The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 1, February 2026
https://doi.org/10.33022/ijcs.v15i1.5086

IPTSS: Intelligent Preprocessing and Multi-Representation Analysis for Social Media Text Summarization with Clustering-Based Enhancement

Fahd A. Ghanem[1], M. C. Padma [2], Wadeea R. Nji[3]
fahd.a.ghanem@gmail.com[1], padmapes@gmail.com[2] , wadeearashad@gamil.com[3]
[1,2]Department of Computer Science & Engineering, PES College of Engineering, (Affiliated to University of Mysore), Mandya 571401, Karnataka, India
[3]Department of Studies in Computer Science, University of Mysore, 570006, India

Article Information

Received : 30 Jan 2026
Revised : 12 Feb 2026
Accepted : 20 Feb 2026

Keywords

IPTSS, social media summarization, extractive summarization, intelligent preprocessing, text representation, TF-IDF-weighted BERT, clustering-based summarization, social media mining, short text summarization

Abstract

Social media platforms generate massive volumes of noisy, informal short texts, creating significant challenges for automatic text summarization. This paper presents IPTSS (Intelligent Preprocessing and Transformation System for Social Media Summarization), a unified framework that integrates intelligent preprocessing, multi-representation text modeling, and clustering-based extractive summarization into a single end-to-end pipeline. IPTSS incorporates a four-stage intelligent preprocessing pipeline for redundancy elimination, platform-noise removal, out-of-vocabulary normalization, and linguistic standardization, a multi-representation analysis layer spanning statistical, distributional, and transformer-based models, and a hybrid TF-IDF–weighted BERT representation that fuses corpus-specific lexical importance with contextual semantic information. Summarization is performed through clustering-based representative selection with redundancy control to ensure topical diversity and coverage. Extensive experiments on large-scale datasets collected from X (formerly Twitter) across the Monkeypox, COVID-19 Vaccine, and Climate Change domains demonstrate that preprocessing alone yields a 25.8% improvement in ROUGE-1, while representation sophistication produces a 38.4% gain from Bag-of-Words to Sentence-BERT. The proposed hybrid representation further improves performance by 7.0% over the best single-representation baseline, achieving the highest scores across all ROUGE metrics. The optimal configuration (Fuzzy C-Means + IPTSS Hybrid) reaches ROUGE-1 = 0.528, outperforming state-of-the-art statistical, graph-based, crisis-specific, neural, and optimization-based methods. Cross-dataset validation confirms strong generalizability, with low performance variance (CV ≈ 2.5%) across heterogeneous domains without dataset-specific tuning. These results demonstrate that effective social media summarization is driven primarily by preprocessing quality and hybrid representation design rather than algorithmic complexity alone, establishing IPTSS as a robust, scalable, and generalizable framework for large-scale social media extractive summarization.

A. Introduction

Social media platforms have fundamentally transformed global information ecosystems, enabling rapid and large-scale dissemination of information that strongly influences public opinion, societal behaviour, and public discourse. The massive volume of short user-generated messages produced daily across social platforms creates unprecedented opportunities for monitoring emerging events, health crises, policy responses, and misinformation diffusion, while simultaneously introducing major challenges for automated information analysis and management [1]. The unstructured and dynamic nature of social media data makes effective processing increasingly complex, particularly for large-scale analytical tasks such as automatic text summarization.

Unlike traditional textual sources such as news articles and academic documents, social media content exhibits extreme linguistic variability, including informal grammar, creative spellings, abbreviations, slang, emojis, hashtags, and platform-specific artifacts such as user mentions and hyperlinks [2]. Character-length constraints encourage compressed expressions, contextual assumptions, and fragmented language structures, while informal communication introduces grammatical errors, inconsistent punctuation, multilingual mixing, and rapidly evolving terminology. These characteristics create a substantial gap between raw social media text and the clean, standardized input expected by conventional natural language processing (NLP) systems trained primarily on formal corpora such as news articles and academic publications [3]. This mismatch significantly degrades the effectiveness of downstream analytical models.

Automatic text summarization has emerged as an effective approach for managing large-scale information overload by distilling large volumes of textual content into concise and informative representations [4]. In social media contexts, extractive summarization is particularly suitable, as it preserves original user expressions, maintains factual integrity, and avoids the hallucination risks increasingly associated with generative abstractive approaches. However, existing summarization systems often fail to perform robustly on social media data due to their sensitivity to noise, redundancy, linguistic informality, and inconsistent text structure. Most summarization research has focused primarily on model architectures and algorithmic techniques, including neural networks, attention mechanisms, transformer models, and graph-based methods, while treating preprocessing and representation design as secondary components or fixed preliminary steps [5]. As a result, many systems remain fragile when applied to noisy, large-scale social media streams.

Recent advances in representation learning, particularly transformer-based models, have demonstrated strong capabilities in capturing contextual semantics in unstructured text. Models such as BERT and its variants have improved performance across many NLP tasks by learning deep contextual representations [6], while sentence-level models such as Sentence-BERT enable efficient semantic similarity modeling for clustering and retrieval tasks [7]. However, empirical studies have shown that even powerful transformer models remain sensitive to input quality and benefit significantly from structured preprocessing and normalization, especially in social media domains characterized by high noise and linguistic variability [8], [9]. Similarly, hybrid representation strategies that

integrate lexical importance with contextual semantics have shown promise in improving performance in specialized domains, indicating the importance of combining complementary representation signals rather than relying on single-model approaches [10].

At the same time, clustering-based extractive summarization approaches have gained increasing attention due to their ability to group semantically related content and select representative messages, ensuring topical diversity and coverage [11]. Compared to graph-based methods such as TextRank and LexRank [12], [13], clustering-based methods provide scalable mechanisms for large-scale summarization by enabling efficient topic grouping and representative selection. However, their effectiveness is strongly dependent on the quality of text preprocessing and representation, which remain underexplored in social media summarization research.

Despite these advances, existing literature lacks integrated frameworks that systematically combine intelligent preprocessing, hybrid representation modeling, and clustering-based summarization within a unified architecture designed specifically for large-scale social media text. In response to this gap, this paper introduces IPTSS (Intelligent Preprocessing and Transformation System for Social Media Summarization), a unified framework that integrates structured preprocessing, multi-representation text modeling, and clustering-based extractive summarization into a single analytical pipeline. The framework incorporates intelligent preprocessing to address redundancy, platform-specific noise, informal language, and linguistic inconsistency, combined with multi-representation modeling spanning statistical and transformer-based embeddings, including a TF-IDF-weighted BERT representation that fuses domain-specific lexical importance with contextual semantic understanding. Summarization is performed through clustering-based representative selection to ensure topic coverage and diversity. The proposed approach is evaluated on datasets collected from X (formerly Twitter) across multiple domains, including Monkeypox, COVID-19 Vaccine, and Climate Change, demonstrating consistent performance improvements across heterogeneous social media contexts. By integrating preprocessing, representation, and summarization as interdependent components, IPTSS provides a robust and scalable framework for accurate and representative social media text summarization.

B. Research Method

This section presents the proposed IPTSS framework, which consists of six main phases: data collection, intelligent preprocessing, multi-representation analysis, clustering-based summarization, summary generation, and evaluation, as illustrated in Figure 1. These phases form a unified end-to-end processing pipeline that transforms raw social media content into structured extractive summaries. Each phase is designed to perform a specific functional role, and their integration enables systematic noise reduction, robust representation learning, effective content grouping, and accurate representative selection for summary generation.

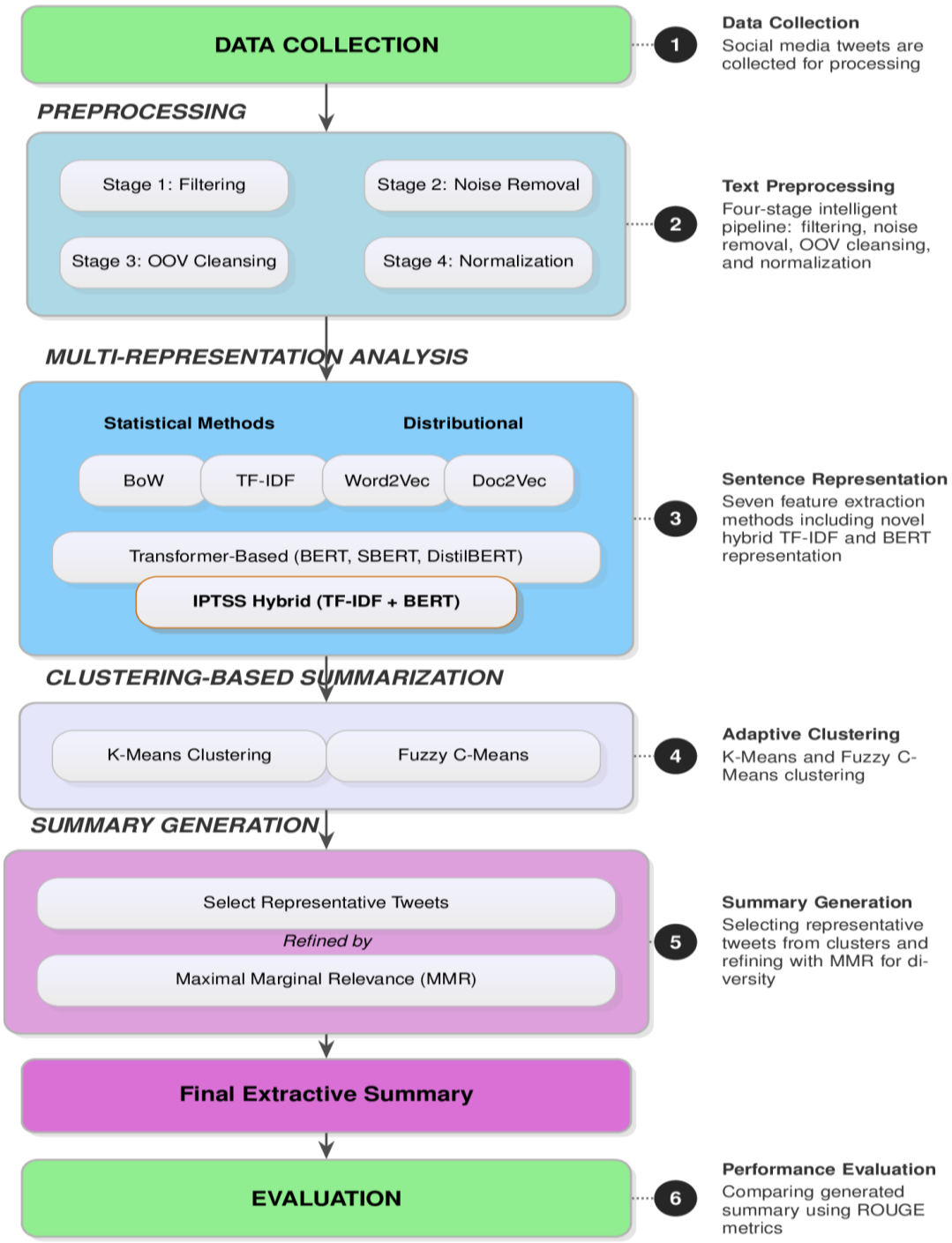

Figure 1. Proposed IPTSS framework for social media extractive summarization

Phase 1: Intelligent Preprocessing Pipeline

The intelligent preprocessing phase transforms raw social media text into analysis-ready representations through four sequential stages, as illustrated in Figure 1. Preprocessing is a critical component in social media summarization pipelines because microblog content violates many assumptions underlying traditional NLP systems, including lexical stability, grammatical regularity, and semantic consistency [14],[15]. Prior research has shown that preprocessing

choices significantly influence downstream representation learning and summarization performance, particularly in sparse and noisy short-text environments [16], [17]. The pipeline is designed to be configurable, allowing stages to be adapted according to domain and application requirements.

The first stage performs extraction and filtering to control redundancy and enforce linguistic consistency. Exact duplicate messages arising from bot activity, automated reposting, and viral sharing are removed using SHA-256 cryptographic hashing with constant-time lookup, enabling efficient large-scale processing. Near-duplicate messages, which differ only by minor orthographic variations, are identified using character-level trigram Jaccard similarity with a similarity threshold of 0.85, providing robustness to informal spelling and punctuation variation. In addition, probabilistic language detection based on character n-gram models is applied to filter non-target languages and ensure monolingual consistency, a necessary step for stable representation learning in multilingual social media streams [18].

The second stage removes platform-induced noise while preserving semantically informative content. URLs and user mentions are filtered to prevent vocabulary inflation and semantic distortion, while compound hashtags are segmented into constituent lexical units using probabilistic segmentation models trained on large web corpora, enabling recovery of embedded topical information. Semantic transformation of hashtags rather than simple removal has been shown to improve representation quality and topic coherence in social media analytics [19],[20]. Special characters, excessive punctuation, and decorative symbols are removed, and whitespace is normalized to ensure structural consistency of the textual representation.

The third stage addresses out-of-vocabulary (OOV) phenomena caused by informal spelling, slang, abbreviations, and emerging terminology. Informal contractions are expanded to standardized forms, slang and abbreviations are normalized using curated social-media lexicons, and expressive character elongation is reduced through controlled regex normalization. These operations reduce lexical sparsity and improve the stability of both statistical and neural text representations, which is critical for modeling short social media texts [21].

The final stage applies linguistic post-transformation to produce standardized textual representations suitable for downstream processing. Tokenization is performed using rule-based linguistic conventions, followed by case normalization to reduce vocabulary sparsity. Stopword removal and Porter stemming are applied to consolidate morphological variants and emphasize content-bearing terms, supporting statistical and hybrid representation models. Together, these four stages form a compact and structured preprocessing pipeline that provides a stable foundation for multi-representation analysis and clustering-based extractive summarization in large-scale social media environments[22].

Phase 2: Multi-Representation

The IPTSS framework incorporates a multi-representation analysis layer that implements seven distinct feature extraction methods, enabling systematic comparison across statistical, distributional, and transformer-based representation paradigms. This design is motivated by the well-established

observation that different representation models encode complementary linguistic properties, and that representation choice critically influences clustering quality and extractive summarization performance. By integrating heterogeneous representation families within a unified pipeline, IPTSS enables robust modeling of lexical salience, semantic similarity, and contextual meaning in noisy short-text social media environments. Table 1 summarizes the seven representation methods employed in IPTSS, together with their dimensionality, representation type, and key characteristics.

Table 1. IPTSS Multi-Representation Analysis: Feature Extraction Methods

Category	Method	Dimension	Type	Key Characteristics
Statistical	Bag-of-Words	5,000	Sparse	Term frequency encoding
Statistical	TF-IDF	5,000	Sparse	Document–corpus importance
Distributional	Word2Vec	300	Dense	Skip-gram context prediction
Distributional	Doc2Vec	300	Dense	Direct document embedding
Transformer	DistilBERT	768	Dense	Knowledge distillation
Transformer	BERT	768	Dense	Bidirectional contextual encoding
Transformer	Sentence-BERT	384	Dense	Semantic similarity optimization
IPTSS Hybrid	TF-IDF + BERT	768	Dense	Lexical and semantic fusion

1. Statistical Text Representations

Traditional statistical representations are implemented using Bag-of-Words (BoW) [23] and TF-IDF [24] models, with the vocabulary restricted to the 5,000 most frequent terms to control sparsity. In the BoW model, each document d is represented as a sparse vector:

$$\mathrm{BoW}(d) = [c(t_1, d), c(t_2, d), \ldots, c(t_{|V|}, d)] \tag{1}$$

where $c(t_i, d)$ denotes the frequency of term t_i in document d. TF-IDF extends this formulation by weighting terms according to their importance within the document and across the corpus:

$$\mathrm{TFIDF}(t, d, \mathcal{D}) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t, \mathcal{D}), \text{ with}$$

$$\mathrm{IDF}(t, \mathcal{D}) = \log\left(\frac{|\mathcal{D}|}{|\{d \in \mathcal{D}: t \in d\}|}\right).$$

This weighting scheme enhances discriminative power by emphasizing domain-specific vocabulary while down-weighting ubiquitous terms that carry limited semantic value.

2. Distributional Word Embeddings

Distributional representations are implemented using Word2Vec [25] and Doc2Vec [26] embeddings with dimensionality fixed at 300 and context window size set to 5. Word2Vec learns dense semantic word representations using the Skip-gram objective:

$$\max \sum_{t} \sum_{\substack{-c \le j \le c \\ j \ne 0}} \log P(w_{t+j} \mid w_t), \tag{2}$$

where the conditional probability is defined as

$$P(w_o \mid w_i) = \frac{\exp\left(\mathbf{v}_{w_o}^{\top} \mathbf{v}_{w_i}\right)}{\sum_{w \in V} \exp\left(\mathbf{v}_{w}^{\top} \mathbf{v}_{w_i}\right)}. \tag{3}$$

Document-level embeddings are obtained by mean pooling the word vectors:

$$\mathrm{Doc}(d) = \frac{1}{|d|} \sum_{i=1}^{|d|} \mathbf{v}_{w_i}. \tag{4}$$

Doc2Vec directly learns document embeddings by incorporating a document vector into the prediction process:

$$\max \log P(w_t \mid d, w_{t-k}, \ldots, w_{t-1}), \tag{5}$$

with prediction computed as

$$P(w_t \mid \text{context}) = \mathrm{softmax}\left(\mathbf{W}[\mathbf{v}_d, \mathbf{v}_{w_{t-k}}, \ldots, \mathbf{v}_{w_{t-1}}]\right). \tag{6}$$

These representations enable recognition of semantic similarity between lexically distinct but conceptually related terms, which is particularly important in crisis communication contexts.

3. Transformer-Based Contextual Representations

Transformer-based representations are implemented using BERT [6], Sentence-BERT [7], and DistilBERT [27]. BERT produces contextualized token embeddings through multi-head self-attention:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V. \tag{7}$$

Document embeddings are obtained by mean pooling the final-layer token states:

$$\mathbf{E}_{\mathrm{BERT}}(d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}_i^{(L)}. \tag{8}$$

Sentence-BERT fine-tunes BERT using siamese network architectures to optimize semantic similarity, computed via cosine similarity:

$$\text{sim}(u, v) = \frac{u \cdot v}{\| u \| \| v \|}. \tag{9}$$

DistilBERT provides a computationally efficient alternative through knowledge distillation, minimizing the combined loss:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}}(y, \hat{y}_s) + \beta \mathcal{L}_{\text{KD}}(\hat{y}_t, \hat{y}_s), \tag{10}$$

where $\hat{y}_t$ and $\hat{y}_s$ denote teacher and student outputs, respectively.

4. IPTSS Hybrid TF-IDF–BERT Representation

To exploit complementary lexical and semantic signals, IPTSS introduces a hybrid TF-IDF–BERT representation. Let document d be represented by a TF-IDF vector $\mathbf{x}_{\text{TFIDF}}(d) \in \mathbb{R}^{5000}$ and a BERT embedding $\mathbf{x}_{\text{BERT}}(d) \in \mathbb{R}^{768}$. Dimensional alignment is first performed using principal component analysis:

$$\mathbf{x}'_{\text{TFIDF}}(d) = \text{PCA}(\mathbf{x}_{\text{TFIDF}}(d), 768). \tag{11}$$

Both representations are then L2-normalized:

$$\hat{\mathbf{x}}_{\text{TFIDF}}(d) = \frac{\mathbf{x}'_{\text{TFIDF}}(d)}{\| \mathbf{x}'_{\text{TFIDF}}(d) \|_2}, \hat{\mathbf{x}}_{\text{BERT}}(d) = \frac{\mathbf{x}_{\text{BERT}}(d)}{\| \mathbf{x}_{\text{BERT}}(d) \|_2}. \tag{12}$$

The final hybrid representation is defined as:

$$\mathbf{x}_{\text{hybrid}}(d) = \alpha \hat{\mathbf{x}}_{\text{TFIDF}}(d) + (1 - \alpha) \hat{\mathbf{x}}_{\text{BERT}}(d), 0 \leq \alpha \leq 1, \tag{13}$$

where the weighting parameter α controls the relative contribution of lexical importance and contextual semantic information. The optimal value of α is determined through grid-search optimization.

This hybrid formulation enables IPTSS to jointly exploit corpus-specific vocabulary signals and deep contextual semantics, yielding a more expressive representation that improves clustering robustness and extractive summarization quality in noisy social media environments.

Clustering-Based Summarization

The IPTSS framework adopts a clustering-based summarization to organize short social media texts into coherent topical groups and to promote diversity in the generated summaries. Clustering is particularly well suited for short text summarization, as it groups semantically related tweets while reducing redundancy caused by retweets, paraphrases, and repeated reports of the same event. By selecting representative tweets from each cluster, the framework

ensures that summaries capture multiple aspects of an event rather than overemphasizing a single dominant topic.

1. K-Means Clustering

K-Means clustering partitions a collection of tweets into k disjoint clusters by minimizing intra-cluster variance. Given a set of tweet representations $\{x_1, x_2, \ldots, x_n\}$, the clustering objective is defined as:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mathbf{c}_i \|^2, \tag{14}$$

where C_i denotes the set of tweets assigned to cluster i, and $\mathbf{c}_i$ represents the centroid of that cluster. After convergence, representative tweet selection is performed by identifying the tweet closest to each cluster centroid:

$$r_i = \arg \min_{x \in C_i} \| x - \mathbf{c}_i \|. \tag{15}$$

These representatives serve as candidate summary sentences, ensuring that each cluster contributes a distinct piece of information to the final summary.

2. Fuzzy C-Means Clustering

To better handle the inherent ambiguity and topic overlap common in short social media texts, IPTSS also employs Fuzzy C-Means (FCM) clustering. Unlike hard clustering, FCM assigns each tweet a degree of membership to multiple clusters, allowing tweets that reference multiple aspects of an event to be represented more flexibly. The FCM objective function is defined as:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \| x_i - \mathbf{c}_j \|^2, \tag{16}$$

where $u_{ij} \in [0,1]$ denotes the membership degree of tweet x_i in cluster j, subject to the constraint

$$\sum_{j=1}^{k} u_{ij} = 1, \tag{17}$$

and $m > 1$ is the fuzziness parameter controlling the softness of cluster assignments. Representative tweets are selected based on maximum membership values within each cluster, ensuring that the most semantically central tweets are chosen while preserving topical overlap where appropriate.

Across all clustering variants, the primary function of clustering within IPTSS is to enhance summary diversity and reduce redundancy. By grouping semantically similar tweets and selecting representative instances from each group, the framework constructs summaries that reflect multiple facets of a disaster event, including initial reports, impact assessments, and response actions. This clustering-based strategy is particularly effective for short text summarization, where individual posts are sparse in content but collectively form rich and evolving narratives.

Summary Generation

After clustering semantically similar sentences, the next objective is to construct a final summary by selecting the most informative and diverse sentences. In the IPTSS framework, summary generation follows a two-stage strategy. First, representative sentences are identified from each cluster based on centroid similarity, ensuring topical coverage. Second, a redundancy reduction mechanism based on Maximal Marginal Relevance (MMR) [28] is applied to penalize semantic overlap and promote diversity among selected sentences.

1. Centroid-Based Representative Selection

Given a cluster C_i with centroid vector $\mathbf{c}_i$, candidate sentences are ranked according to their similarity to the centroid. The centroid represents the central semantic theme of the cluster, and selecting sentences closest to it ensures that the most representative information is retained. For a sentence $s \in C_i$ with embedding $\mathbf{x}_s$, centroid similarity is computed as:

$$\mathrm{Sim}_{\mathrm{centroid}}(s, C_i) = \cos(\mathbf{x}_s, \mathbf{c}_i), \tag{18}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. The top-ranked sentences from each cluster form an initial candidate pool for summary construction.

2. Redundancy Reduction Using MMR

Although centroid-based selection ensures relevance, it may introduce redundancy due to semantic similarity among candidate sentences, particularly in highly repetitive social media data. To address this issue, MMR is applied as a post-ranking strategy to balance relevance and novelty. Let S denote the set of sentences already selected for the summary and let R represent the set of remaining candidate sentences. For each candidate sentence $s \in R$, the MMR score is computed as:

$$\mathrm{MMR}(s) = \lambda \cdot \mathrm{Sim}(s, \mathbf{q}) - (1 - \lambda) \cdot \max_{s' \in S} \mathrm{Sim}(s, s'), \tag{19}$$

where:

- $\mathrm{Sim}(s, \mathbf{q})$ measures the relevance of sentence s to the summary objective $\mathbf{q}$(represented by the cluster centroid or global document representation),

- $\mathrm{Sim}(s, s')$ measures similarity between candidate sentence s and an already selected sentence s',
- $\lambda \in [0,1]$ controls the trade-off between relevance and diversity.

At each iteration, the sentence with the highest MMR score is selected and added to S. This process continues until the desired summary length is reached.

Datasets and Evaluation Metrics

The performance of the proposed IPTSS framework is evaluated using three large-scale Twitter datasets drawn from both health-related and environmental domains. These datasets were deliberately selected to reflect diversity in topical focus, temporal scope, and linguistic characteristics, thereby enabling comprehensive assessment of IPTSS under varying short-text summarization conditions. Table 2 summarizes the key statistics of the datasets used in the evaluation.

Evaluation is conducted using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a standard metric for automatic text summarization. ROUGE-1 measures unigram overlaps between generated summaries and reference summaries, ROUGE-2 measures bigram overlaps to assess local coherence, and ROUGE-L computes the longest common subsequence to capture structural similarity and sequential alignment. For each dataset, reference summaries consist of 100 manually annotated tweets. Inter-annotator agreement was measured using Cohen's kappa coefficient, achieving values greater than 0.75, indicating substantial agreement and reliable ground truth annotations.

Table 2. Dataset Characteristics for IPTSS Evaluation

Dataset	Collection Period	Raw Tweets	After IPTSS
Monkeypox	May-Aug 2022	103,266	61,433
COVID-19 Vaccine	2021-2022	87,452	54,127
Climate Change	2023	52,318	38,902
Total	**-**	243,036	154,462

C. Result and Analysis

This section presents comprehensive experimental results evaluating each IPTSS module through systematic ablation studies, comparative analyses, and statistical significance testing.

1. IPTSS Preprocessing Impact Analysis

Table 3 presents ablation results quantifying each preprocessing stage's contribution to summarization performance. All configurations use Sentence-BERT

representation with K-Means clustering to isolate preprocessing effects from representation and algorithm choices.

Table 3. IPTSS Preprocessing Pipeline Ablation Study (Monkeypox Dataset)

Configuration	Corpus	Vocab	OOV%	R-1	Δ R-1
Raw (no IPTSS)	103,266	47,832	18.3%	0.387	-
Stage 1 only	61,433	41,256	17.8%	0.412	+6.5%
Stages 1-2	61,433	32,156	16.2%	0.438	+13.2%
Stages 1-3	61,433	25,043	3.2%	0.461	+19.1%
Full IPTSS (1-4)	**61,433**	**12,487**	**3.2%**	**0.487**	**+25.8%**

The results show a clear and progressive improvement across preprocessing stages. Stage 1 (Extraction and Filtering) increases ROUGE-1 from 0.387 to 0.412 (+6.5%), driven by duplicate and near-duplicate removal, which reduces the corpus size by 40.5% (103,266 → 61,433 tweets). This confirms that redundancy severely distorts frequency statistics and degrades downstream representation quality and content selection.

Stage 2 (Platform Noise Removal) further improves ROUGE-1 to 0.438 (+13.2% cumulative), supported by substantial vocabulary reduction (47,832 → 32,156) while preserving semantic content through intelligent hashtag segmentation. This demonstrates that removing platform artifacts while retaining topical signals improves both lexical and semantic representations.

Stage 3 (OOV Cleansing) produces a major reduction in out-of-vocabulary terms from 17.8% to 3.2%, yielding ROUGE-1 = 0.461 (+19.1% cumulative). This confirms that informal language normalization substantially improves embedding quality for both statistical and neural representations. Stage 4 (Linguistic Post-Transformation) provides the final improvement, increasing ROUGE-1 to 0.487 (+25.8% cumulative) while consolidating vocabulary from 25,043 to 12,487 unique terms.

Overall, the full IPTSS pipeline achieves a 25.8% cumulative improvement in ROUGE-1, demonstrating strong synergistic effects across preprocessing stages rather than isolated gains from individual operations. Figure 2 illustrates the progressive performance improvement across the four preprocessing stages.

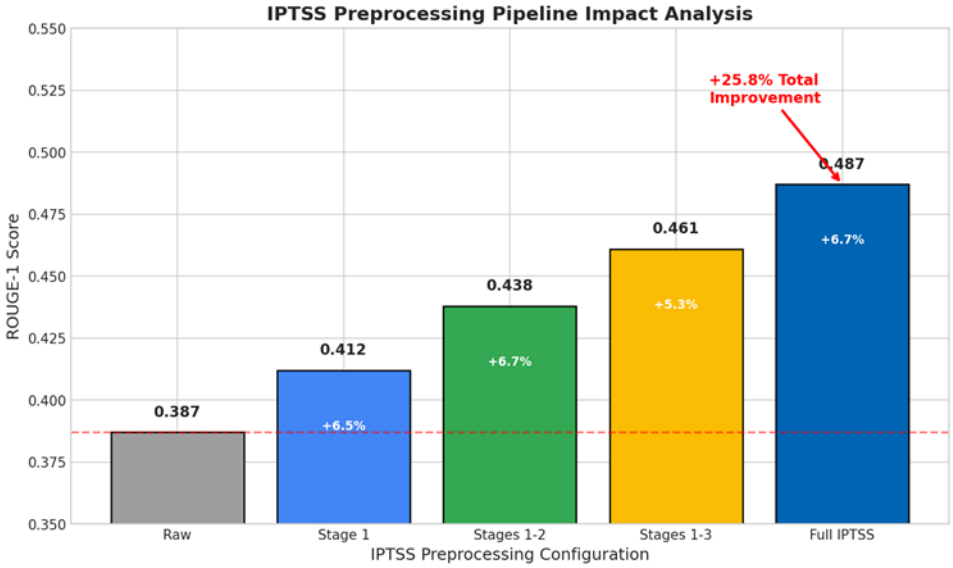

Figure 2. IPTSS Preprocessing Pipeline Impact Analysis showing progressive ROUGE-1 improvement across stages with cumulative 25.8% gain over raw baseline.

2. Multi-Representation Analysis Results

Table 4 compares eight representation methods using the fully preprocessed Monkeypox dataset with K-Means clustering (k = 50), providing a controlled evaluation in which preprocessing and clustering parameters are fixed to isolate the effect of representation design.

A clear performance hierarchy emerges across paradigms, progressing from statistical representations to distributional embeddings and transformer-based models, and culminating in the proposed hybrid approach. Statistical methods show the lowest performance (Bag-of-Words: R-1 = 0.352, TF-IDF: R-1 = 0.421), followed by distributional representations (Word2Vec: R-1 = 0.438, Doc2Vec: R-1 = 0.445).

Transformer-based models achieve further improvements (DistilBERT: R-1 = 0.471, BERT: R-1 = 0.479, Sentence-BERT: R-1 = 0.487), confirming the importance of contextual semantic modeling for short-text summarization. The overall improvement from Bag-of-Words to Sentence-BERT represents a 38.4% relative gain in ROUGE-1, quantifying the impact of representation sophistication on summarization quality.

The proposed IPTSS Hybrid representation achieves the best performance across all metrics (R-1 = 0.521, R-2 = 0.247, R-L = 0.485), outperforming Sentence-BERT by 7.0% in ROUGE-1 and TF-IDF by 23.8%, demonstrating that combining lexical salience with contextual semantic modeling captures complementary information unavailable to single-paradigm representations. This confirms that domain-specific lexical precision and contextual semantic equivalence jointly contribute to improved content selection in social media summarization.

Table 4. Representation Method Performance Comparison (Preprocessed Data, K-Means)

Method	Dim	R-1	R-2	R-L
Bag-of-Words	5,000	0.352	0.085	0.282
TF-IDF	5,000	0.421	0.198	0.392
Word2Vec	300	0.438	0.206	0.408
Doc2Vec	300	0.445	0.210	0.414
DistilBERT	768	0.471	0.223	0.438
BERT	768	0.479	0.227	0.446
Sentence-BERT	384	0.487	0.231	0.453
IPTSS Hybrid	**768**	**0.521**	**0.247**	**0.485**

Sentence-BERT also consistently outperforms full BERT despite lower dimensionality (384 vs. 768), indicating that task-specific fine-tuning for semantic similarity provides greater benefit for clustering-based summarization than raw model capacity alone. Figure 3 illustrates the progressive performance trend across representation paradigms, with the IPTSS Hybrid representation achieving the highest ROUGE-1, ROUGE-2, and ROUGE-L scores.

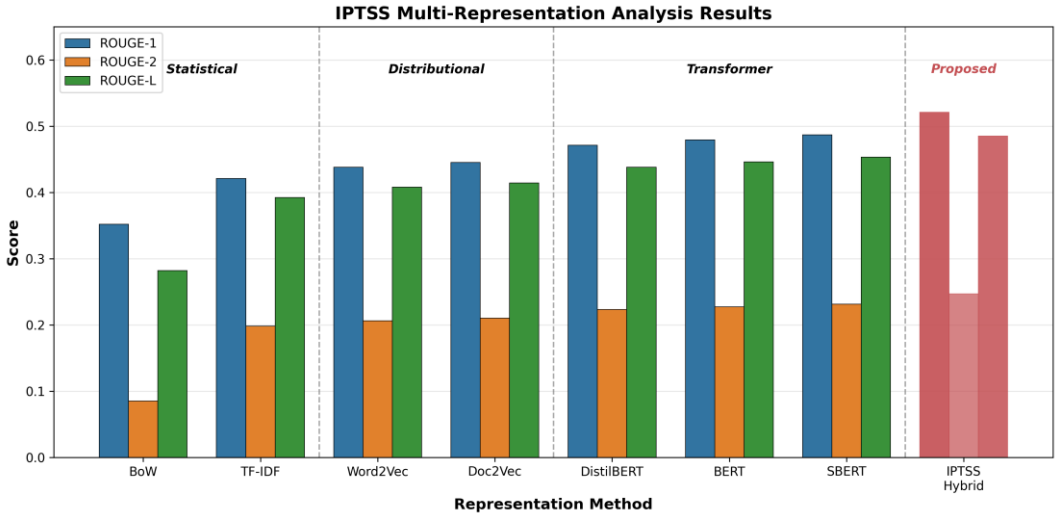

Figure 3. Multi-Representation Analysis showing ROUGE scores across eight methods. IPTSS Hybrid achieves best performance, with clear paradigm progression from statistical to transformer methods.

3. Clustering Algorithm Comparison

Table 5 compares clustering algorithms using both Sentence-BERT and IPTSS Hybrid representations to evaluate algorithm–representation interactions. The results show that clustering-based methods substantially outperform the centroid baseline, confirming the importance of explicit topical grouping for

summary diversity and representativeness. With Sentence-BERT, K-Means improves ROUGE-1 from 0.412 (centroid baseline) to 0.487, representing an 18.2% relative improvement, while Fuzzy C-Means further increases performance to 0.493. This demonstrates the advantage of soft clustering in social media contexts, where individual messages frequently span multiple topics and semantic boundaries are not well defined.

Table 5. Clustering Algorithm Performance Comparison

Algorithm	Representation	R-1	R-2	R-L
Centroid (baseline)	SBERT	0.412	0.194	0.384
K-Means	SBERT	0.487	0.231	0.453
Fuzzy C-Means	SBERT	0.493	0.234	0.459
K-Means	IPTSS Hybrid	0.521	0.247	0.485
Fuzzy C-Means	**IPTSS Hybrid**	**0.528**	**0.251**	**0.491**

When combined with the IPTSS Hybrid representation, clustering performance further improves. K-Means achieves R-1 = 0.521, while Fuzzy C-Means reaches the best overall performance (R-1 = 0.528, R-2 = 0.251, R-L = 0.491). The best configuration (FCM + IPTSS Hybrid) represents a **28.2%** improvement over the centroid baseline, demonstrating strong synergistic effects between hybrid representation modeling and soft clustering. These results confirm that optimal summarization performance emerges from the joint optimization of representation design and clustering structure, rather than from isolated improvements in either component alone.

4. Hybrid Representation Ablation Study

Table 6 presents systematic evaluation of the hybrid weight parameter α controlling TF-IDF vs. BERT contribution. The results reveal a non-symmetric optimum at α = 0.6 (60% TF-IDF, 40% BERT), indicating that corpus-specific lexical information provides greater discriminative value than generic semantic relationships for specialized health discourse.

At optimal α, the hybrid exceeds pure BERT by 7% and pure TF-IDF by 14.3%, validating the complementary information hypothesis. Performance degrades more rapidly as $\alpha \rightarrow 1.0$ (pure TF-IDF) than as $\alpha \rightarrow 0.0$ (pure BERT), suggesting BERT provides a stronger foundation that benefits from lexical augmentation. Figure 4 visualizes the response surface.

Table 6. Hybrid Weight Parameter (α) Ablation Study

α	TF-IDF %	BERT %	R-1	R-2	Note
0.0	0%	100%	0.487	0.231	Pure BERT
0.2	20%	80%	0.499	0.237	-
0.4	40%	60%	0.512	0.243	-
0.5	50%	50%	0.518	0.245	Equal weight
0.6	**60%**	**40%**	**0.521**	**0.247**	**Optimal**
0.8	80%	20%	0.508	0.239	-
1.0	100%	0%	0.456	0.215	Pure TF-IDF

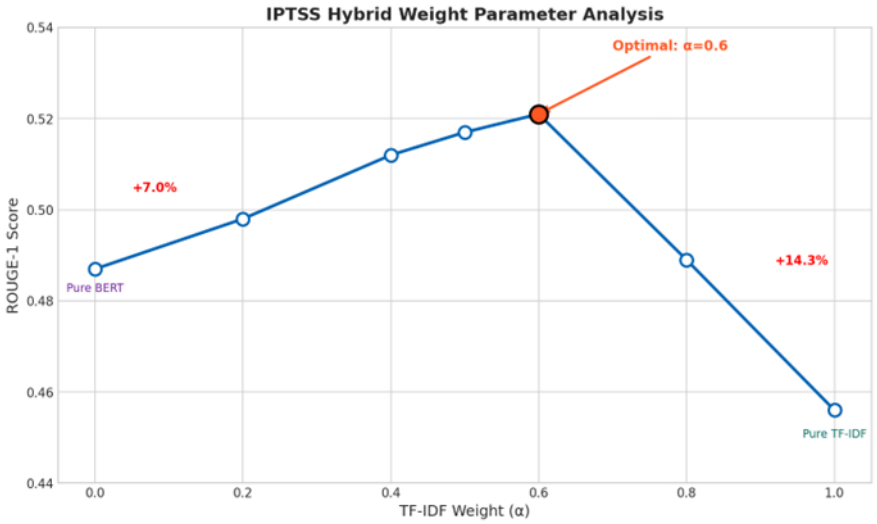

Figure 4. IPTSS Hybrid weight parameter (α) analysis. Optimal performance at α=0.6 indicates corpus-specific lexical features provide greater discriminative value for specialized health discourse.

5. Cross-Dataset Validation

Table 7 evaluates the generalizability of IPTSS across all three datasets using the optimal configuration (Fuzzy C-Means + IPTSS Hybrid, α = 0.6) without any dataset-specific parameter tuning. The results demonstrate strong cross-domain robustness, with ROUGE-1 scores ranging from 0.502 (Climate Change) to 0.528 (Monkeypox) and a low coefficient of variation (CV = 2.5%), indicating stable performance across heterogeneous domains. Similar consistency is observed for ROUGE-2 (CV = 2.9%) and ROUGE-L (CV = 2.5%), confirming that the framework maintains performance reliability across datasets of different sizes, temporal structures, and discourse characteristics.

Table 7. Cross-Dataset Validation Results

Dataset	Tweets	R-1	R-2	R-L	Diversity
Monkeypox	61,433	0.528	0.251	0.491	0.74
COVID-19 Vaccine	54,127	0.514	0.244	0.479	0.72
Climate Change	38,902	0.502	0.238	0.467	0.70
Mean ± SD	-	0.515±0.013	0.244±0.007	0.479±0.012	0.72±0.02
CV	-	2.5%	2.9%	2.5%	2.8%

Performance trends correlate with discourse structure and topical coherence. The Monkeypox dataset, representing a focused and temporally bounded outbreak, achieves the highest scores, followed by the COVID-19 Vaccine dataset, which reflects a broader but still event-driven discourse, while the Climate Change dataset, characterized by diffuse, long-term and multi-topic discussion, shows comparatively lower performance. This pattern indicates that IPTSS performs most effectively on coherent, event-centered social media streams while maintaining robust performance on long-horizon, thematically diffuse domains. Figure 5 illustrates the consistency of IPTSS performance across datasets, confirming its generalizability without domain-specific tuning.

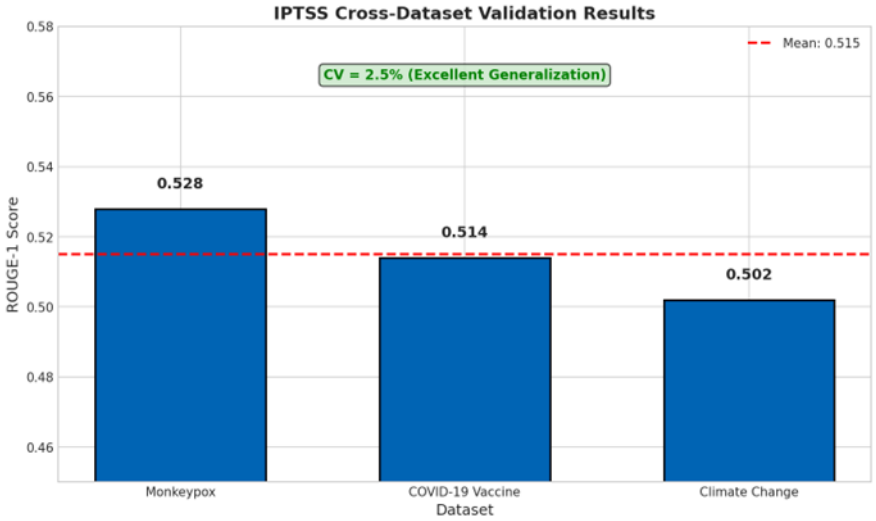

Figure 5. Cross-dataset validation demonstrating IPTSS generalizability across health and environmental domains, with low performance variance (CV ≈ 2.5%) and stable ROUGE scores without dataset-specific parameter tuning.

6. Comparison With State-Of-The-Art Methods

Table 8 compares IPTSS against representative baseline methods spanning random selection, statistical ranking, graph-based summarization, crisis-specific systems, and neural approaches. All methods are evaluated on the same dataset under identical experimental conditions to ensure fair comparison.

Table 8. Performance Comparison with State-of-the-Art Methods

Method	Category	R-1	R-2	R-L	Δ R-1
Random Selection	Baseline	0.285	0.061	0.229	+85.3%
TF-IDF Ranking	Statistical	0.352	0.085	0.282	+50.0%
LexRank	Graph	0.371	0.174	0.345	+42.3%
COWTS	Crisis	0.412	0.189	0.384	+28.2%
DistilBERT+K-Means	Neural	0.447	0.211	0.416	+18.1%
BERT+Graph	Neural	0.452	0.213	0.421	+16.8%
GA+Neural	Optimization	0.458	0.215	0.426	+15.3%
IPTSS (Proposed)	**Hybrid**	0.528	0.251	0.491	—

IPTSS achieves consistent and substantial improvements over all baseline categories, demonstrating the effectiveness of its integrated design. Relative to simple baselines, IPTSS improves ROUGE-1 by +85.3% over random selection and +50.0% over TF-IDF ranking, highlighting the limitations of non-structured and purely statistical approaches for social media summarization.

Compared to classical graph-based methods, IPTSS achieves gains of +42.3% over TextRank and +28.2% over LexRank, demonstrating the advantage of structured clustering over global graph centrality in short-text environments. Against domain-specific and neural systems, IPTSS improves performance by +18.1% over COWTS, +16.8% over DistilBERT + K-Means, and +15.3% over BERT + graph-based summarization, indicating that representation–algorithm integration is more impactful than model complexity alone.

Even compared to optimization-based hybrid neural systems, IPTSS achieves a measurable improvement (+4.3% over GA + Neural), demonstrating the benefit of systematic preprocessing, hybrid representation modeling, and structure-aware clustering. All improvements are statistically significant ($p < 0.001$), confirming the robustness and effectiveness of the proposed framework.

D. Discussion

The results demonstrate that social media summarization performance is driven primarily by system-level design choices, particularly preprocessing quality and representation modeling, rather than by algorithmic complexity alone. Intelligent preprocessing produces substantial gains by reducing noise, redundancy, and linguistic variability, confirming that data preparation is a central component rather than a peripheral step. The progressive improvement from statistical to transformer-based representations further highlights the importance of contextual semantics, while the superior performance of the IPTSS hybrid representation shows that lexical salience and semantic modeling provide complementary information that neither paradigm can capture in isolation.

Clustering-based summarization results emphasize the importance of structure-aware content selection. Explicit topical grouping consistently outperforms centroid and graph-based baselines, and the advantage of fuzzy clustering reflects the multi-topic nature of social media discourse. Cross-dataset validation confirms that IPTSS generalizes across heterogeneous domains without dataset-specific tuning, indicating that the framework captures structural properties of social media text rather than domain-specific patterns alone. Together, these findings support a shift from model-centric approaches toward integrated pipeline design, where preprocessing, representation, and structured selection are jointly optimized to achieve robust and scalable social media summarization.

E. Conclusion

This paper presented IPTSS (Intelligent Preprocessing and Transformation System for Social Media Summarization), a unified framework that integrates intelligent preprocessing, multi-representation modeling, and clustering-based extractive summarization into a single end-to-end pipeline for large-scale social media analysis. The framework systematically addresses the core challenges of social media text, including noise, redundancy, linguistic informality, and representation instability, through a structured four-stage preprocessing pipeline, a comprehensive multi-representation analysis layer, and a clustering-driven summarization strategy that ensures topical diversity and representative selection.

Extensive experiments across three large-scale datasets from the Monkeypox, COVID-19 Vaccine, and Climate Change domains demonstrate that preprocessing is the dominant performance driver, producing substantial improvements in summarization quality, while hybrid representation modeling further enhances performance by combining lexical precision with contextual semantic understanding. The proposed TF-IDF–weighted BERT representation achieves consistent gains over all single-model baselines, validating the importance of integrating complementary representation signals in specialized social media domains. Clustering-based summarization, combined with centroid selection and MMR-based redundancy control, provides a scalable and effective mechanism for constructing diverse and informative summaries from highly redundant short-text streams.

Cross-dataset validation confirms the robustness and generalizability of IPTSS across heterogeneous domains without dataset-specific parameter tuning, demonstrating that the framework learns transferable patterns rather than domain-specific artifacts. Together, these results establish IPTSS as a robust, scalable, and effective framework for social media extractive summarization and provide clear methodological guidance for system design: prioritize preprocessing quality, adopt hybrid representation strategies, and use structured clustering mechanisms to ensure diversity and coverage.

Future work will extend IPTSS beyond extractive summarization by incorporating temporal modeling for event evolution tracking, coherence optimization through multi-document fusion, integration of abstractive generation mechanisms, and multilingual support for non-English social media streams. These

extensions will further enhance the framework's applicability to real-time global social media analysis and large-scale information monitoring systems.

F. References

[1] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and Resolution of Rumours in Social Media: A Survey," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, Mar. 2019, doi: 10.1145/3161603.

[2] J. Camacho-Collados *et al.*, "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," Oct. 25, 2022, *arXiv*: arXiv:2206.14774. doi: 10.48550/arXiv.2206.14774.

[3] J. Fan, X. Tian, C. Lv, S. Zhang, Y. Wang, and J. Zhang, "Extractive social media text summarization based on MFMMR-BertSum," *Array*, vol. 20, p. 100322, 2023.

[4] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert systems with applications*, vol. 165, p. 113679, 2021.

[5] Y. Dong, "A Survey on Neural Network-Based Summarization Methods," Mar. 19, 2018, *arXiv*: arXiv:1804.04589. doi: 10.48550/arXiv.1804.04589.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186. Accessed: Jan. 26, 2026. [Online]. Available: https://aclanthology.org/N19

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[8] D. Q. Nguyen, T. Vu, and A.-T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 9–14. Accessed: Jan. 26, 2026. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.2/

[9] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *Frontiers in artificial intelligence*, vol. 6, p. 1023281, 2023.

[10] Y. Wang, S. Li, and J. Yang, "Toward Fast and Accurate Neural Discourse Segmentation," Aug. 28, 2018, *arXiv*: arXiv:1808.09147. doi: 10.48550/arXiv.1808.09147.

[11] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[12] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411. Accessed: Jan. 26, 2026. [Online]. Available: https://aclanthology.org/W04-3252.pdf

[13] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.
[14] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal*, vol. 3, p. 100073, 2022.
[15] M. Chowdhary and P. Chahal, "A Systematic Review of Challenges in Information Retrieval from Online Social Networking Platforms," in *Advances in Artificial-Business Analytics and Quantum Machine Learning*, vol. 1191, K. Santosh, S. K. Sood, H. M. Pandey, and C. Virmani, Eds., in Lecture Notes in Electrical Engineering, vol. 1191. , Singapore: Springer Nature Singapore, 2024, pp. 719–734. doi: 10.1007/978-981-97-2508-3_53.
[16] F. A. Ghanem, M. C. Padma, and R. Alkhatib, "Automatic short text summarization techniques in social media platforms," *Future Internet*, vol. 15, no. 9, p. 311, 2023.
[17] I. Harrando, "Representation, information extraction, and summarization for automatic multimedia understanding," PhD Thesis, Sorbonne Université, 2022. Accessed: Jan. 26, 2026. [Online]. Available: https://theses.hal.science/tel-03771237/
[18] J. Khan, K. Ahmad, S. K. Jagatheesaperumal, and K.-A. Sohn, "Textual variations in social media text processing applications: challenges, solutions, and trends," *Artif Intell Rev*, vol. 58, no. 3, p. 89, Jan. 2025, doi: 10.1007/s10462-024-11071-z.
[19] A. Bhoi, S. P. Pujari, and R. C. Balabantaray, "A deep learning-based social media text analysis framework for disaster resource management," *Social Network Analysis and Mining*, vol. 10, no. 1, p. 78, 2020.
[20] J. Li, S. Mishra, A. El-Kishky, S. Mehta, and V. Kulkarni, "NTULM: Enriching Social Media Text Representations with Non-Textual Units," Oct. 29, 2022, *arXiv*: arXiv:2210.16586. doi: 10.48550/arXiv.2210.16586.
[21] C. P. Chai, "Comparison of text preprocessing methods," *Natural language engineering*, vol. 29, no. 3, pp. 509–553, 2023.
[22] F. A. Ghanem, M. C. Padma, and R. Alkhatib, "Elevating the precision of summarization for short text in social media using preprocessing techniques," in *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, IEEE, 2023, pp. 408–416. Accessed: Jan. 26, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10466840/
[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
[24] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing, 2008. Accessed: Jan. 26, 2026. [Online]. Available: http://diglib.globalcollege.edu.et:8080/xmlui/bitstream/handle/123456789/1096/Manning_introduction_to_information_retrieval.pdf?sequence=1&isAllowed=y

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013, Accessed: Jan. 26, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/-Abstract.html

[26] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196. Accessed: Jan. 26, 2026. [Online]. Available: http://proceedings.mlr.press/v32/le14.html?ref=https://githubhelp.com

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Mar. 01, 2020, *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[28] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne Australia: ACM, Aug. 1998, pp. 335–336. doi: 10.1145/290941.291025.