the next generation of trustworthy and deployable natural language processing systems for understanding the digital public sphere.

Funding Statement: The publication fees are covered by ELKE at the University of West Attica.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] A. Papagiannopoulou, and C. Angeli, "Summarizing User Comments on Social Media Using Transformers", In 11th European Conference Proceedings on Social Media (ECSM-2024), 198-205, ACI, Brighton, UK, 2024.
- [2] H. P. Luhn, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, vol. 2, no. 2, 1958, pp. 159-165.

- [3] S. Gupta, and S. K. Gupta, "Abstractive Summarization: An Overview of the State of The Art", *Expert Systems with Applications* 121, 2019, pp. 49–65.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.
- [5] Q. Li, and Q. Zhang, "Abstractive Event Summarization on Twitter", In *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*, Association for Computing Machinery: 22–23, 2020.
- [6] Z. Kerui, H. Haichao, L. Yuxia, "Automatic text summarization on social media", *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2020.
- [7] I. S. Blekanov, N. Tarasov and S. S. Bodrunova, "Transformer-Based Abstractive Summarization for Reddit and Twitter: Single Posts vs. Comment Pools in Three Languages", *Future Internet* 14, 69, 2022.
- [8] S. Gao, X. Chen, P. Li, Z. Ren, L. Bing, D. Zhao, and R. Yan, "Abstractive Text Summarization by Incorporating Reader Comments", In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019, 33(01): 6399-6406.
- [9] P. Bhandarkar, K. T. Thomas, "Text Summarization Using Combination of Sequence-To-Sequence Model with Attention Approach", *Springer Science and Business Media Deutschland GmbH*, 2023, 283–293.
- [10] Z. Liang, J. Du, and C. Li, "Abstractive social media text summarization using selective reinforced Seq2Seq attention model", *Neurocomputing*, 410, 2020, 432–440.
- [11] Q. Wang, and J. Ren, "Summary-aware attention for social media short text abstractive summarization", *Neurocomputing*, 425, 2021, 290–299.
- [12] A. Gupta, D. Chugh, and R. Katarya, "Automated News Summarization Using Transformers", In *Sustainable Advanced Computing*, Volume 840, 2022. ISBN: 978-981-16-9011-2.
- [13] J. Devlin, M-W. Chang, K. Lee, K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding", *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [14] M. H. Su, C. H. Wu, H. T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization", *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28, 2020, 2061–2072.
- [15] D. Singhal, K. Khatte, A. Tejaswini, R. Jayashree, "Abstractive Summarization of Meeting Conversations", In *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, Institute of Electrical and Electronics Engineers Inc.
- [16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narag, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *The Journal of Machine Learning Research*, Volume 21, Issue 1, 2019, 140:1-140:67.
- [17] A. Pal, L. Fan, V. Igodifo, "Text Summarization using BERT and T5", https://anjali001.github.io/Project_Report.pdf, 2022.
- [18] Nguyen, M. T., Nguyen, V. C., Vu, H. T., Nguyen, V. H. (2020) Transformer-based Summarization by Exploiting Social Information. in *Proceedings - 2020 12th International Conference on Knowledge and Systems Engineering, KSE 2020*, Institute of Electrical and Electronics Engineers Inc.: 2020, 25–30.

- [19] R. Rawat, P. Rawat, V. Elahi, and A. Elahi, "Abstractive Summarization on Dynamically Changing Text", In 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1158-1163.
- [20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing", *ACM Computing Surveys* 55, 2021, 1-35.
- [21] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, et al, "OpenPrompt: An Open-source Framework for Prompt-learning", In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2022, pp. 105-113.
- [22] X. L. Li, and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation", In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, 4582–4597.
- [23] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning", In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, 3045–3059.
- [24] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, C. Yue, H. Zhang, Y. Liu, X. Li, B. Ge, D. Zhu, Y. Yuan, D. Shen, T. Liu, and S. Zhang, "Prompt Engineering for Healthcare: Methodologies and Applications", <https://arxiv.org/html/2304.14670v2>, 2024.
- [25] A. Papagiannopoulou, and C. Angeli, "Designing a summarization system on social comments using transformers", In 12th International Conference Proceedings on Computer Technologies and Development (TechDev), 2023, 1-5, IEEE, Rome, Italy.
- [26] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", In Text Summarization Branches Out, 2004, 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [27] J-P. Ng, and V. Abrecht, "Better Summarization Evaluation with Word Embeddings for ROUGE", In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal. Association for Computational Linguistics, 2015, 1925–1930.
- [28] Y. Gao, W. Zhao, S. Eger, "SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pages 1347–1354, Association for Computational Linguistics.
- [29] N. Egan, O. Vasilyev, J. Bohannon, "Play the Shannon game with language models: a human-free approach to summary evaluation", Proceedings of the AACL conference on artificial intelligence, 2022, 10599–10607.
- [30] Y. Liu, Q. Jia, K. Zhu, "Reference-free summarization evaluation via semantic correlation and compression ratio". Proceedings of the 2022 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, 2022b, 2109–2115.
- [31] W. Kryściński, B. McCann, C. Xiong, et al, "Evaluating the factual consistency of abstractive text summarization", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pages 9332–9346, Association for Computational Linguistics.