Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(1), 239-258
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Generating Abstractive Summaries from Social Media Discussions Using Transformers

Afrodite Papagiannopoulou[1], Chrissanthi Angeli[1], Mazida Ahmad[2]

[1]Department of Electrical and Electronics Engineering, University of West Attica, Athens, Greece
[2]Department of Computing, University Utara Malaysia, Sintok, Malaysia
Email: apapagiannop@uniwa.gr, angeli@uniwa.gr, c_angeli@otenet.gr, mazida@uum.edu.my

How to cite this paper: Papagiannopoulou, A., Angeli, C. and Ahmad, M. (2025) Generating Abstractive Summaries from Social Media Discussions Using Transformers. *Open Journal of Applied Sciences*, **15**, 239-258.
https://doi.org/10.4236/ojapps.2025.151016

Received: December 19, 2024
Accepted: January 23, 2024
Published: January 26, 2024

Abstract

The rise of social media platforms has revolutionized communication, enabling the exchange of vast amounts of data through text, audio, images, and videos. These platforms have become critical for sharing opinions and insights, influencing daily habits, and driving business, political, and economic decisions. Text posts are particularly significant, and natural language processing (NLP) has emerged as a powerful tool for analyzing such data. While traditional NLP methods have been effective for structured media, social media content poses unique challenges due to its informal and diverse nature. This has spurred the development of new techniques tailored for processing and extracting insights from unstructured user-generated text. One key application of NLP is the summarization of user comments to manage overwhelming content volumes. Abstractive summarization has proven highly effective in generating concise, human-like summaries, offering clear overviews of key themes and sentiments. This enhances understanding and engagement while reducing cognitive effort for users. For businesses, summarization provides actionable insights into customer preferences and feedback, enabling faster trend analysis, improved responsiveness, and strategic adaptability. By distilling complex data into manageable insights, summarization plays a vital role in improving user experiences and empowering informed decision-making in a data-driven landscape. This paper proposes a new implementation framework by fine-tuning and parameterizing Transformer Large Language Models to manage and maintain linguistic and semantic components in abstractive summary generation. The system excels in transforming large volumes of data into meaningful summaries, as evidenced by its strong performance across metrics like fluency, consistency, readability, and semantic coherence.

Keywords

Abstractive Summarization, Transformers, Social Media Summarization,

Transformer Language Models

1. Introduction

The proliferation of social media platforms has created unprecedented opportunities for sharing and analyzing vast amounts of social data. The influence of social media is becoming increasingly significant, particularly in shaping public discourse. Users' posts and comments play a critical role in molding opinions on a wide range of important topics, including politics, economics, and societal issues. Social media data is invaluable in identifying patterns of social behavior, which can inform decisions in social, business, and governmental contexts.

People often express themselves more openly and candidly online, feeling secure in the absence of face-to-face interaction. This openness is fueled by the anonymity offered by the internet, the perceived privacy of online communication, and the lack of racial stereotypes in digital interactions. Social media posts frequently reflect real-time events. When significant incidents occur, social media platforms are inundated with content related to the unfolding event, serving as a dynamic and immediate source of information. These posts take the form of articles, discussions, and conversations, making them time-intensive and challenging to read in their entirety. Summarizing them is essential to extract valuable insights efficiently. Similarly, summarizing social media user comments is equally important, as these comments reflect public opinion on specific events. Users communicate through diverse modalities, including text, audio, images, and videos, making these platforms essential channels for exchanging news, opinions, and sentiments. Platforms such as Facebook, Instagram, Twitter (X), Reddit, LinkedIn and many others, have transformed communication dynamics, influencing daily habits and enabling the extraction of actionable insights for business, political, and economic decision-making.

Among the various modes of digital interaction, text-based posts hold a position of particular prominence. Over the past quarter-century, Natural Language Processing (NLP) has matured into a sophisticated discipline, integrating insights from computer science, artificial intelligence, and linguistics to analyze and interpret natural language effectively. While NLP has proven adept at processing traditional media, its application to social media content presents a distinct set of challenges stemming from the inherently informal, diverse, and unstructured nature of user-generated text.

Social media posts are characterized by the frequent use of slang, abbreviations, emojis, misspellings, and non-standard grammatical structures, all of which complicate accurate interpretation by conventional NLP models. The brevity of posts and comments further limits contextual clues, often obscuring the intended meaning or sentiment. Additionally, the pervasive presence of irrelevant or noisy data—such as spam, advertisements, and off-topic contributions—

further impedes meaningful analysis. The dynamic and rapidly evolving lexicon of social media, driven by trends, hashtags, and emerging colloquialisms, demands that NLP systems adapt continuously to remain effective. Moreover, capturing the subtleties of sentiment, sarcasm, and irony in social media content poses a significant challenge due to their nuanced and context-dependent nature. These complexities underscore the necessity for innovative approaches and specialized methodologies tailored to extracting and processing information from social media, ensuring that the unique attributes of this digital medium are adequately addressed.

A crucial application of NLP in this context is the summarization of user comments, which helps address the overwhelming volume of content. Abstractive summarization, a cutting-edge approach, has demonstrated its effectiveness by generating concise, human-like summaries. Unlike extractive summarization, which selects words and phrases directly from the original text without considering their deeper meaning, abstractive summarization focuses on understanding the essence of the text. It then uses new words and phrases to craft a concise version that appears entirely different from the original while preserving its core meaning. By condensing user opinions, abstractive methods provide users with a clear overview of recurring themes, prevailing sentiments, and critical insights, thereby enhancing comprehension and engagement.

For businesses, summarized feedback offers actionable insights into customer sentiment, preferences, and pain points. It facilitates quicker response times, enabling decision-makers to detect trends, prioritize issues, and adapt strategies effectively. Furthermore, summaries reduce cognitive load, helping users and organizations navigate high volumes of data efficiently. Summarization thus emerges as a pivotal tool, not only for improving user experience on social media platforms but also for enabling informed, agile decision-making in an increasingly data-driven world.

Summarizing involves distilling the essence of an extended text into a more concise form while retaining its original meaning. This process can be executed through various methods. One approach, known as *manual summarization*, entails human effort and is often time-intensive. Alternatively, advancements in algorithms and artificial intelligence have enabled the development of *Automatic Text Summarization*, a technique that has garnered significant attention from researchers, particularly in recent years [1].

The inception of automated text summarization can be traced back to 1958 [2], marking the beginning of a field that has since evolved to encompass a wide array of applications. Early research primarily targeted the summarization of formal texts, including books, journals, scientific articles, and technical reports. These endeavors employed linguistic, heuristic, and statistical methodologies to distill essential content into concise summaries. In contemporary society, the internet has irreversibly supplanted traditional avenues of information dissemination, particularly in domains such as news, political and economic affairs, advertising, and

the exchange of opinions. This paradigm shift has redirected scholarly attention toward the development of summarization systems tailored to web content, microblogs, and social media networks. A multitude of algorithms have been devised to generate text summaries. However, in the field of social media summarization, it is essential to note that transformer-based abstractive summarization systems have been underexplored.

The aim of this work is to create an automated system capable of summarizing user feedback from social media, extracting key themes, and identifying pressing issues, thereby enabling businesses and individuals to make timely and well-informed decisions. This study pushes the boundaries of current methodologies in three key areas. First, it generates summaries that are not only coherent but also meaningful, tailored to the distinctive features of social media interactions, such as (a) the linguistic intricacies of the content, (b) the clustering of related ideas, and (c) the evolving nature of discussions. Second, unlike previous research that primarily centers on Reddit and Twitter, this work broadens the scope by exploring the creation of varied datasets and training strategies. Finally, it moves beyond traditional evaluation metrics by employing a multidimensional framework to assess the quality of the generated summaries.

The remainder of this paper is structured as follows: Section 2 provides an overview of the research context. Section 3 details the methodology, outlining the proposed system and its underlying architecture. Section 4 illustrates the experimental settings and model's results. Section 5 explores the evaluation perspective for summary generation, broadening the scope of analysis. Finally, Section 6 concludes the study, summarizing the key findings and discussing potential avenues for future enhancements.

2. Related Work

The advancement of Artificial Intelligence (AI) and Neural Networks has significantly propelled the progress of Natural Language Processing (NLP). This evolution is primarily attributed to two key factors: (a) the capability to process the vast volumes of information available on the internet, and (b) the remarkable computational power accessible today.

In the domain of text summarization, substantial research has focused on leveraging Recurrent Neural Networks (RNNs) and their variant Long Short-Term Memory (LSTM) RNNs, along with Convolutional Neural Networks (CNNs) [3]. For abstract summarization of social media content, methodologies utilizing RNNs have employed frameworks such as the Attentional Encoder-Decoder model [4], or mechanisms of attentional awareness [5]. These approaches effectively filter pertinent information while addressing the unique characteristics of social media data.

An innovative technique described in [6] combines sequence-to-sequence modeling with attention mechanisms. In this model, the encoder is fortified with an LSTM architecture, while the decoder integrates attention layers. This

enhancement allows more direct access to the input sequence, facilitating the generation of concise and relevant summaries.

The advent of transformers and their attention mechanism [7] has precipitated a transformative shift in deep learning applications, particularly resonating within natural language processing (NLP) research [8]. Initially, transformer models were employed for tasks such as summarizing texts, including articles, books, and official documents. For instance, a two-stage, transformer-based method was introduced [9] to generate abstractive summaries of Chinese articles. This approach adeptly produces coherent and variable-length abstracts tailored to user requirements. The process begins with a pre-trained BERT model and a bidirectional LSTM to segment the input text. Subsequently, a mining-based BERTSUM model extracts the most salient information from these segments. The model undergoes a two-stage training regimen, ultimately utilizing a Document Transformer. The input to the Document Transformer comprises the outputs from the export model, while its output generates concise and coherent header summaries.

Another noteworthy transformer-based approach [10] processes conversational dialogues to generate abstractive summaries of encounters, such as human interactions. Additionally, comparative analyses have been conducted [11] among three pre-trained transformer models—BART, PEGASUS, and T5—using news articles sourced from the web as input. These models, fine-tuned for summarization tasks, demonstrated impressive performance, producing fluent summaries. Evaluation metrics, such as ROUGE, revealed that the T5 model outperformed its counterparts.

Transformers have also been employed to synthesize summaries from diverse datasets, including the Wikihow knowledge base [12] and web documents enriched with insights from social media [13]. Pre-trained models like BERT and T5 are instrumental in this context, facilitating the generation of high-quality summaries that reflect the evolving capabilities of transformer-based architectures.

In the domain of social media research, there is a growing focus on leveraging transformer-based models to generate summaries of online content. Efforts to summarize user comments beneath individual posts, however, remain in their infancy. The unique characteristics of social media content—encompassing both posts and user interactions—pose challenges that transformer models are uniquely equipped to address.

Current studies predominantly draw from platforms such as Twitter and Reddit [11] [14] [15], alongside Chinese social media platforms like Sina Weibo [16]. These summaries tend to center on events and the content of posts, with comparatively less attention paid to user comments. Typically, these approaches employ pre-trained BERT models on the encoder side, paired with either non-pre-trained transformers [14] [15] or alternative technologies [16] on the decoder side. Although the pre-trained T5 model demonstrates versatility across various natural language processing tasks, its application to summarizing user comments remains relatively limited [11]. During the same period, an advanced abstract summarization model was introduced [17], employing a transformer-based architecture to

generate concise summaries of individual sentences from review texts. This approach was further enhanced by integrating the Universal Sentence Encoder with statistical techniques and a graph reduction algorithm, enabling the selection of the most pertinent sentences to effectively encapsulate the essence of the entire text within the summary.

Prompt engineering has recently emerged as a groundbreaking and effective approach to enhancing the performance and adaptability of language models. Rooted in the capabilities of large language models (LLMs) to discern linguistic patterns and structures, this technique addresses the challenges posed by the substantial time and computational demands required for model training. In response, researchers have pivoted toward prompt engineering as a practical alternative.

Innovative frameworks, such as OpenPrompt [18], have been developed to facilitate prompt learning within pre-trained language models (PLMs). OpenPrompt accommodates a range of tasks—sorting and generating—while supporting various PLM architectures, including masked language models (MLM), language models (LM), and sequence-to-sequence models (Seq2Seq), alongside modular prompt components. Additionally, advanced mechanisms for summary generation have been introduced, leveraging an entity chain intermediate representation [19]. In this approach, summaries are derived as a function of the input and the constructed entity chain. To address the challenges posed by the increasing scale of contemporary language models, prefix-tuning has emerged as an innovative solution [20], enabling efficient adaptation without necessitating extensive modifications to the underlying model. In contrast to fine-tuning, which adjusts all model parameters and requires storing a complete model copy for each task, prefix-tuning preserves the language model's parameters, modifying only a small, task-specific continuous vector known as the "prefix." This approach facilitates personalization by enabling the creation of distinct prefixes tailored to individual users' data and requirements, thereby supporting the generation of specialized text. A notable variation of this concept is "prompt-tuning" [21], as opposed to "prefix-tuning." Given the substantial cost of sharing and maintaining large models, reusing a frozen model across multiple downstream tasks offers a practical solution to reduce computational and storage overhead.

Prompt engineering has emerged as a highly effective approach in the healthcare domain, addressing limitations in the performance of traditional machine learning and deep learning methods in tackling NLP tasks within the medical field [22]. As an innovative and promising paradigm in natural language processing, prompt-based learning seeks to bridge the shortcomings of conventional techniques while accommodating the intricate demands of natural language understanding and processing [23].

3. Methodology

Social media comments and opinions significantly influence economic, political,

and business decisions, but their vast volume makes them difficult to manage. Abstractive summarization has proven effective in generating concise, human-like summaries that highlight key themes, perspectives, and insights, enhancing comprehension and reducing cognitive load. For businesses, summarizing user feedback provides actionable insights, streamlines analysis, and supports agile decision-making by identifying trends, assessing sentiment, and prioritizing critical issues.

Transformers excel at summarizing informal text, such as conversational messages or social media posts, due to their ability to understand context, slang, and fragmented language through self-attention mechanisms. Trained on diverse datasets, they adapt to informal styles while preserving tone and intent in summaries. Their parallel processing enhances efficiency for large-scale tasks, and fine-tuning on domain-specific data makes them robust against noise, ensuring meaningful content extraction.

Transformer-based systems provide an effective solution for generating human-like summaries, leveraging their exceptional ability to process informal input through advanced contextual understanding, adaptability, and robustness. These systems excel at converting disorganized or fragmented informal text into coherent summaries, maintaining the original tone, sentiment, and essential details. Furthermore, they prove invaluable for contemporary applications such as conversational AI, social media analytics, and beyond.

The objective of the proposed work is to develop an automated system that can summarize user feedback from social media platforms, identify key themes, and highlight urgent issues to assist businesses and individuals in making timely and informed decisions [24].

3.1. T5-Architecture for Summarization

The **T5 (Text-to-Text Transfer Transformer)** [25] model follows the traditional Transformer architecture. It is an encoder-decoder model with multi-head attention. The differences are (**Figure 1**):

1) *Placement of Layer Normalization*: Layer normalization is applied directly preceding each attention mechanism and feedforward transformation, positioned outside the residual pathway.

2) *Exclusion of Additive Bias in Layer Normalization*: The implementation of LayerNorm utilizes only scaling, omitting the additive bias component.

3) *Position Embedding Strategy*: A straightforward positional embedding technique is employed, wherein a scalar value is added to the corresponding logit utilized in the computation of attention weights.

4) *Application of Dropout*: Dropout regularization is incorporated throughout the network, including attention weights, feedforward layers, and skip connections, among other components.

It works well for summarization tasks due to its sequence-to-sequence architecture, pre-training on the span-corruption objective, and flexible task setup. It

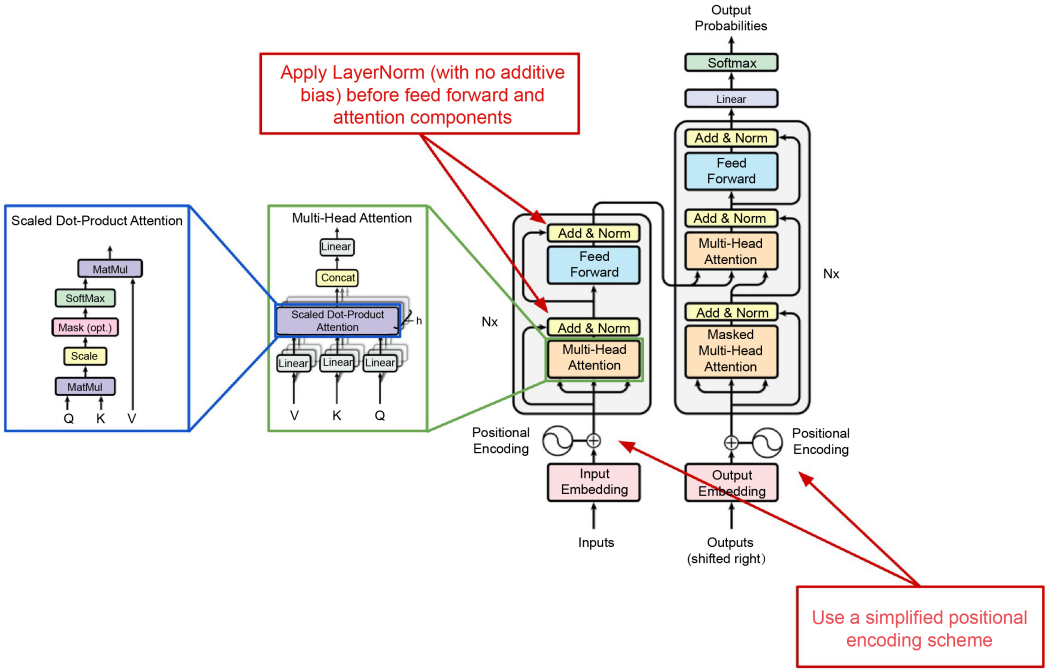

Figure 1. Differences of T5 and the traditional transformer architecture[1].

is widely used for abstractive summarization due to its powerful architecture and its ability to generate coherent, concise summaries. The key points of T5 are the following:

Unified Text-to-Text Framework for Summarization: T5 frames all tasks, including summarization, as text-to-text problems, unifying input and output as text sequences. This consistent format simplifies task setup and enables seamless transfer across NLP tasks like translation, question answering, and summarization.

Encoder-Decoder Architecture: T5's encoder-decoder architecture is tailored for sequence-to-sequence tasks like summarization. The encoder captures the input text's context and meaning, while the decoder generates a coherent summary conditioned on the encoder's output and previously generated words.

Pre-training on Span-Corruption Task: T5 is pre-trained using a span-corruption objective, where masked text spans are predicted based on context. This teaches T5 to generate and rephrase content, making it highly effective for abstractive summarization and creating concise summaries.

Task-Specific Prefixes and Training Data for Summarization: T5 uses the prompt "summarize:" to indicate summarization tasks, training on datasets like CNN/Daily Mail to learn effective summary structures and styles. Fine-tuning enhances its ability to generate high-quality summaries tailored to the domain.

Autoregressive Text Generation for Coherent Summaries: The T5 decoder generates summaries autoregressively, producing one word at a time by using previously generated words as context. Masked self-attention ensures each token

[1]https://cameronrwolfe.substack.com/p/t5-text-to-text-transformers-part.

depends only on prior tokens, enabling sequential generation. Techniques such as beam search or top-k sampling enhance summary quality by exploring multiple candidate sequences and selecting the most probable one based on model scores.

Controlling Output Length and Content in Summaries: T5 can be fine-tuned for summaries of varying lengths based on the use case, such as concise news summaries or detailed medical/legal ones. Tailoring datasets and applying length-control techniques like setting maximum length or nucleus sampling during decoding enable T5 to meet specific length requirements.

Handling Different Summarization Requirements: T5 can be fine-tuned for both short and multi-sentence summaries, including news or research articles. By training on domain-specific datasets (e.g., medical or legal), T5 can generate summaries tailored to the most relevant information in those fields.

3.2. Model Structure

The unique characteristics of social media content, previously highlighted, render the task of summary generation particularly compelling. Unlike formal texts, articles, or documents enriched with structured grammar and linguistic precision, social media posts are typically brief, replete with abbreviations, slang, special characters, and emoticons. Moreover, their often redundant and repetitive nature can confuse readers.

This study centers on crafting human-like summaries of social media discussions, aiming to distill and streamline the information conveyed to the public while preserving the original meaning and core ideas [26]. To accomplish this, it leverages the T5 transformer model, an encoder-decoder architecture renowned for its efficacy in abstractive summarization. As discussed earlier, T5 reframes all problems into a text-to-text format, making it highly adaptable for supervised or unsupervised tasks. Its versatility extends to a wide range of natural language processing tasks, including translation, language inference, information extraction, and summarization. The model's training utilizes teacher forcing, processing input sequences alongside corresponding target sequences.

The choice to adopt a pre-trained model, rather than constructing one from scratch, was guided by several advantages: (a) pre-trained models deliver superior results with thoughtful data preprocessing; (b) they offer a robust foundation for fine-tuning across diverse datasets; and (c) they facilitate the creation of specialized models with minimal adjustments to training and fine-tuning, enabling faster and more efficient outcomes.

The project's core design involves the following steps (**Figure 2**):

Data Collection: Data is downloaded from social media platforms and modified to suit the model's training needs. The dataset must align with the specific type of summarization, such as news or social media summarization, and contain high-quality, coherent summaries. A suitable dataset also matches the required summary length for the task, ensuring the model generates relevant outputs. Poor-quality or too-small datasets can lead to poor model performance and overfitting.

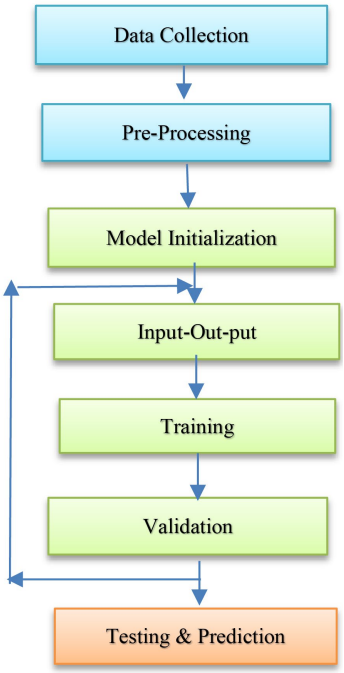

Figure 2. Model Structure.

Pre-processing: Pre-processing is a crucial step in preparing raw text data for analysis. It transforms messy, inconsistent text into a clean, standardized format, enhancing the performance of NLP models by improving accuracy and efficiency.

Model Initialization: The model is a transformer-based encoder-decoder designed for abstractive text summarization, using a pre-trained T5 model for conditional generation. The T5 tokenizer converts text into token IDs, tokenizes the dataset, adds attention masks, and ensures consistent input lengths for the model.

Input-Output: The process involves encoding input text to create contextual embeddings that capture semantic relationships. These embeddings are passed to the decoder, where cross-attention layers focus on relevant input parts while generating the summary. This setup fine-tunes a T5 transformer for text summarization, handling data tokenization, preparation, optimization, and training with validation. Model checkpoints are saved at each epoch if performance improves.

Training and Optimization: This training process completes one epoch of T5 model training on a batched dataset, incorporating mixed-precision training for faster computation. It handles input tokens, attention masks, summary tokens, and their respective masks. The forward pass computes loss and predictions, followed by gradient scaling for mixed-precision. The learning rate is adjusted using a scheduler, and gradients are cleared before each pass. The model is optimized with Adam and a decreasing learning rate schedule.

Validation: The validation process evaluates a trained model on a validation dataset after each epoch, calculating average validation loss and perplexity (PPL). It prepares the model, avoids gradient calculation to save memory, and computes the loss and prediction scores for each batch. The average loss and PPL are then

calculated to assess model performance, helping track learning progress and avoid overfitting.

Testing and Prediction: The testing process evaluates a trained model on a test dataset by calculating average test loss and perplexity, and generating predictions. The model is switched to evaluation mode, and variables for test loss and predictions are initialized. Each batch is processed on the GPU with gradient calculations disabled to save memory and improve speed. Predictions are made and the loss is accumulated across batches to compute the average test loss.

4. Experimental Settings and Results

4.1. Data Preprocessing

There are numerous platforms that provide datasets for natural language processing, primarily suited for tasks such as data mining, event detection, sentiment analysis, and emotion analysis. However, datasets specifically tailored for summarization in social media are not readily accessible. The process of locating and downloading the right datasets has been challenging due to restrictions on data downloads imposed by social media platforms. Nonetheless, selecting an appropriate dataset is crucial for summarization tasks, as its quality, relevance, and structure significantly impact the model's ability to generate effective summaries. This project uses a dataset of Facebook news posts accompanied by user comments below each post. The data was processed to eliminate irrelevant elements while preserving and organizing the useful ones. The raw data has 7 columns namely "created_time", "from_id", "from_name", "message", "post_name", "post_title", "post_num" and 1,781,576 rows. The adjustments made were to retain 3 of the 7 columns [27]. Therefore, the inclusion of 'post_title' and 'post_number', which serve as identifiers for specific posts and facilitate the grouping of posts and events, was considered essential. Additionally, the 'message' column was utilized as the primary source of information, as it contains the user comments that need to be summarized.

4.2. T5 Fine-Tuning Configuration and Performance Metrics

The proposed model was trained on PyCharm with NVDIA GeForce 4070 GPU with 12 BG RAM. T5 pre-trained model fine-tuned for the purposes of the study and its variants are tested according to the computational power available. The summary length is calculated according to the input list length. The dataset is split into Train, Validation and Test with 80%, 20% and 10% ratio size respectively. The selected parameters were chosen based on the available computational power and resources. The system is trained using the standard T5-small model with 6 layers, 8 attention heads, and a feedforward network depth of 2048.

DROPOUT RATE = 0.09.

LEARNING RATE = $1^{e}-3$.

EPOCHS = 11.

BATCH SIZE = 16.

The Train, Validation loss and Validation PPL were calculated to assess the performance of the model (Table 1). Loss is a crucial metric that quantifies the error produced by a model, playing a fundamental role in deep learning and neural network training. After splitting the dataset into Training, Validation, and Test subsets, we assessed both Training and Validation losses for all samples. These metrics consistently decrease over multiple epochs, with the gap between validation and training loss gradually narrowing. This occurs because, as the network learns from the data, the regularization loss tied to model weights reduces. As a result, the difference between the two losses becomes minimal, suggesting a better fit for most data samples. Additionally, conceptual scoring methods are used to assess abstractive summaries, further strengthening the measurement process.

Table 1. Train, Validation Loss and Validation Perplexity (PPL).

Epoch	Train Loss	Validation Loss	Validation PPL
1	2.336	0.985	2.679
2	1.054	0.916	2.499
3	0.988	0.883	2.417
4	0.944	0.864	2.373
5	0.911	0.851	2.343
6	0.883	0.843	2.324
7	0.862	0.840	2.316
8	0.846	0.835	2.305
9	0.830	0.834	2.302
10	0.821	0.833	2.301
11	0.814	0.832	2.298

5. Evaluation

5.1. Evaluation Metrics

Automatic evaluation metrics, widely recognized as the most traditional methods, are commonly used in automatic summarization tasks. These metrics follow established guidelines to assess the quality of generated summaries. They are particularly popular during evaluation processes, as they are easier to implement and often serve as indicators of the underlying performance of language models [28]. Numerous automated language model evaluation metrics have been developed, initially for machine translation assessments, and are now also applied to other linguistic tasks like summarization. A key metric in this field is ROUGE [29], which evaluates the similarity between generated and reference texts. BLEU [30] extends this by adding a brevity penalty to measure the match between machine-translated text and human reference translations. To address issues in BLEU, NIST [31] and SacreBLEU [32] were introduced. However, these metrics are primarily suitable for short, single-document summaries and fail to analyze syntactic

or contextual details. As a result, alternative evaluation frameworks have emerged, offering greater flexibility. CIDEr [33] compares the similarity of a generated sentence with a set of human-written reference sentences. METEOR [34] improves upon BLEU by extending monogram matching. chrF [35] and chrF++ [36] utilize character-level n-grams and focus on morpho-syntactic effects. Additionally, integrated metrics have been introduced, where word symbols or n-grams represent real-valued vectors that can be learned using various neural network models. These metrics evaluate summaries by comparing the similarity of representations between the generated and reference texts. Word Mover's Distance [37] measures the "distance" between two texts based on the minimum amount of movement required for words to match between the documents using word embeddings. BERTScore [38] calculates similarity between the generated and reference texts using embeddings from the BERT language model, specifically by computing the cosine similarity between contextualized token embeddings. FACTCC [39] assesses factual consistency in summaries or generated texts, particularly in tasks like abstractive summarization, to identify errors that deviate from the source material. The BLANC [40] metric, designed for summary evaluation, does not require a reference text. Instead, it measures the summary's quality by evaluating how well it aids a machine learning model in tasks like filling in missing parts of the original text, quantifying informativeness and coherence. SUPERT [41], developed for summarization tasks, uses unsupervised techniques to evaluate summary quality without relying on reference summaries, making it valuable when high-quality references are unavailable. As an extension of BLANC, the Shannon evaluation metric [42] further examines consistency in generated documents. Inspired by Claude Shannon's work in information theory, this metric measures the information content, efficiency, and uncertainty of a document. Finally, SDC (Semantic Distance-based Clustering [43], evaluates summary quality by measuring the semantic similarity between summary sentences and key sentences in the source document, without relying on reference summaries.

5.2. Evaluation Results and Analysis

The advent of neural networks, the introduction of transformers, and the development of large language models have made Natural Language Generation (NLG) a highly complex field in terms of evaluating generated texts. If human evaluation were feasible, it would primarily rely on linguistic criteria, making it multi-dimensional. However, due to the vast amount of information available on the internet, human evaluation has been supplanted by automated evaluation systems. While efforts are being made to create multi-dimensional automatic evaluators, similarity-based metrics continue to be the dominant approach in NLG evaluation today.

This paper represents evaluation metric that emphasize the inherent characteristics of the generated text, such as fluency, coherence, and informativeness, as well as semantics, instead of directly comparing it to a reference. This section begins by introducing the metrics used to assess this research, followed by an

evaluation of the system's performance based on these metrics, and concludes with an analysis of the results (Table 2).

Table 2. Evaluation Scores for the generated summaries.

Evaluation Scores				
SUPERT	Shannon Consistency	Shannon Coherency	SDC	Rouge-WE
0.9990	0.9512	0.9519	0.5746	0.9568

SUPERT is a tool for summarization tasks that uses unsupervised methods to assess summary quality without requiring reference summaries. It evaluates how well a summary captures key content from the source document by utilizing sentence embeddings and clustering. The method groups sentences from the source based on their semantic similarity, identifying central topics. SUPERT then checks how well the summary aligns with these important clusters, assigning a higher score if the summary covers the key points effectively.

The Shannon evaluation metric, based on Claude Shannon's information theory, measures the efficiency and information content of messages, documents, or signals. Although Shannon did not create a specific metric for text evaluation, his concept of entropy is widely used to assess information richness, redundancy, and uncertainty. In text evaluation, low entropy indicates redundancy (predictability), while high entropy suggests richness (diversity and unique information).

SDC (Semantic Distance-based Clustering) is a reference-free metric for evaluating summaries by measuring how semantically similar the summary sentences are to key sentences in the original document. It works by converting both source document sentences and summary sentences into vector representations using sentence embeddings like BERT. These vectors are then clustered by semantic similarity, with each cluster representing a central theme or topic. The metric calculates the semantic distance between each summary sentence and the central sentences of the clusters, with higher scores indicating better alignment with the document's key points.

ROUGE-WE [44], or "ROUGE with Word Embedding," extends the traditional ROUGE metric by incorporating word embeddings, enabling it to measure semantic similarity rather than just exact word matches. Unlike classic ROUGE, it allows "soft" matching, recognizing synonyms and related words, making it better at capturing the meaning of a summary. By using embeddings like Word2Vec or GloVe, ROUGE-WE aligns words based on semantic proximity, evaluating overlap of semantically similar n-grams rather than exact matches.

Observing the results (Table 2) we can reach the following conclusions:

SUPERT scores typically range from 0 to 1, with 1 being the highest possible score. A score of 0.9990 indicates that the summary is almost perfectly aligned with the essential content of the source text. This suggests that the summary includes nearly all critical information clusters and represents the main ideas effectively. It is highly coherent, informative, and likely covers the important aspects

without significant omissions or errors. The summary's content closely matches the essential information clusters from the source, indicating excellent representation and coverage of the source text's main ideas.

Shannon evaluation scores are based on Shannon entropy, a concept from information theory, applied to evaluate the quality of summaries in terms of consistency and coherence. A high consistency score (0.9512 in this case) indicates that the abstract retains most of the key information from the document provenance with very few inconsistencies or inaccuracies. A score of 0.9512 indicates that the summary is 95.12% consistent with the original content. Shannon Coherency assesses how well structured and logically connected the sentences or ideas of the abstract are. It measures whether the summary flows smoothly and presents the information in a coherent manner. A high coherence score (0.9519 here) means that the summary maintains logical progression and that ideas are presented in a clear and connected manner. It also indicates that the summary is 95.19% coherent, meaning that the information is presented with great clarity and logical consistency. In conclusion the summary closely follows the information in the source document, with a very high level of factual accuracy. In addition, it is well organized and maintains a logical flow between sentences or sections, making it easy to follow and understand. Both scores approaching 1.0 reflect an excellent quality summary that is both highly consistent with the original content and presented in a coherent, structured manner.

SDC score refers to the result of an evaluation based on the Semantic Distance-based Clustering (SDC) metric for summary evaluation. It measures the semantic distance between sentences in the summary and the most important clusters (or themes) of sentences in the source document. A lower score in SDC indicates that the summary is semantically closer to the key sentences (or clusters) in the document, meaning it captures the most important information more accurately. An SDC score of 0.5746 represents the average semantic distance between the summary sentences and the central topics of the document. The average score indicates room for improvement in how well the summary captures the most important information. Such a result is absolutely expected given the specificity, difficulty and linguistic limitations of the original text.

A ROUGE-WE score of 0.9568 indicates a strong similarity between the generated summary and the reference summary when evaluated using word embeddings. Here's what this entails:

High Semantic Similarity: A score nearing 1.0 signifies that the generated summary effectively reflects the semantic meaning of the reference summary. This suggests it includes many relevant words or phrases that are semantically related, even if the exact wording varies.

Accurate Content Representation: The high score indicates that the generated summary successfully conveys the main ideas or themes present in the reference, showcasing strong performance in capturing semantic content.

In summary, metrics focusing on the meaning of the text yield better results. A

ROUGE-WE score of 0.9568 highlights the model's ability to produce content that closely aligns with the intended message, demonstrating its effectiveness in generating meaningful summaries.

6. Conclusions

The rapid development of information volume on web and social media has led to an increasingly strong demand for automatic summarization systems and evaluation metrics for them. In the case of social media, both automatic summarization and automatic evaluation are difficult processes due to the nature of the texts. These challenges have led to the need for a more comprehensive assessment approach using a combination of many old and modern tools. In this paper, Non-classical methods have been used to evaluate the automatic text generated.

Summary evaluation metrics, like those selected above, provide an essential framework for assessing the quality of generated summaries without relying on human-written references. These metrics offer insights into various aspects of summary quality, including fluency, coherence, informativeness, and topic coverage.

SDC is designed to assess the quality of summaries without relying on gold-standard reference summaries. The key idea behind SDC is to measure how semantically similar the sentences in the summary are to the most important sentences in the source document. However, like many semantic metrics, it doesn't evaluate the fluency or readability of the text, and its performance is sensitive to the quality of the embeddings and clustering algorithm used.

While Shannon entropy (derived from Claude Shannon's pioneering work) is not a full-fledged text evaluation metric by itself, it offers a valuable tool for assessing the informativeness, predictability, and compression efficiency of text. It is widely used in various applications, such as language modeling, data compression, and summarization tasks, but it should often be combined with other metrics that can account for semantics, coherence, and readability to provide a more complete evaluation of a text or summary. However, it needs to be supplemented with other evaluation metrics to capture the full quality of the text.

SUPERT is a powerful metric for evaluating summaries, especially useful when reference summaries are not available. It assesses how well a generated summary captures the important content from the original document using sentence embeddings and clustering. While it provides valuable insights into the semantic coverage of summaries, it doesn't explicitly evaluate fluency or readability, so it's often combined with other metrics in practical applications. The metrics indicate that the model performs exceptionally well in terms of consistency, coherence, and overall quality (with very high Shannon Consistency, Shannon Coherency, and SUPERT scores). However, the SDC score is lower, implying that while the model's text is coherent and consistent, it may have room for improvement in semantic diversity or could benefit from a broader range of expressions.

In general, these high scores suggest that the model's outputs are reliable,

logical, and high-quality, though it could improve in terms of diversity to make the text less repetitive or more varied in expression. Finally, ROUGE-WE is beneficial when assessing summaries or text generation tasks where semantic understanding matters, not just exact word matching. This makes it a valuable tool for evaluating tasks like machine translation, summarization, or text generation where synonyms and paraphrasing are common. This is the main reason why ROUGE-WE shows quite high results.

While the results achieved during both the training and evaluation phases are commendable, there remains scope for further refinement. Notably, the system exclusively processes comments in English, thereby disregarding any input provided in other languages.

Acknowledgements

The publication fees were totally covered by ELKE at the University of West Attica.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gupta, S. and Gupta, S.K. (2019) Abstractive Summarization: An Overview of the State of the Art. *Expert Systems with Applications*, **121**, 49-65. https://doi.org/10.1016/j.eswa.2018.12.011

[2] Luhn, H.P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, **2**, 159-165. https://doi.org/10.1147/rd.22.0159

[3] Gao, S., Chen, X., Li, P., Ren, Z., Bing, L., Zhao, D., *et al.* (2019) Abstractive Text Summarization by Incorporating Reader Comments. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 6399-6406. https://doi.org/10.1609/aaai.v33i01.33016399

[4] Liang, Z., Du, J. and Li, C. (2020) Abstractive Social Media Text Summarization Using Selective Reinforced Seq2seq Attention Model. *Neurocomputing*, **410**, 432-440. https://doi.org/10.1016/j.neucom.2020.04.137

[5] Wang, Q. and Ren, J. (2021) Summary-Aware Attention for Social Media Short Text Abstractive Summarization. *Neurocomputing*, **425**, 290-299. https://doi.org/10.1016/j.neucom.2020.04.136

[6] Bhandarkar, P. and Thomas, K.T. (2022) Text Summarization Using Combination of Sequence-to-Sequence Model with Attention Approach. In: Smys, S., Lafata, P., Palanisamy, R. and Kamel, K.A., Eds., *Computer Networks and Inventive Communication Technologies*, Springer, 283-293. https://doi.org/10.1007/978-981-19-3035-5_22

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[8] Gupta, A., Chugh, D. and Katarya, R. (2022) Automated News Summarization Using Transformers. In Aurelia, S., Hiremath, S.S., Subramanian, K. and Biswas, S.K., Eds., *Sustainable Advanced Computing*, Springer, 249-259. https://doi.org/10.1007/978-981-16-9012-9_21

[9] Su, M., Wu, C. and Cheng, H. (2020) A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 2061-2072. https://doi.org/10.1109/taslp.2020.3006731

[10] Singhal, D., Khatter, K., A, T. and R, J. (2020) Abstractive Summarization of Meeting Conversations. 2020 *IEEE International Conference for Innovation in Technology* (*INOCON*), Bangluru, 6-8 November 2020, 1-4. https://doi.org/10.1109/inocon50539.2020.9298305

[11] Blekanov, I.S., Tarasov, N. and Bodrunova, S.S. (2022) Transformer-Based Abstractive Summarization for Reddit and Twitter: Single Posts vs. Comment Pools in Three Languages. *Future Internet*, **14**, Article 69. https://doi.org/10.3390/fi14030069

[12] Pal, A., Fan, L. and Igodifo, V. (n.d.) Text Summarization Using BERT and T5. https://anjali001.github.io/Project_Report.pdf

[13] Nguyen, M., Nguyen, V., Vu, H. and Nguyen, V. (2020) Transformer-Based Summarization by Exploiting Social Information. 2020 12*th International Conference on Knowledge and Systems Engineering* (*KSE*), Can Tho, 12-14 November 2020, 25-30. https://doi.org/10.1109/kse50997.2020.9287388

[14] Li, Q. and Zhang, Q. (2020) Abstractive Event Summarization on Twitter. Companion *Proceedings of the Web Conference* 2020, Taipei, 20-24 April 2020, 22-23. https://doi.org/10.1145/3366424.3382678

[15] Kerui, Z., Haichao, H. and Yuxia, L. (2020) Automatic Text Summarization on Social Media. *Proceedings of the* 2020 4*th International Symposium on Computer Science and Intelligent Control*, Newcastle upon Tyne, 17-19 November 2020, 1-5. https://doi.org/10.1145/3440084.3441182

[16] Tampe, I., Mendoza, M. and Milios, E. (2021) Neural Abstractive Unsupervised Summarization of Online News Discussions. In: Arai, K. Ed., *Intelligent Systems and Applications*, Springer International Publishing, 822-841. https://doi.org/10.1007/978-3-030-82196-8_60

[17] Rawat, R., Rawat, P., Elahi, V. and Elahi, A. (2021) Abstractive Summarization on Dynamically Changing Text. 2021 5*th International Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 8-10 April 2021, 1158-1163. https://doi.org/10.1109/iccmc51019.2021.9418438

[18] Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., *et al*. (2022) Openprompt: An Open-Source Framework for Prompt-Learning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Dublin, 22-27 May 2022, 105-113. https://doi.org/10.18653/v1/2022.acl-demo.10

[19] Narayan, S., Zhao, Y., Maynez, J., Simões, G., Nikolaev, V. and McDonald, R. (2021) Planning with Learned Entity Prompts for Abstractive Summarization. *Transactions of the Association for Computational Linguistics*, **9**, 1475-1492. https://doi.org/10.1162/tacl_a_00438

[20] Li, X.L. and Liang, P. (2021) Prefix-Tuning: Optimizing Continuous Prompts for Generation. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), Online, 1-6 August 2021, 4582-4597. https://doi.org/10.18653/v1/2021.acl-long.353

[21] Lester, B., Al-Rfou, R. and Constant, N. (2021) The Power of Scale for Parameter-Efficient Prompt Tuning. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, 7-11 November 2021, 3045-3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

[22] Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., *et al*. (2024) Prompt Engineering for Healthcare: Methodologies and Applications.

[23] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. (2023) Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, **55**, 1-35. https://doi.org/10.1145/3560815

[24] Papagiannopoulou, A. and Angeli, C. (2024) Encoder-Decoder Transformers for Textual Summaries on Social Media Content. *Automation, Control and Intelligent Systems*, **12**, 48-59. https://doi.org/10.11648/j.acis.20241203.11

[25] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narag, S., Matena, M., Zhou, Y., Li, W., and Liu P. J. (2019) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, **21**, 1-67. https://dl.acm.org/doi/abs/10.5555/3455716.3455856

[26] Papagiannopoulou, A. and Angeli, C. (2024) Summarizing User Comments on Social Media Using Transformers. *European Conference on Social Media*, **11**, 198-205. https://doi.org/10.34190/ecsm.11.1.2046

[27] Papagiannopoulou, A. and Angeli, C. (2023) Designing a Summarization System on Social Comments Using Transformers. 2023 12*th International Conference on Computer Technologies and Development* (*TechDev*), Rome, 14-16 October 2023, 1-5. https://doi.org/10.1109/techdev61156.2023.00008

[28] Lu, L., Liu, Y., Xu, W., Li, H. and Sun, G. (2023) From Task to Evaluation: An Automatic Text Summarization Review. *Artificial Intelligence Review*, **56**, 2477-2507. https://doi.org/10.1007/s10462-023-10582-5

[29] Lin, C.Y. (2004) Rouge: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, Barcelona, 22 July 2004, 74-81.

[30] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[31] Martin, A.F. and Przybocki, M.A. (2001) The NIST Speaker Recognition Evaluations: 1996-2001. 2001: *A Speaker Odyssey—The Speaker Recognition Workshop*, Crete, 18-22 June 2001, 39-43.

[32] Post, M. (2018) A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, 31 October-1 November, 2018, 186-191. https://doi.org/10.18653/v1/w18-6319

[33] Oliveira dos Santos, G., Colombini, E.L. and Avila, S. (2021) CIDEr-R: Robust Consensus-Based Image Description Evaluation. *Proceedings of the Seventh Workshop on Noisy User-Generated Text* (*W-NUT* 2021), Online, 11 November 2021, 351-360. https://doi.org/10.18653/v1/2021.wnut-1.39

[34] Lavie, A. and Agarwal, A. (2007) Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, 23 June 2007, 228-231. https://doi.org/10.3115/1626355.1626389

[35] Popović, M. (2015) ChrF: Character N-Gram F-Score for Automatic MT Evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, 17-18 September 2015, 392-395. https://doi.org/10.18653/v1/w15-3049

[36] Popović, M. (2017) ChrF++: Words Helping Character N-Grams. Proceedings of the Second Conference on Machine Translation, Copenhagen, 7-11 September 2017, 612-618. https://doi.org/10.18653/v1/w17-4770

[37] Kusner, M., Sun, Y., Kolkin, N., *et al.* (2015) From Word Embeddings to Document Distances. *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning*, Lille, 6-11 July 2015, 957-966.

[38] Zhang, T., Kishore, V., Wu, F., *et al* (2019) BERTScore: Evaluating Text Generation with BERT. arXiv. https://doi.org/10.48550/arXiv.1904.09675

[39] Kryscinski, W., McCann, B., Xiong, C. and Socher, R. (2020) Evaluating the Factual Consistency of Abstractive Text Summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, 16-20 November 2020, 9332-9346. https://doi.org/10.18653/v1/2020.emnlp-main.750

[40] Recasens, M. and Hovy, E. (2010) BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering*, **17**, 485-510. https://doi.org/10.1017/s135132491000029x

[41] Gao, Y., Zhao, W. and Eger, S. (2020) SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1347-1354. https://doi.org/10.18653/v1/2020.acl-main.124

[42] Egan, N., Vasilyev, O. and Bohannon, J. (2022) Play the Shannon Game with Language Models: A Human-Free Approach to Summary Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 10599-10607. https://doi.org/10.1609/aaai.v36i10.21304

[43] Liu, Y., Jia, Q. and Zhu, K. (2022) 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*, Seattle, 10-15 July 2022, 2109-2115. https://doi.org/10.18653/v1/2022.naacl-main.153

[44] Ng, J. and Abrecht, V. (2015) Better Summarization Evaluation with Word Embeddings for Rouge. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1925-1930. https://doi.org/10.18653/v1/d15-1222