IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 July 2023, accepted 28 September 2023, date of publication 5 October 2023, date of current version 17 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322226

RESEARCH ARTICLE

Warm-Starting for Improving the Novelty of Abstractive Summarization

AYHAM ALOMARI[1,2], AHMAD SAMI AL-SHAMAYLEH[3], NORISMA IDRIS[4], AZNUL QALID MD SABRI[4], (Senior Member, IEEE), IZZAT ALSMADI[5], (Senior Member, IEEE), AND DANAH OMARY[6]

[1]Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman 11931, Jordan
[2]MEU Research Unit, Middle East University, Amman 11831, Jordan
[3]Department of Data Science and Artificial Intelligence, Al-Ahliyya Amman University, Amman 19328, Jordan
[4]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[5]Department of Computing and Cybersecurity, Texas A&M University-San Antonio, San Antonio, TX 78224, USA
[6]Department of Electrical Engineering, University of North Texas, Denton, TX 76210, USA

Corresponding author: Aznul Qalid Md Sabri (aznulqalid@um.edu.my)

This work was supported by the Ministry of Higher Education, Malaysia, under Grant JPT(BKPI)1000/016/018/25(58).

ABSTRACT Abstractive summarization is distinguished by using novel phrases that are not found in the source text. However, most previous research ignores this feature in favour of enhancing syntactical similarity with the reference. To improve novelty aspects, we have used multiple warm-started models with varying encoder and decoder checkpoints and vocabulary. These models are then adapted to the paraphrasing task and the sampling decoding strategy to further boost the levels of novelty and quality. In addition, to avoid relying only on the syntactical similarity assessment, two additional abstractive summarization metrics are introduced: 1) NovScore: a new novelty metric that delivers a summary novelty score; and 2) NSSF: a new comprehensive metric that ensembles Novelty, Syntactic, Semantic, and Faithfulness features into a single score to simulate human assessment in providing a reliable evaluation. Finally, we compare our models to the state-of-the-art sequence-to-sequence models using the current and the proposed metrics. As a result, warm-starting, sampling, and paraphrasing improve novelty degrees by 2%, 5%, and 14%, respectively, while maintaining comparable scores on other metrics.

INDEX TERMS Abstractive summarization, novelty, warm-started models, deep learning, metrics.

I. INTRODUCTION

Abstractive summarization is one of the two main types of text summarization. It entails understanding a lengthy article and condensing its meanings using alternative words. The other type is extractive summarization, in which the most salient sentences are extracted from the source article without being altered to form a summary. The difficulty in understanding the text and efficiently conveying its meanings renders abstractive summarization more complex and challenging [1], [2].

In theory, novelty (otherwise known as abstractiveness) is the essential aspect that distinguishes abstractive summaries since the resulting summary must be written using different phrases and expressions from those in the input text. High copying rates in the models' outputs highlight the issue of novelty, making them seem extractive rather than abstract.

Current progress in the abstractive summarization field shows impressive results with the utilization of Deep Learning (DL) and Transfer Learning (TL) approaches. Using massive datasets and high machine capabilities, most pretrained models have been trained to either understand a text or to freely/directedly generate text, utilizing the Transformer's encoder [1], decoder, or both parts. Some of the pretrained models include the Bidirectional Encoder Representations from Transformers (BERT) [2], the Generative Pre-trained Transformer (GPT-2) [3], and the Bidirectional and Auto-Regressive Transformers (Bart) [4], respectively.

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

These pretrained models can transfer their learned knowledge to other models and be finetuned to downstream tasks.

As a result, fine-tuning pretrained models to varied tasks has revolutionized the fields of Natural Language Processing (NLP) by emulating human results in text generation, drawing academics to enrich the fields, and conduct substantial research. For abstractive summarization, the results typically attain significant improvement based on syntactical and semantical similarity, but they are still poor in terms of generating novel phrases that do not appear in the source article. One possible explanation is that pretrained sequence-to-sequence models that have been trained/fine-tuned to generate abstractive summaries, such as BART and the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) [5], utilize the same vocabulary and training dataset in both the encoder and decoder, resulting in summaries with few novel words. Warm-starting [6] is the other approach that could be utilized for sequence-to-sequence tasks where an encoder-only pretrained model is combined with a decoder-only pretrained model to form a full encoder-decoder model. Warm-starting specifically refers to the process of initializing a model with pretrained weights and parameters and then fine-tuning it for a specific task or domain. Instead of training a model from scratch, this method applies the knowledge and representations obtained during a large-scale pretraining phase to a distinct but related task. In warm-starting models, the encoder is responsible for capturing and encoding the input text into meaningful representations so the model can comprehend the text's semantic meaning and contextual relationships. In contrast, the decoder component receives and interprets these high-level contextualized representations before employing them to generate the required output sequence based on the current task.

In this paper, we describe how to successfully warm-start a variety of models by leveraging various pretrained models' checkpoints for encoders and decoders that have been trained using diverse vocabulary, training sets, learning strategies and objectives to generate coherent summaries with more novel words. In addition, we adapt these models to paraphrase the generated summary while maintaining its syntactic and semantic meaning, resulting in a remarkable enhancement in novelty levels. Finally, we employ the Nucleus Sampling decoding strategy [7] to encourage our models to generate uncommon expressions that increase the amounts of novel words in the generated summaries while maintaining comparable Rouge scores.

Even though the field of abstractive summarization is witnessing a flourishing era in the development of its models and findings, it still lacks an efficient metric for measuring all crucial aspects of the summary. In recent years, the Rouge metric [7] has served as the *de facto* measurement for the vast majority of studies. This metric counts the number of words that overlap between the candidate and the reference summary, which is frequently written by humans. This evaluation provides a realistic estimation of how well the generated summary matches the reference summary in terms of syntax. In light of this, a comprehensive evaluation of abstractive summarization cannot be conducted using the Rouge measure alone. This is because Rouge does not take into consideration other essential aspects of abstractive summarization, such as semantical similarity, faithfulness, and novelty, especially when novel words in the generated and reference summaries are not the same [8].

In this paper, we define a new novelty metric, NovScore, that yields an overall novelty score based on 1-gram, 2-gram, 3-gram, and 4-gram groups to reflect the novelty value of the generated summary. In addition, we define NSSF, a new comprehensive metric that ensembles Novelty, Syntactic, Semantic, and Faithfulness features into a single score in an attempt to simulate human assessment in providing a comprehensive evaluation.

The main contributions of this paper, as shown in Fig. 1, can be summed up as follows:

1) The identification of three novel approaches for abstractive summarization to boost novelty and quality levels. These approaches are:
   - Warm-starting models using various checkpoints.
   - Paraphrasing; i.e., abstractive-then-abstractive strategy.
   - Sampling decoding strategy.
2) The development of two new abstractive summarization metrics:
   - *NovScore* metric to provide a precise novelty score by computing the weighted n-gram scores that are normalized and averaged by the lengths of generated and reference summaries.
   - *NSSF* metric to provide a comprehensive score that balances syntactical and semantical similarity with faithfulness and novelty in a single overall score to improve evaluating abstractive summarization.
3) Our models outperform current state-of-the-art pretrained sequence-to-sequence models on the CNN/Daily Mail corpus in terms of novelty while performing comparably in other metrics.

The following sections are organized as follows: Section II discusses the related work. Section III explains the warm-started models. The new metrics are described in Section IV. Section V and Section VI detail the experimental setups and results. Finally, Section VII concludes the paper.

II. RELATED WORK

A. EFFECTS OF DL AND TL APPROACHES ON ABSTRACTIVE SUMMARIZATION NOVELTY

Many studies [9], [10], [11], [12] began applying various neural DL-based approaches to improve the results of abstractive summarization and address various challenges such as long-term dependencies, out of vocabulary words, inaccurate factual details, repeated statements, fake information, and difficulties in preserving semantic relevancy, key information, faithfulness, and controlling the output [13], [14]. However,

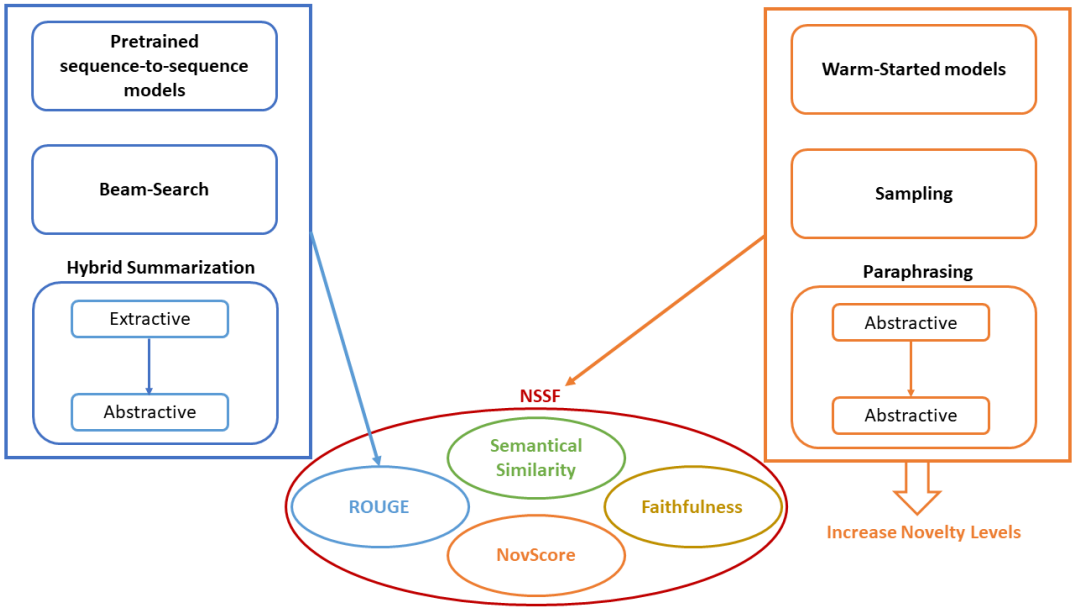


FIGURE 1. Current research trend (left) vs. Our models (right).

the most effective DL-based models have achieved reasonable with extremely minimal novelty results that tend to be more extractive than abstractive. Therefore, Deep Reinforcement Learning (DRL) techniques have begun to address the novelty problem [15], [16], [17], [18], [19]. However, most of this research has failed to attain acceptable novelty rates [13]. In 2017, the Transformer [1] was introduced, followed by a flood of pretrained models that revolutionized the abstractive summarization field through the use of TL-based approaches [2], [3], [4], [5], [20]. As a result, novelty and quality levels have increased, but they still need to be raised further to close the gap with human abstractiveness. Potential contributors to the low novelty problem could be the use of the DL-based pointer-generator mechanism [10], [11], which forces the model to include specific sorts of information from the input text, such as named entities and rare words, to enhance faithfulness. Moreover, many researchers have combined extractive and abstractive summarization methods, usually by utilizing DRL-based approaches, which generate hybrid summaries containing more words found in the input text [15], [16], [18], [21], [22], [23]. Additionally, the novelty of the outcomes can be affected by the novelty of the training dataset. Finally, finetuning pretrained sequence-to-sequence models on abstractive summarization, such as BART, T5, and PEGASUS [4], [5], [24], in which the decoder is trained using the same vocabulary and training dataset as the encoder, could help reduce the number of novel words in the generated summary.

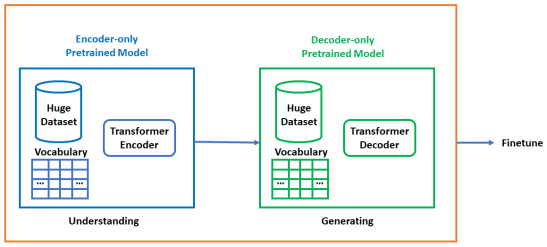
B. PRETRAINED ENCODER-DECODER VS WARM-STARTED MODELS

Multiple pretrained models have been trained using a variety of architectures, features, datasets, and pretraining tasks. Some are proposed as encoder-only for classification and understanding tasks, others as decoder-only for generation tasks, and others as encoder-decoder for sequence-to-sequence tasks. To fine-tune a pretrained model on a sequence-to-sequence task, such as abstractive summarization, one approach is to utilize an encoder-decoder pretrained model, such as BART, T5, or PEGASUS, which has been trained using both the Transformer's encoder and decoder components, as shown in Fig. 2b. The second way is wisely choosing suitable and compatible standalone encoder-only (such as BERT) and decoder-only (such as GPT-2) checkpoints, combining them to construct an encoder-decoder model and then training it on the task in hand, a process known as warm-starting [6], which is illustrated in Fig. 2a.

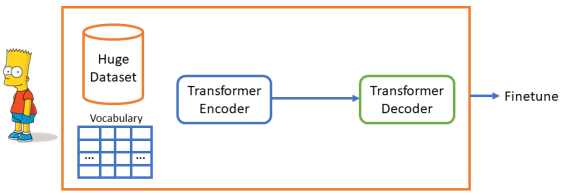
However, encoder-only checkpoints could also be utilized as decoders if the cross-attention layers are randomly initialized while the self-attention layer and the language model head are initialized with the weight parameters of the pretrained model. Section III-A-II describes this behaviour in detail. The authors of [6] evaluated the usefulness of employing BERT, the Robustly Optimized BERT Approach (RoBERTa) [25], and GPT-2 checkpoints to warm-start various models. The experiments were conducted on a variety of tasks, including abstractive summarization. The study concludes that warm starting is efficient as long as the encoder is initialized properly. Moreover, the research demonstrates the importance of sharing weights and vocabulary between the encoder and the decoder, which improves performance and results. The study, however, does not examine the influence of warm-starting on novelty levels.

C. NOVELTY OF DATASETS

As previously mentioned, the novelty degree of the summary generated by a specific model can be influenced by the degree of novelty in the training and evaluation datasets.



(a) Warm-Started Model (ours).



(b) Encoder-Decoder Pretrained Model.

FIGURE 2. **Warm-started models vs Encoder-Decoder models.**

TABLE 1. **Characteristics of CNN/daily mail and XSum datasets.**

		CNN/DM	XSum
Split Size	Train	287,227	204,045
	Validation	13,368	11,332
	Test	11,490	11,334
Mean (Sentences)	Input	30.71	19.77
	Summary	3.78	1.00
Mean (words)	Input	685.17	431.07
	Summary	51.99	23.26
Novelty %	1-g	11.90	34.88
	2-g	50.03	78.78
	3-g	70.26	92.03
	4-g	80.04	96.80
NovScore		64.29	85.57
Density		3.8	1.2

Grusky et al. [26] use Coverage and Density metrics to assess the novelty of datasets. After analyzing the novelty levels of the most prominent abstractive summarization datasets, XSum [27] and CNN/Daily Mail [10], [28], and as shown in Table 1, XSum has high levels of novelty as its NovScore is significantly higher than CNN/Daily Mail. In addition, CNN/Daily Mail has a greater Density, indicating low ratios of novel n-grams. This implies that CNN/Daily Mail is biased towards extractive summaries [5], [26], impacting the novelty of models' outputs. This makes raising the novelty levels of models trained on CNN/DM more challenging.

D. ABSTRACTIVE SUMMARIZATION EVALUATION METRICS

This section sums up the existing evaluation metrics that focus on the three main aspects we consider in our work. First, we discuss syntactical similarity metrics with a particular emphasis on Rouge. Then, we discuss current semantical similarity metrics. Finally, we discuss existing novelty metrics, which we extend as illustrated in Section IV-C.

1) SYNTACTICAL SIMILARITY METRICS

Several metrics are proposed to assess the syntactical similarity, which estimates the number of shared words or phrases between two texts. Examples of such metrics that could be used in text summarization include Rouge [7], Meteor [29], and BLEU [30].

Despite the fact that Rouge [7] was defined roughly two decades ago, it is still used to evaluate the vast majority (95% [31]) of current abstractive summarization research. Rouge provides an excellent first impression of how accurately the candidate summary captures the reference summary (recall) and how much of the candidate summary is relevant (precision). Formally speaking, Rouge has three metrics: Recall ($Rouge_R$), Precision ($Rouge\_P$), and F-measure ($Rouge\_F$). The Recall metric is calculated as follows:

$$Rouge_R = \frac{ow}{rw} \tag{1}$$

where ow and rw denote the number of words that overlap between the candidate and reference summaries, and the total number of words in the reference summary, respectively.

However, this metric has the disadvantage of favouring longer summaries. Precision (i.e., focus), $Rouge_P$, resolves this by determining how much of the candidate summary is relevant, i.e., the suitability of the candidate summary. Precision is calculated as follows:

$$Rouge_P = \frac{ow}{gw} \tag{2}$$

where gw denotes the candidate summary's total word count. Unlike Recall, Precision prefers shorter summaries.

The F-measure metric, $Rouge_F$, balances the Recall and Precision results by computing their harmonic mean as follows:

$$Rouge_F = 2\frac{P_{ROUGE} \cdot R_{ROUGE}}{P_{ROUGE} + R_{ROUGE}} \tag{3}$$

$Rouge_F$ is the most often used metric for calculating the $Rouge^1_F$, $Rouge^2_F$ and $Rouge^L_F$ scores.

$Rouge^1_F$ quantifies the overlap of unigrams, i.e., individual words, in the candidate and reference summaries. $Rouge^2_F$ measures the overlap of bigrams, i.e., every two consecutive words. Finally, $Rouge^L_F$ determines the longest common sequence between the candidate and reference summaries. In particular, $Rouge^1_F$ and $Rouge^2_F$ measure informativeness, whereas $Rouge^L_F$ assesses fluency [6].

The authors of [32] re-evaluated available abstractive summarization metrics based on reliability and human judgments and concluded that $Rouge_F^2$ has the strongest correlation to human evaluations among all existing metrics.

2) SEMANTICAL SIMILARITY METRICS

In recent years, semantic-based measures such as VERT [33], MoverScore [34], and BERTScore [35] have gained popularity for evaluating abstractive summarization models [17], [36], [37]. VERT compares candidate and reference summaries by combining the scores for document-level similarity and word-level dissimilarity to determine the summary's semantical similarity. MoverScore incorporates contextualized embeddings and Earth Mover's Distance [38] to quantify the semantic distance between texts. Lastly, BERTScore computes semantical similarity using BERT contextual embedding vectors at the token level.

The researchers of [39] re-evaluated existing abstractive text summarization metrics and found that BERTScore has a strong correlation with human evaluations and is the most robust measure for assessing the performance of models. This assumption, however, has been contested by [40], who postulated that when comparing whole sentences, word-to-word similarity metrics, such as BERTScore, do not provide semantically relevant sentence embeddings. This implies that their sentence-to-sentence similarities are inaccurate. They also argued that BERT's design is unsuitable for comparing complete sentences in texts. Sentence-BERT [39] was recently introduced as a means for facilitating trustworthy sentence-based semantic textual similarity. Specifically, using Siamese and triplet network structures, a pretrained BERT network was employed to generate semantically relevant sentence embeddings that can subsequently be compared using cosine-similarity.

3) NOVELTY METRICS

Syntactic and semantic metrics provide no insight into the novelty of the summaries. Producing summaries with high copy rates degrades the quality of the findings since novelty is a desirable characteristic that should be retained. Consequently, evaluating novelty is essential for conveying a sense of the level of abstractiveness, thereby enhancing the quality of the evaluation process.

Intuitively, novelty metrics determine the percentage of words in the output summary that do not overlap with the input text. According to Chen and Bansal [16], the novelty score is defined as follows:

$$M(S, T, n) = \frac{||U(s, n) - U(T, n)||}{||(S, n)||} \cdot 100\% \qquad (4)$$

where M is the novelty metric, U calculates unique words, n is the n-grams, S is the candidate summary, T is the input text, and $||X||$ is the number of words in X.

More accurately, the researchers of [15] normalized Chen and Bansal's metric by multiplying it by a new factor, the length ratio between candidate and reference summaries. This adjustment is intended to discourage preferring summaries that are too brief. Their metric is defined as follows:

$$L(S, T, R, n) = \frac{||U(S, n) - U(T, n)||}{||U(S, n)||} \cdot \frac{||S||}{||R||} \qquad (5)$$

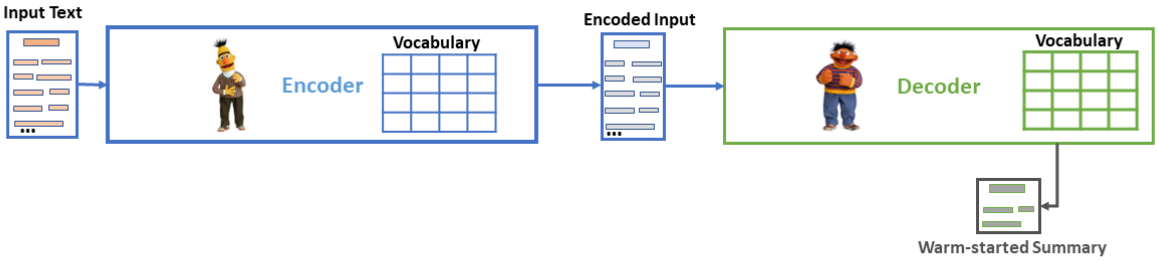
where L is the normalized novelty metric, and R is the reference summary. This metric, however, favours more extended summaries. To bypass this behaviour, we define a new novelty metric, *NovScore*, based on weighted n-gram novelty scores that are normalized and averaged throughout the lengths of candidate and reference summaries, as detailed in Section IV. Nevertheless, Rouge scores are inversely related to novelty scores. That is, as Rouge scores drop, the novelty score rises. Additionally, a reliable evaluation cannot be made just based on a single metric, such as syntactical similarity, semantical similarity, or novelty. Therefore, more investigation into the search for a compromise measurement is necessary. For that purpose, we define a new comprehensive metric, *NSSF*, that encompasses four key aspects of abstractive summarization assessment to provide a reliable evaluation. Detailed explanations of this metric are provided in Section IV.

III. LEARNING MODELS

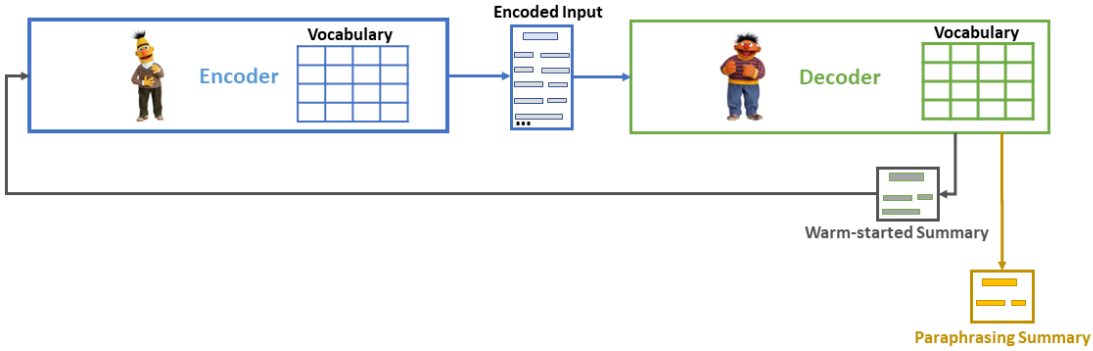
Pretrained sequence-to-sequence models have significantly improved the results of text generation tasks through various aspects including quality, readability, coherency, and generating correct results semantically. For abstractive summarization, additional aspects should be considered, such as improving the quality of the generated summary by considering semantic correlation, faithfulness, fact correctness, conciseness, and generating novel words and sentences. However, most pretrained sequence-to-sequence models produce high-quality summaries with low novelty ratings [13].

To address this problem, we warm-started fifteen models using six individual pretrained models for the model's two major components, the encoder and the decoder, each of which had been trained with distinct objectives and vocabulary. Therefore, the likelihood of the decoder generating a summary using phrases other than those used by the encoder increases, thereby raising the novelty level. As a result, compared to State-of-the-Art (SotA) pretrained sequence-to-sequence models, including BART, PEGASUS, and ProphetNet, our models have significantly improved novelty scores while maintaining comparable Rouge and other metrics scores, demonstrating the significance of our models.

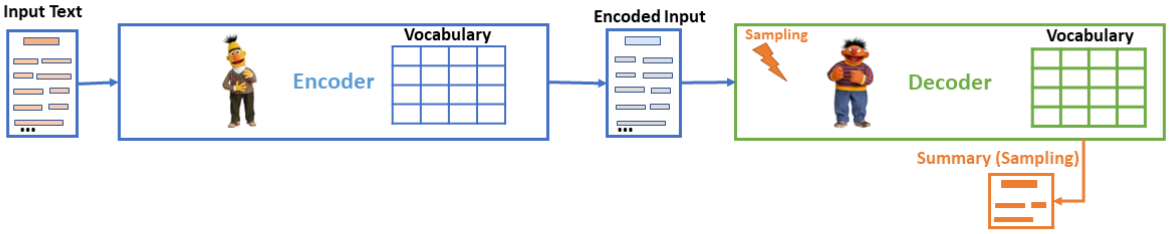
Moreover, to improve the novelty even further, we leveraged the idea of regenerating a new summary based on the prior output. This is accomplished by supplying the model with its own output, which can be defined as paraphrasing. This idea enables the model to focus on the essential parts of the generated summary, which was initially focused on the important parts of the article. By implementing this concept, we were able to generate more concise summaries with more



(a) Warm-started models.



(b) Warm-started models with paraphrasing.



(c) Warm-started models with sampling.

FIGURE 3. Warm-started models (a) with paraphrasing (b) and sampling (c).

new words, focusing on the most relevant elements of the article.

Additionally, because the Nucleus sampling decoding strategy is more suitable for open-ended generation tasks [41], we employ it instead of Beam-Search to increase the level of freedom during generation. This further enhances the novelty of the candidates. Fig. 3 depicts our models, using the bert2ernie model as an example.

A. WARM-STARTED MODELS

1) ENCODERS

The Transformer encoder is designed to read the entire input text at once, with each token considering the context in both directions. This feature has been exploited to efficiently comprehend and represent the meanings of texts by several pretrained encoder models trained on a variety of objectives and datasets. Google's BERT, for example, has pretrained on masked language modeling and next sentence prediction tasks using 16GB of English Wikipedia BooksCorpus datasets. BERT is the first encoder-only pretrained model, representing a quantum leap in the era of pretrained models. This fundamental model is then refined in various ways, as illustrated in Table 2.

We focused our experiments on five high-performing encoders, which are BERT, RoBERTa, A Lite BERT (ALBERT) [42], the eXtreme Learning Network model (XLNet) [43], and the Enhanced Representation through kNowledge IntEgration model (Ernie 2.0) [44]. The details of the models' dimensions are shown in Table 4.

2) DECODERS

For the decoder part, we leverage GPT2, BERT, RoBERTa, and Ernie 2.0. GPT2 is trained on open-ended autoregressive generation with the objective of predicting the next word in a sequence of a few token starters (causal language modeling) utilizing only the Transformer's unidirectional "left-to-right" self-attention decoder. Theoretically, GPT2 is claimed to be the optimal design for use as a decoder. However, with a few adjustments, bi-directional encoder-only pretrained models such as BERT, RoBERTa, and

TABLE 2. Example of pretrained models' enhancements to BERT using Transformer Encoders.

Model	Institute(s)	Goal(s)	Method(s)
RoBERTa	-Facebook AI -University of Washington	To enhance the overall performance	-New training strategy and design choices -Altering key hyperparameters
ALBERT	-Google Research -Toyota Technological Institute	-To conserve more memory, accelerate training, and improve model scaling -To increase the coherence of multiple sentences	-Parameter reduction techniques (Factorized embedding parameterization and Cross-layer parameter sharing) -Inter-sentence coherence loss
XLNet	-Carnegie Mellon University -Google AI Brain	To address the BERT problems of: 1-Relying on input data corruption 2-The independence assumption (all masked tokens are separately reconstructed) 3-The pretrain-finetune discrepancy (special symbols inputs contained in BERT but not in downstream tasks)	Combining the benefits of autoregressive language modelling (Transformer-XL) and autoencoding (BERT) while avoiding their drawbacks
Ernie 2.0	Baidu	-To mine training datasets for lexical, syntactical, semantical, and co-occurring information.	Building and learning pretraining tasks Incrementally using continual multi-task learning

TABLE 3. Example of pretrained models' enhancements to BERT using Transformer Decoders.

Model	Institute(s)	Goal(s)	Method(s)
RoBERTa	-Facebook AI -University of Washington	To enhance the overall performance	-New training strategy and design choices -Altering key hyperparameters
ALBERT	-Google Research -Toyota Technological Institute	-To conserve more memory, accelerate training, and improve model scaling -To increase the coherence of multiple sentences	-Parameter reduction techniques (Factorized embedding parameterization and Cross-layer parameter sharing) -Inter-sentence coherence loss
XLNet	-Carnegie Mellon University -Google AI Brain	To address the BERT problems of: 1-Relying on input data corruption 2-The independence assumption (all masked tokens are separately reconstructed) 3-The pretrain-finetune discrepancy (special symbols inputs contained in BERT but not in downstream tasks)	Combining the benefits of autoregressive language modelling (Transformer-XL) and autoencoding (BERT) while avoiding their drawbacks
Ernie 2.0	Baidu	-To mine training datasets for lexical, syntactical, semantical, and co-occurring information.	Building and learning pretraining tasks Incrementally using continual multi-task learning

Ernie 2.0 may simply be adapted as decoders. To compare the encoder and decoder architectures of the Transformer, encoder blocks consist solely of a bi-directional self-attention layer and two feed-forward layers. By contrast, decoder blocks have a unidirectional self-attention layer, a cross-attention layer, and two feed-forward layers. In addition, a language model head layer follows the decoder blocks, converting the last decoder block's output vectors to logit vectors. As a result, the following steps have been taken to enable an encoder to function as a decoder: First, we alter the self-attention layers to operate unidirectionally, similar to the decoder, and initialize them with the weights from the encoder's self-attention layer. A cross-attention layer is then added between the self-attention layer and the two feed-forward layers. As suggested by [6], we randomize the initialization of this layer's weights, which are subsequently trained while finetuning the model on the summarization task. Finally, we add a language model head layer on top of the last block of the decoder and initialize it with the weights of the encoder's word embeddings. Table 3 highlights the necessary adjustments for this approach. It is worth mentioning that the hidden size of the encoder and decoder in warm-started models must match in order for them to communicate and perform dot products on their respective vectors.

B. PARAPHRASING IN ABSTRACTIVE SUMMARIZATION

Many research [15], [16], [18], [21], [22], [23] utilized hybrid approaches to text summarization employing an extractive-then-abstractive strategy, in which the most salient sentences are extracted from the input text and subsequently paraphrased. In our models, however, we employ an abstractive-then-abstractive strategy in order to increase novelty while maintaining other key aspects. This is the first attempt to employ the abstractive-then-abstractive technique in the abstractive summarization domain to the best of our knowledge. The technique is shown in Fig. 3b.

C. SAMPLING IN ABSTRACTIVE SUMMARIZATION

Nucleus Sampling [41] is proposed as an alternative to Beam-Search to avoid text degeneration by truncating the

TABLE 4. The specifications of the checkpoints used in our experiments.

Checkpoint	Usage	Hidden Size	Vocab size	Hidden Layers	Maximum Position Embeddings	Filter size	Heads	Parameters
BERT-base	Encoder/ Decoder	768	30,522	12	512	3072	12	110M
RoBERTa-base	Encoder/ Decoder	768	50,265	12	514	3072	12	125M
AlBERT-base	Encoder	768	30,000	12	512	3072	12	12M
XLNet-base	Encoder	768	32,000	12	-	3072	12	110M
Ernie 2.0	Encoder/ Decoder	768	30522	12	512	3072	12	-
GPT2-base	Decoder	768	50,257	12	1024	-	12	124M

unreliable tail of the probability distribution, resulting in higher-quality text. Instead of following the distribution of high-probability next words, Nucleus Sampling selects tokens from the smallest feasible set of words whose cumulative probability mass exceeds a predetermined threshold. This behaviour resembles that of the human generation, which favours the use of surprising words to improve fluency. As a result, Nucleus Sampling appears to generate writing that is more natural, fluent, and human-like than conventional Beam-Search.

Nevertheless, Beam-Search is the most common decoding strategy used in research involving directed-generation tasks, such as abstractive summarization, where the output is a constrained transformation of the input with a predicted length. On the other hand, the Sampling decoding strategy is more typically utilized in research involving open-ended text generation tasks such as story generation and text continuation. We defy this trend by adopting Nucleus Sampling on abstractive summarization to encourage generating uncommon phrases during the decoding process and hence raise novelty levels.

Therefore, we conducted a separate set of experiments using the Nucleus sampling decoding strategy to enhance novelty degrees, as shown in Fig. 3c. Section VI discusses the findings.

IV. EVALUATION METRICS

A. MOTIVATION

Currently, no automated abstractive summarization evaluation metric can sufficiently capture all key aspects of the summary. Every metric only focuses on a specific aspect. For example, the syntactic-based metric, Rouge, focuses on the sufficiency of the information by computing the hard overlapping between generated and reference summaries. Whereas semantic-based metrics, such as BERTScore, focus on the semantical similarity, i.e., soft overlapping, using contextual embeddings. Consequently, models respond differently to diverse metrics, confounding evaluation and model preference.

To date, only human evaluation of summaries has been proven to be reliable. It prioritizes the incorporation of relevant content from the source article within an acceptable length [32]. This relevance is satisfied for abstract summarization by favouring syntactical and semantical similarity with high novelty ratings. However, manually evaluating summaries is time-consuming and costly.

In order to provide a reliable automated evaluation that replicates human evaluation, we incorporated various measures that cover the most critical aspects of human review. Specifically, we integrated four independent measures of abstractive summarization: syntactical similarity, semantical similarity, faithfulness to the input text, and novelty.

B. NSSF

In this paper, we have defined NSSF, a new abstractive summarization metric that combines all the aforementioned metrics to form an overall value. We began by comparing the candidate summary to the reference summary in terms of syntactical and semantical similarity. The candidate is then compared against the input text to get its faithfulness and novelty scores.

Specifically, as concluded in Section II-D, we used $Rouge_F^2$ and an MPNET-based sentence-transformers model, as recommended by [32] and [40], to assess the candidate's syntactical and semantical similarity to the reference, respectively.

For semantical similarity measurement, in which summaries sentence embeddings are constructed and subsequently compared semantically, we chose a model based on MPNet pretrained model [45] for several reasons: First, MPNet combines the benefits of BERT and XLNet while avoiding their limitations. Second, because MPNET is not used in any of our models, thus there is no bias favouring any model. Finally, the chosen model1 is appropriate for our models and dataset settings because it maps sentences to a 768-dimensional dense vector space, which is the same dimension as all of our models, and it accepts sentence lengths up to 512, enabling it to accept candidate summary, reference summary, and input document, thereby facilitating the measurement of semantical similarity and faithfulness.

To measure the candidate's faithfulness and fact-consistency with the input text, we compared the sentence

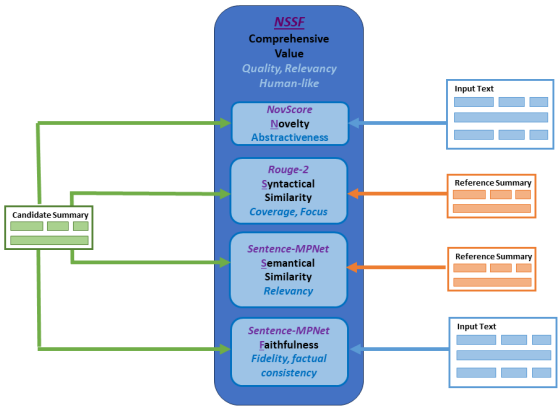


FIGURE 4. Composition of NSSF metric.

embeddings of the candidate and input text utilizing the same model1 used to assess semantical similarity.

To evaluate novelty, we introduced a new metric, *NovScore*, which provides an overall novelty score based on n-grams, where $n \in \{1, 2, 3, 4\}$. Section IV-C discusses this metric in greater depth.

However, there are trade-off behaviours between these measurements. For instance, as indicated in [46], [47], [48], a trade-off between novelty and faithfulness is observed. An increase in the number of novel terms in the generated summary that are not present in the input text makes it prone to contain hallucinatory and contradictory content. Additionally, a trade-off between Rouge and novelty is found.

High novelty summaries are more likely to use synonyms for terms in the reference summaries, resulting in lower overlap and syntactical similarity scores, and thereby poorer Rouge results. Therefore, balancing these measurements by obtaining a high NSSF score is challenging.

Overall, the NSSF value represents all essential aspects that a trustworthy generated abstractive summary should retain. With this value, we can see how closely the generated summary matches human-authored summaries. Fig. 4 illustrates NSSF in detail.

C. FORMULA SPECIFICATION

First, to measure syntactical similarity between the candidate (S) and reference (R) summaries, we used the $Rouge_F^2$ score, as follows:

$$Synt_{SIM}(S, R) = Rouge_F^2(S, R) \tag{6}$$

Second, to evaluate faithfulness and fact consistency, *sentence_mpnet_similarity* is used to calculate the semantical similarity between the candidate summary and the input article by constructing single vector embeddings and then compare them using cosine similarity [40]. The definition of *sentence_mpnet_similarity* is as follows:

$$FF(S, T) = sentence\_mpnet\_similarity(S, T) \tag{7}$$

Third, *sentence_mpnet_similarity* was used to give insights into the semantical similarity dimension of the generated summary to the reference summary. The definition of *sentence_mpnet_similarity* is as follows:

$$Sem_{Sim}(S, R) = sentence\_mpnet\_similarity(S, R) \tag{8}$$

Fourth, to measure the novelty of summaries, we extended (4) and (5) which were described in Section II-D-III. First, we argue that the proper ratio should be taken concerning the entire summary, not only the unique terms in the summary. Based on this assumption, we adjusted (5) by dividing the novel words by the total number of summary terms:

$$N(S, T, R, n) = \frac{||U(S, n) - U(T, n)||}{||(S, n)||} \cdot \frac{||S||}{||R||} \tag{9}$$

where N is the updated normalized novelty metric.

Moreover, to avoid the bias towards longer summaries, we propose, first, having a new value, V, that normalizes (4) by multiplying it by a new factor:

$$V(S, T, R, n) = \frac{||U(S, n) - U(T, n)||}{||(S, n)||} \cdot \frac{||R||}{||S||} \tag{10}$$

Next, we calculated the harmonic mean between N and V to avoid favouring short or long summaries of (10) and (9), respectively. In general, the Harmonic Mean (HM) of n positive numbers is defined as:

$$F = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1} \tag{11}$$

Setting n equal to 2 gives:

$$Nov_F(S, T, R, n) = 2\frac{N(S, T, R, n) \cdot V(S, T, R, n)}{N(S, T, R, n) + V(S, T, R, n)} \tag{12}$$

Finally, the harmonic mean of the four independent metrics was used to calculate *NSSF*. This score reflects a comprehensive assessment of the summaries. Higher levels of syntactical similarity, semantical similarity, faithfulness, and novelty result in a higher NSSF score, indicating that the summaries are of higher quality, more relevant, and more human-like.

Setting n equal to 4 yields:

$$NSSF(s) = \frac{4}{\frac{1}{NovScore(S,T,R)} + \frac{1}{Syn\_Sim(S,R)} + \frac{1}{Sem\_Sim(S,R)} + \frac{1}{FF(S,T)}}$$

It is worth noting that we accorded each of the four metrics equal weight when calculating the NSSF score, as we deemed them equally significant to the final outcome.

As a result, the NSSF value captures the four most essential aspects of abstractive summarization. In addition, NSSF eliminates the drawbacks of the Rouge metric and the previously described trade-off between novelty, Rouge, and faithfulness.

Table 5 details cases in which a single metric resulted in erroneous evaluations, as well as how these evaluations are corrected to yield a more reliable NSSF score.

TABLE 5. Examples of producing error results by different single metrics and how these issues are corrected by the NSSF metric.

Input Article	Bayern Munich moved a step closer to a third straight Bundesliga title thanks to a 3-0 win over Eintracht Frankfurt. Robert Lewandowski boosted his own chances of winning the golden boot award at the end of his first season with Bayern, and a second in a row after finishing the league's top-scorer with Borussia Dortmund last season, by netting two goals in a commanding victory for Pep Guardiola's men. Bayern also had a goal disallowed and they struck the woodwork before Thomas Muller added a third (. . .) Lewandowski flies through the air in celebration after scoring the opener during Saturday's Bundesliga clash. The Poland international punches the air with delight maintained their advantage at the top of the table. (. . .)Thomas Muller added a third, with few doubts from start to finish that they would pick up all three points once again. (. . .)Bayern had the ball in the back of the net again shortly before half-time, but after consultation with his linesman, the referee ruled out Muller's effort for offside. (. . .) The Poland international punches (. . .) The hosts did not have long to wait for their second, though, with Gotze's shot blocked by Makoto Hasebe, but the ball fell kindly for Lewandowski to tuck in for his 16th goal of the season. (. . .) Frontman Thomas Muller celebrates after scoring Bayern Munich's third goal with eight minutes to go .		
Reference Summary	Robert Lewandowski scored twice as Bayern Munich claimed 3-0 victory. Thomas Muller scored with eight minutes remaining to complete win. Bayern Munich maintained their lead at the top of the Bundesliga.		
Problematic Candidate Summary 1	Robert Lewandowski scored twice as Bayern Munich lose 3-0. Thomas Muller scored with the beginning eight minutes to complete win . Bayern Munich maintained their lead at the bottom of the Bundesliga table.		
Single Evaluation Error	**Problem**	**NSSF Correction Metrics**	**Justification**
Rouge High Score: *Summaries that have high levels of extractiveness and different meanings achieve high Rouge scores*	The candidate summary has nearly identical terms to the reference summary, but with a few contradicting words that modify the entire meaning.	Semantic Similarity and Faithfulness metrics will correct this score.	Semantic Similarity will decrease the NSSF score because it assesses full semantic sentence similarity between the candidate and the reference. Faithfulness will decrease the NSSF score because it assesses full semantic sentence similarity between the candidate and the article.
Problematic Candidate Summary 2	With two goals, Robert Lewandowski led Bayern Munich to a 3-0 triumph. Thomas Muller struck the game-winning goal in the 82nd minute. Bayern Munich remained in first place in the Bundesliga.		
Single Evaluation Error	**Problem**	**NSSF Correction Metrics**	**Justification**
Rouge Low Score: *Summaries with high levels of novelty achieve low Rouge scores.*	The candidate summary has the same meaning as the reference summary but uses different terms. This behavior should be rewarded based on the principles of abstractive summarization, but it actually received a low rating.	Semantic Similarity and NovScore will correct this score.	Semantic Similarity will increase the NSSF score because it assesses full semantic sentence similarity between the candidate and the reference. NovScore will correct the NSSF score because it evaluates the degree of novelty between the candidate summary and the input article.
Problematic Candidate Summary 3	Robert Lewandowski netted two goals as Bayern Munich won 3-0 over Eintracht Frankfurt. Thomas Muller celebrates after scoring Bayern Munich's third goal with eight minutes to go. Bayern Munich maintained their advantage at the top of the table.		
Single Evaluation Error	**Problem**	**NSSF Correction Metrics**	**Justification**
Semantic Similarity High Score: *Summaries with high levels of extractiveness achieve high Semantic Similarity Scores.*	The candidate summary has the same meaning as the reference summary using the same terms as the input article, violating one of the fundamental principles of abstractive summarization.	NovScore and Rouge will correct this score.	NovScore will decrease the NSSF score because it evaluates the degree of novelty between the candidate summary and the input article. Rouge will decrease NSSF score because it measures the number of overlapped words between the candidate summary and the reference summary (not the input article).
Problematic Candidate Summary 4	Cristiano Ronaldo scored three goals as Real Madrid defeated Barcelona by a score of 3-0. In the 85$^{\text{th}}$ minute, the game-winning goal was scored by Leonel Messi. Juventus maintained their advantage atop the Seria A standings.		

TABLE 5. *(Continued.)* **Examples of producing error results by different single metrics and how these issues are corrected by the NSSF metric.**

Single Evaluation Error	Problem	NSSF Correction Metrics	Justification
NovScore High Score: *Summaries with many novel words that change the meaning achieve a high NovScore rating.*	The candidate summary has a different meaning than the reference summary due to the use of so many different terms that do not exist in the input article, resulting in hallucinatory and contradictory content.	Rouge, Semantic Similarity, and Faithfulness will correct this score	As Rouge and Semantic Similarity compare the syntactical and semantical similarities of the candidate summary to the reference summary, they will both decrease the NSSF score because the candidate summary and reference summary have distinct meanings. Similarly, the faithfulness metric will correct the NSSF score because the candidate summary content contradicts the input text details, resulting in a low score.
Problematic Candidate Summary 5	Bayern Munich advanced closer to a third consecutive Bundesliga championship with a 3-0 victory over Eintracht Frankfurt. Robert Lewandowski increased his prospects of winning the golden boot award at the conclusion of his first season with Bayern, and a second in a row after finishing as the league's leading scorer with Borussia Dortmund last season, by scoring two goals in a decisive victory for Pep Guardiola's squad. Lewandowski leaps into the air in jubilation after scoring the game's first goal on Saturday in the Bundesliga. The Poland international strikes the air with glee as his team maintains their table-topping position. Thomas Muller celebrates after netting the third goal for Bayern Munich with eight minutes remaining.		
Single Evaluation Error	**Problem**	**NSSF Correction Metrics**	**Justification**
Faithfulness High Score: *Long Summaries with the same meaning and/or words achieve high Faithfulness rating.*	The candidate summary has the same meaning as the source article using the same or different words, but it is significantly longer than the reference summary, thus violating one of the primary characteristics of summarization (i.e., producing concise text).	NovScore and Rouge will correct this score	Both NovScore and Rouge scores are calculated based on the length of the candidate summary, which will result in low scores in this case, thereby correcting the NSSF score.

V. EXPERIMENTAL SETUPS

A. DATASET

We evaluate our models using the non-anonymous version of CNN/Daily Mail, the most widely used dataset for abstractive single-document summarization. As illustrated in Table 1, this corpus consists of 287k training pairs, 13k validation pairs, and 11k testing pairs. Each entry contains a CNN or Daily Mail English news article accompanied by a multi-sentence summary. This corpus includes documents of 30 sentences and 700 words and summaries of 3-4 sentences and 50 words. However, we truncated the input articles to 512 tokens during training following the encoders' maximum position embeddings.

B. TRAINING DETAILS

In all our experiments, we employ five HuggingFace pretrained checkpoints for the encoder part: BERT (bert-base-uncased),[1] RoBERTa (roberta-base),[2] AlBERT (albert-base-v2),[3] XLNet (xlnet-base-cased),[4] and Ernie 2.0 (nghuyong/ernie-2.0-en).[5] For the decoder, we employ four checkpoints: GPT2 (gpt2),[6] BERT (bert-base-uncased),[7] RoBERTa (roberta-base),[8] and Ernie 2.0 (nghuyong/ernie-2.0-en).[9] Then we selected the top fifteen models.

Since all the models leverage the base checkpoints, most of them have a hidden size of 768, 12 hidden layers, a filter size of 3072, and 12 attention heads. We utilize an 8-batch size and the Adam optimizer with betas equal to (0.9,0.999) and epsilon equal to 1e-08. The learning rate, warmup steps, and total finetune epoch are set at 5e-05, 2000, and 3, respectively.

We follow the majority of encoders' maximum capacity and limit the length of input articles to 512 tokens. Likewise, the decoders' maximum length is limited to 128 tokens, with a 2.0 length penalty during inference. As a result, most of our models generate summaries with lengths between 50 and 69 tokens. For decoding parameters, we employ Beam-Search with a beam size of 4, remove the duplicated trigrams, and sample with p=0.99. For the framework versions,

[1] https://huggingface.co/bert-base-uncased
[2] https://huggingface.co/roberta-base
[3] https://huggingface.co/albert-base-v2
[4] https://huggingface.co/xlnet-base-cased
[5] https://huggingface.co/nghuyong/ernie-2.0-en
[6] https://huggingface.co/gpt2
[7] https://huggingface.co/bert-base-uncased
[8] https://huggingface.co/roberta-base
[9] https://huggingface.co/nghuyong/ernie-2.0-en

TABLE 6. Results of different models on the CNN/Daily mail dataset. The first two blocks represent the baseline models. The remaining blocks represent our models grouped by the decoder used. R is short for Rouge. Bold values are the top performing models in our models. Underlined values are the overall top-performing models in each measurement.

Model	Syntactical Similarity	Semantic Similarity	Faithfulness	NovScore	NSSF	R-1	R-Lsum
Baseline Models - Pretrained							
Bart	21.28	**78.504**	82.925	20.262	33.019	44.16	**41.785**
Pegasus	**21.47**	76.493	80.59	30.756	38.251	44.17	36.871
ProphetNet	21.17	76.581	80.716	34.257	**39.264**	**44.2**	39.115
Bert2Bert_base	17.84	N/A	N/A	N/A	N/A	39.02	36.29
Bert2GPT_base	4.96	N/A	N/A	N/A	N/A	25.2	22.99
RoBerta2GPT_base	14.72	N/A	N/A	N/A	N/A	36.35	33.79
Baseline Models - Non-Pretrained							
Pointer-Gen + Coverage	17.28	N/A	N/A	6.15200	N/A	39.53	36.38
Entail (RL)	17.51	N/A	N/A	6.59800	N/A	39.53	36.44
SumGAN	17.65	N/A	N/A	7.69200	N/A	39.92	36.71
WPLoss-1atthead	14.93	N/A	N/A	17.50400	N/A	36.48	31.55
rnnext+abs+RL+rerank	17.8	N/A	N/A	21.18000	N/A	40.88	38.54
ML+RLROUGE+Novel, with LM	17.38	N/A	N/A	28.69300	N/A	40.19	37.52
GPT2-Decoder Models (ours)							
roberta2gpt	17.681	**77.399**	**85.273**	35.178	36.486	39.147	36.885
bert2gpt	16.61	75.983	84.416	35.04	35.165	38.008	35.764
ernie2gpt	16.779	76.084	84.545	34.113	35.123	38.178	35.947
Ernie-Decoder Models (ours)							
ernie2ernie	19.027	74.551	79.159	25.897	34.124	41.275	38.479
bert2ernie	18.798	74.195	78.961	24.712	33.387	41.079	38.219
roberta2ernie	19.227	74.916	79.335	31.204	36.36	41.812	39.002
Bert-Decoder Models (ours)							
bert2bert	18.514	74.209	79.156	28.888	34.862	40.977	38.152
albert2bert	18.585	74.283	79.222	31.115	35.705	41.243	38.453
roberta2bert	19.065	74.79	79.153	34.659	37.275	41.658	38.898
ernie2bert	19.024	74.649	79.257	28.725	35.276	41.414	38.616
xlnet2bert	18.421	74.41	79.409	32.212	35.919	41.034	38.225
Roberta-Decoder Models (ours)							
roberta2roberta	**19.967**	76.208	80.895	33.947	**38.087**	**42.732**	**40.089**
bert2roberta	18.46	74.427	79.548	**36.086**	37.075	41.182	38.581
albert2roberta	18.441	74.42	79.422	35.433	36.874	41.251	38.658
ernie2roberta	18.875	74.258	78.722	32.581	36.416	41.357	38.706

we use Huggingface Transformers 4.12.0.dev0, Pytorch 1.10.0+cu111, Datasets 1.18.3, and Tokenizers 0.10.3.

For the framework versions, we use Huggingface Transformers 4.12.0.dev0, Pytorch 1.10.0+cu111, Datasets 1.18.3, and Tokenizers 0.10.3.

C. BASELINES

Our models are compared to several baselines, which are categorized as non-pretrained models, pretrained models, and warm-started models as follows:

Non-pretrained Models

- **Pointer-Gen + Coverage** [11]: This model uses the pointer-generator network to copy tokens from the input text while having the ability to generate new ones.
- **SumGAN** [18]: This model uses the adversarial network and reinforcement learning policy gradient to combine extractor and abstractor models to generate a coherent summary with novel words.

TABLE 7. Paraphrasing results.

Model	Syntactical Similarity	Semantical Similarity	Faithfulness	NovScore	NSSF	R-1	R-Lsum
GPT2-Decoder Models							
roberta2gpt_paraphrase	15.072	73.964	81.588	46.260	35.167	36.004	233.982
bert2gpt_paraphrase	14.216	72.150	80.264	45.297	33.687	34.970	232.992
ernie2gpt_paraphrase	14.329	72.117	80.211	44.133	33.675	35.091	233.096
Ernie-Decoder Models							
ernie2ernie_paraphrase	17.003	70.593	74.295	31.770	33.923	38.039	235.548
bert2ernie_paraphrase	16.530	69.860	73.566	32.263	33.501	37.508	235.055
roberta2ernie_paraphrase	14.834	65.608	69.114	43.645	33.323	35.710	233.360
Bert-Decoder Models							
bert2bert_paraphrase	16.282	69.796	73.952	36.789	34.350	37.609	235.159
albert2bert_paraphrase	16.387	70.250	74.053	39.555	35.074	37.968	235.526
roberta2bert_paraphrase	14.373	64.828	68.450	47.528	33.154	35.392	233.083
ernie2bert_paraphrase	16.552	69.875	73.515	37.180	34.714	37.771	235.334
xlnet2bert_paraphrase	12.791	61.249	65.326	48.023	30.619	33.269	231.077
Roberta-Decoder Models							
roberta2roberta_paraphrase	17.741	72.313	75.394	41.011	37.091	39.393	236.977
bert2roberta_paraphrase	16.067	69.638	73.449	42.399	35.149	37.526	235.193
albert2roberta_paraphrase	15.986	69.369	73.015	43.983	35.273	37.601	235.310
ernie2roberta_paraphrase	16.553	69.816	73.265	38.960	35.072	37.830	235.420

- **WPLoss-1atthead** [17]: This reinforcement learning-based model employs multi-head attention, pointer dropout, and new loss functions to encourage generating more novel words while maintaining high Rouge scores.
- **rnn-ext+abs+RL+rerank** [16]: This hybrid extractive-abstractive summarization model uses reinforcement learning sentence-level policy gradient to increase novel words and preserve language fluency.
- **ML+RL ROUGE+Novel, with LM** [15]: This reinforcement learning-based model that develops a novelty reward to encourage generating new words.

Pre-trained Models

- **BART-large** [4]: This pretrained sequence-to-sequence model was finetuned for abstractive text summarization and achieved SotA Rouge results.
- **PEGASUS** [5]: This pretrained sequence-to-sequence model trained on the gap sentences generation task, essentially designed for the abstractive summarization. The model achieved SotA Rouge results.
- **ProphetNet** [20]: This pretrained sequence-to-sequence model was trained on an objective related to abstractive summarization, which is predicting future n-gram, to predict multiple future tokens based on previous context tokens. This objective extremely enhanced novelty scores while still achieving SotA Rouge results.

Warm-Started Models Three warm-started models are chosen from [6] as baselines. To distinguish these baseline models from our warm-started models, we label them Bert2Bert_base, Bert2GPT_base, and RoBerta2GPT_base.

VI. EXPERIMENTAL RESULTS

A. ANALYSIS

As seen in Table 6, our warm-started models outperform all baseline models on novelty and faithfulness metrics while attaining comparable syntactical-similarity, semantical-similarity, and NSSF scores to SotA models.

This shows that we achieved the main goal of this research of improving the novelty of abstractive summaries without compromising quality. Particularly, the bert2roberta model achieves the best NovScore while maintaining comparable other scores. RoBerta2GPT has the highest faithfulness metric. PEGASUS and BART perform the best in syntactical- and semantical- similarity metrics, respectively. Finally, the model with the best overall performance is ProphetNet, which produces the most human-like summaries and achieves the highest score on the overall quality metric, NSSF.

B. ABLATION STUDY

To determine the impact of each individual contribution, we compare our models to one another and to baselines. First, we show the effect of warm-starting models with different

TABLE 8. Sampling results.

Model	Syntactical Similarity	Semantical Similarity	Faithfulness	NovScore	NSSF	R-1	R-Lsum
GPT2-Decoder Models							
roberta2gpt_sampling	17.345	77.273	85.104	37.411	36.672	38.876	36.656
bert2gpt_sampling	16.286	75.705	84.150	37.538	35.356	37.719	35.506
ernie2gpt_sampling	16.519	75.863	84.216	36.354	35.367	37.943	35.736
Ernie-Decoder Models							
ernie2ernie_sampling	18.805	74.525	79.245	28.519	35.002	41.083	38.333
bert2ernie_sampling	18.603	74.169	78.981	27.225	34.297	40.897	38.094
roberta2ernie_sampling	19.052	75.075	79.583	33.583	36.986	41.669	38.900
Bert-Decoder Models							
bert2bert_sampling	18.275	74.432	79.502	31.660	35.613	40.861	38.069
albert2bert_sampling	18.186	74.375	79.450	34.408	36.336	40.987	38.226
roberta2bert_sampling	18.686	74.807	79.408	37.131	37.591	41.421	38.687
ernie2bert_sampling	18.975	74.826	79.584	30.299	35.831	41.437	38.616
xlnet2bert_sampling	18.115	74.443	79.643	34.718	36.365	40.786	38.030
Roberta-Decoder Models							
roberta2roberta_sampling	19.440	76.020	80.787	36.403	38.298	42.284	39.686
bert2roberta_sampling	17.946	74.186	79.428	38.742	37.173	40.831	38.251
albert2roberta_sampling	17.972	74.010	79.074	38.715	37.164	40.866	38.307
ernie2roberta_sampling	18.687	74.240	78.634	34.754	36.875	41.246	38.633

encoder and decoder checkpoints and vocabulary on novelty degrees. Then, we demonstrate how paraphrasing affects all measures. Finally, we examine the impact of using sampling with various parameters. Fig. 5 summarizes the results of all methods.

1) DIFFERENT ENCODERS AND DECODERS

To show the effect of having different encoders and decoders with varying vocabularies, training sets, and learning strategies on novelty degrees, twelve of the fifteen warm-started models have been built using different encoder and decoder checkpoints. The impact on novelty scores is illustrated in Fig. 6. It was observed that models that use the same checkpoint for both the encoder and decoder, i.e., bert2bert, roberta2roberta, and ernie2ernie, have the worst NovScore results. It is noticed that Ernie-based models achieve poor novelty results and are hence excluded from the comparisons of RoBERTa-decoder and BERT-decoder models.

2) PARAPHRASING

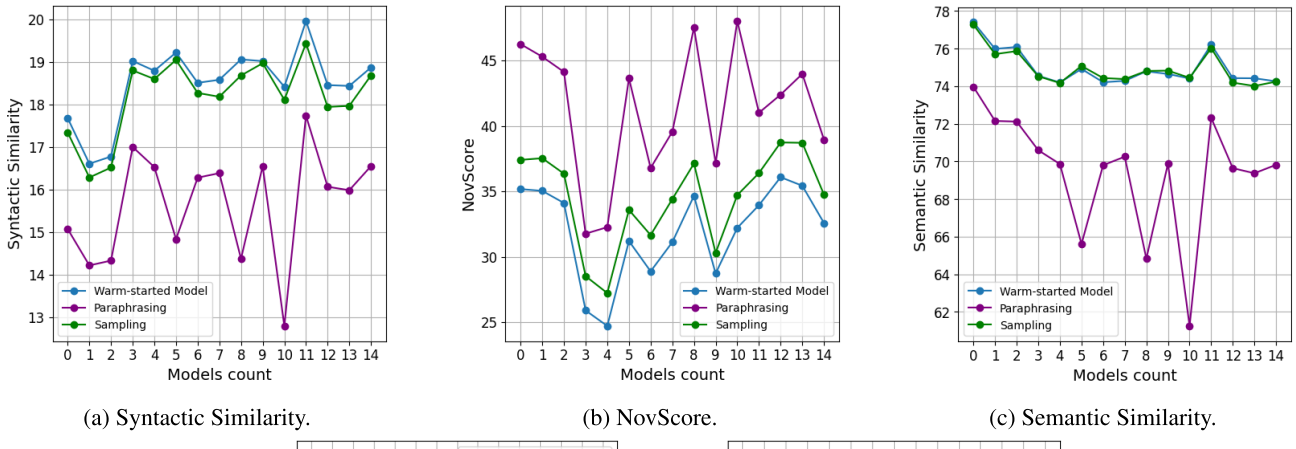
As shown in Table 7, the paraphrasing approach enhanced novelty scores significantly while achieving a predictable decrease in other measures. The xlnet2bert_ Paraph model performed the best in terms of novelty, but poorly in other measurements. As a result, paraphrasing only enhances the novelty side of the summary, leaving plenty of room to improve other aspects in the research space.

3) SAMPLING

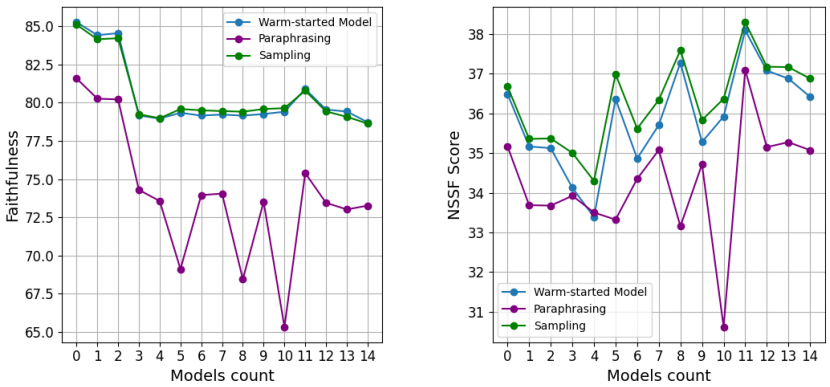
We applied Nucleus sampling, discussed in Section III-C, to encourage producing novel words while maintaining the summary's meaning. The outcomes are shown in Table 8. Compared to the original warm-started models, this strategy improves the overall NSSF score by boosting novelty scores and achieving competing performance on other metrics. As a result, sampling improves the overall novelty and quality of the results.

C. DISCUSSION

Fig. 5 shows the findings of the original warm-started, paraphrasing, and sampling models based on syntactical similarity, NovScore, faithfulness, semantical similarity, and NSSF results. The findings indicate that, compared to warm-starting, paraphrasing achieved the highest levels of novelty. However, this was achieved at the expense of summary quality. In contrast, sampling was successful in better balancing all metrics. By increasing NovScore and NSSF scores, sampling improved the novelty and quality of the summaries. As a result, the findings indicate that although Nucleus sampling is better suited for open-ended generation

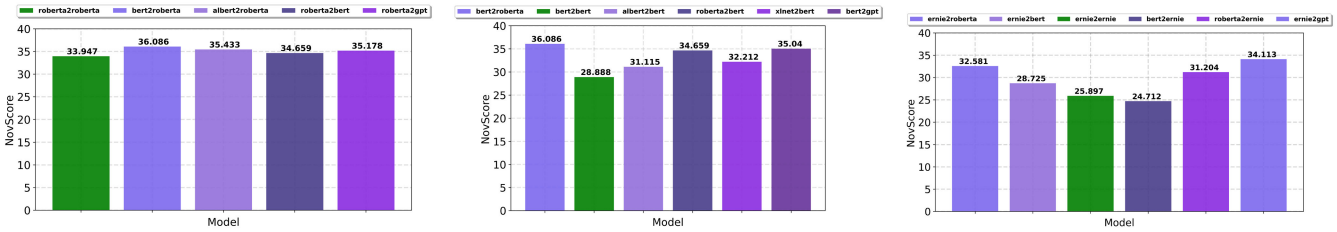


(a) Syntactic Similarity. (b) NovScore. (c) Semantic Similarity.



(d) Faithfulness. (e) NSSF.

FIGURE 5. Performance in different metrics for warm-started models vs paraphrasing vs sampling.



(a) Roberta Models. (b) Bert Models. (c) Ernie Models.

FIGURE 6. NovScore enhancement of warm-started models with different checkpoints (purples) vs with the same checkpoints (green).

tasks, it may also be employed efficiently for directional generation tasks, such as abstractive ATS.

Overall, Fig. 8 depicts the novelty improvement above baseline and SotA models. Additionally, as illustrated in Fig. 7, our warm-started models achieve higher Rouge scores compared to baseline warm-started models, indicating an improvement in the coverage and focus of warm-starting findings in the field of abstractive ATS.

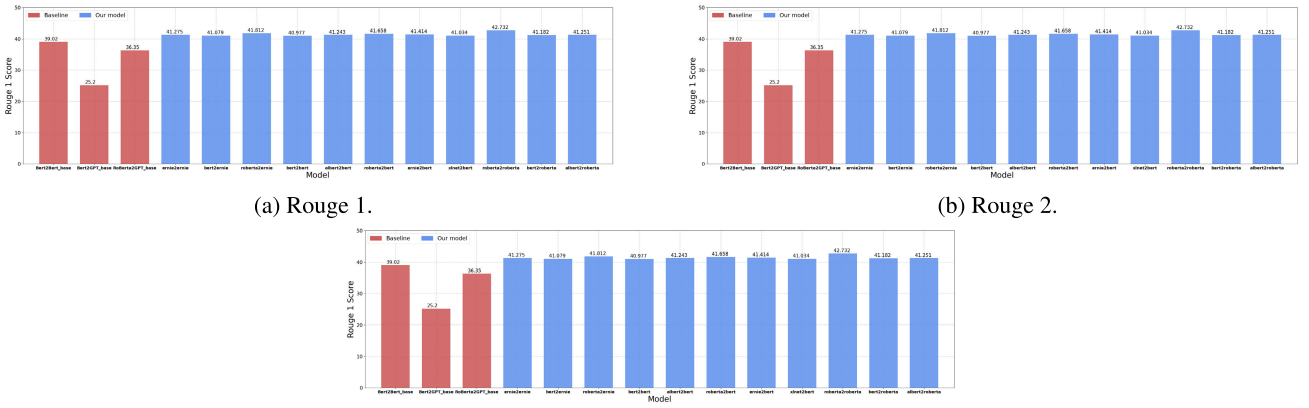
D. TRADE-OFFS

1) NOVELTY VS ROUGE

To be able to achieve high Rouge scores, models must produce summaries using the exact words and phrases as reference summaries. In the CNN/Daily Mail dataset, reference summaries include 88% of the single terms found in the input text, i.e., 12% 1-gram novelty. On the other hand, high novelty scores can be achieved by including a wide range of words in the output summary that are not found in the input text and/or by minimizing the 2-gram, 3-gram, and 4-gram identical similarity to the input text. Fig. 9 demonstrates that NovScore increases as syntactical similarity decreases and vice versa, resulting in a trade-off between these two features, indicating that balancing them is challenging.

2) NOVELTY VS FAITHFULNESS AND FACT CONSISTENCY

Outputs containing fewer overlapping words with the input text are more likely to be unfaithful [46]. That is why extractive summaries are more faithful and factually consistent with the input text. Conversely, summaries with higher novelty levels tend to be less faithful and factual consistent. This is because inaccurately exchanging terminology can modify the



(a) Rouge 1.

(b) Rouge 2.

(c) Rouge LSum.

FIGURE 7. Rouge (1, 2, 3) enhancement of our models compared to warm-started baseline models.

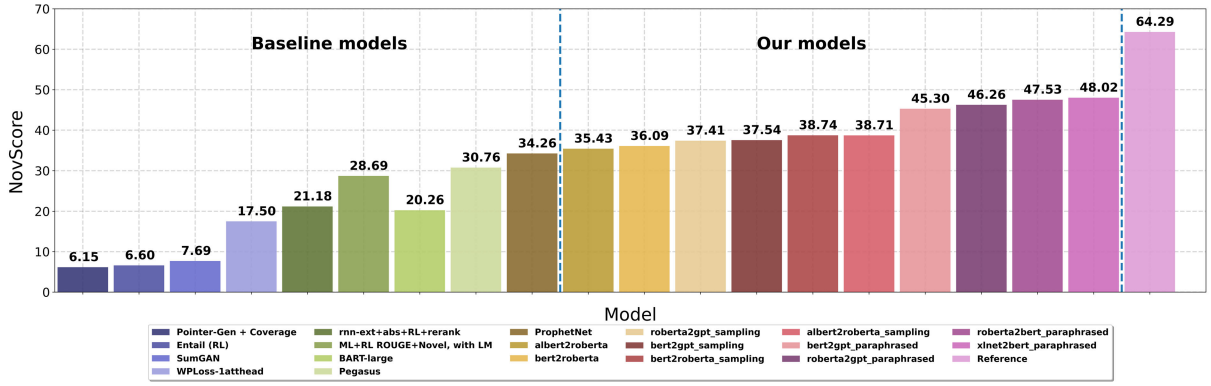


FIGURE 8. Novelty improvement of our models over the baseline models.

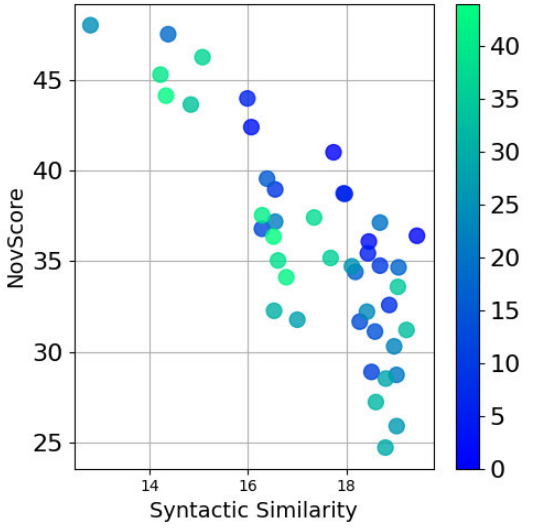


FIGURE 9. The trade-off behaviour between syntactical similarity and NovScore measurements.

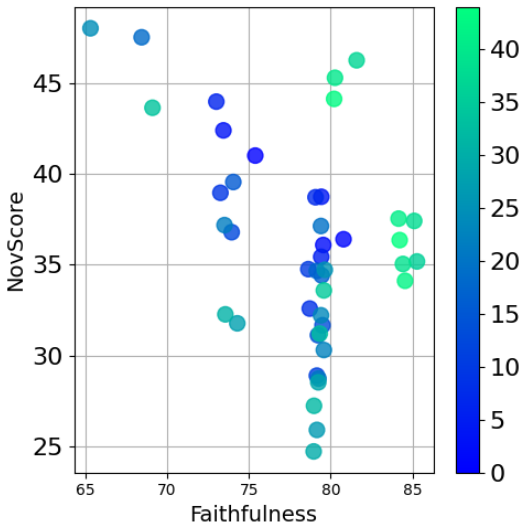


FIGURE 10. The trade-off behaviour between faithfulness and NovScore measurements.

meaning, resulting in lower levels of faithfulness and fact consistency. This tendency can be clearly seen in Fig. 10. Balancing these two aspects is a challenging task.

VII. CONCLUSION

This paper has addressed the issue of limited novelty in summaries generated by models trained on the same dataset

with a shared vocabulary for the encoder and decoder. By implementing warm-starting, in which the encoder and decoder were trained on separate datasets and vocabularies, and by employing other innovative techniques in the field of abstractive text summarization, we successfully increased the novelty levels of the generated summaries while maintaining other evaluation scores. We used warm-starting strategies by initializing fifteen models, twelve of which used separate checkpoints for the encoder and decoder. With this method, the novelty levels of the generated summaries significantly improved. To further enhance novelty, we employed the paraphrasing method and the Nucleus sampling decoding strategy. Paraphrasing creates variations of the summaries, but only focuses on enhancing their novelty. Nucleus sampling provides a broader range of candidate tokens during decoding, thereby facilitating the generation of more novel and creative summaries without sacrificing other evaluation aspects. Besides, we introduced two novel abstractive text summarization metrics: NovScore and NSSF, to reliably evaluate novelty levels and overall summary performance. NovScore evaluates the novelty of the generated summaries, revealing how well the models produce novel content. NSSF, on the other hand, is a comprehensive metric that evaluates the overall performance of the summary by considering multiple factors. As a result, among the models we evaluated, the bert2roberta and bert2roberta_sampling models achieved the goal of this study by obtaining the highest NovScore while maintaining comparable NSSF and other scores, showcasing their ability to produce summaries with a high level of novelty while maintaining good quality in other areas. The ProphetNet model received the highest NSSF rating, demonstrating its ability to produce summaries that closely resemble those written by humans.

ACKNOWLEDGMENT

The authors would like to thank Applied Science Private University, Amman, Jordan, for their support in conducting this research. They also like to thank Al-Ahliyya Amman University, for providing all the necessary support to conduct this research work. They also like to thank Prof. Mohammad A. Omary (Distinguished Professor of Chemistry, Physics, and Mechanical Engineering) from the University of North Texas/USA, for his efforts in proofreading the manuscript and improving its technical writing and readability.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6009.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL HLT*, 2019, pp. 4171–4186.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[5] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11328–11339.

[6] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 264–280, Jun. 2020, doi: 10.1162/tacl_a_00313.

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out (ACL)*, pp. 74–81.

[8] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Exp. Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679, doi: 10.1016/j.eswa.2020.113679.

[9] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for sentence summarization," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2015, pp. 379–389.

[10] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, 2016, pp. 280–290, doi: 10.18653/v1/k16-1028.

[11] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1073–1083, doi: 10.18653/v1/P17-1099.

[12] Z. Cao, W. Li, F. Wei, and S. Li, "Retrieve, Rerank and rewrite: Soft template based neural summarization," in *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, 2018, pp. 152–161.

[13] A. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101276, doi: 10.1016/j.csl.2021.101276.

[14] S. Liu, J. Cao, R. Yang, and Z. Wen, "Key phrase aware transformer for abstractive summarization," *Inf. Process. Manag.*, vol. 59, no. 3, May 2022, Art. no. 102913, doi: 10.1016/j.ipm.2022.102913.

[15] W. Kryscinski, R. Paulus, C. Xiong, and R. Socher, "Improving abstraction in text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1808–1817, doi: 10.18653/v1/d18-1207.

[16] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 675–686, doi: 10.18653/v1/p18-1063.

[17] F. Boutkan, J. Ranzijn, D. Rau, and E. van der Wel, "Point-less: More abstractive summarization with pointer-generator networks," 2019, *arXiv:1905.01975*.

[18] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 8109–8110, Apr. 2018.

[19] M. Zhang, G. Zhou, W. Yu, and W. Liu, "FAR-ASS: Fact-aware reinforced abstractive sentence summarization," *Inf. Process. Manag.*, vol. 58, no. 3, 2021, Art. no. 102478, doi: 10.1016/j.ipm.2020.102478.

[20] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training," 2020, *arXiv:2001.04063*.

[21] W. T. Hsu, C. K. Lin, M. Y. Lee, K. Min, J. Tang, and M. Sun, "A unified model for extractive and abstractive summarization using inconsistency loss," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 132–141.

[22] S. Gehrmann, Y. Deng, and A. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4098–4109, doi: 10.18653/v1/d18-1443.

[23] Q. Wang, P. Liu, Z. Zhu, H. Yin, Q. Zhang, and L. Zhang, "A text abstraction summary model based on BERT word embedding and reinforcement learning," *Appl. Sci.*, vol. 9, no. 21, p. 4701, Nov. 2019, doi: 10.3390/app9214701.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[26] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 708–719, doi: 10.18653/v1/n18-1065.

[27] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1797–1807.
[28] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend NIPS 2015," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.
[29] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380, doi: 10.3115/v1/w14-3348.
[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguistics*, 2002, pp. 311–318, doi: 10.1002/andp.19223712302.
[31] F. Koto, T. Baldwin, and J. H. Lau, "FFCI: A framework for interpretable automatic evaluation of summarization," *J. Artif. Intell. Res.*, vol. 73, pp. 1–53, Apr. 2022.
[32] M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu, and G. Neubig, "Re-evaluating evaluation in text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 9347–9359, doi: 10.18653/v1/2020.emnlp-main.751.
[33] J. Krantz and J. Kalita, "Abstractive summarization using attentive neural techniques," 2018, *arXiv:1810.08838*.
[34] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 563–578, doi: 10.18653/v1/d19-1053.
[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.
[36] H. Jin, T. Wang, and X. Wan, "SemSUM: Semantic dependency guided neural abstractive summarization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8026–8033, doi: 10.1609/aaai.v34i05.6312.
[37] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," 2021, *arXiv:2106.01890*.
[38] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[39] A. R. Fabbri, W. Kryscinski, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 391–409, Apr. 2021.
[40] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992, doi: 10.18653/v1/d19-1410.
[41] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proc. CEUR Workshop*, vol. 2540, 2019, pp. 1–16.
[42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.
[43] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–18.
[44] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 8968–8975, doi: 10.1609/aaai.v34i05.6428.
[45] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," in *Proc. NIPS*, 2020, pp. 16857–16867.
[46] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization,", pp. 5055–5070, 2020, doi: 10.18653/v1/2020.acl-main.454.
[47] F. Ladhak, E. Durmus, H. He, C. Cardie, and K. McKeown, "Faithful or extractive? On mitigating the faithfulness-abstractiveness trade-off in abstractive summarization," 2021, *arXiv:2108.13684*.
[48] T. Goyal, J. Xu, J. Jessy Li, and G. Durrett, "Training dynamics for text summarization models," 2021, *arXiv:2110.08370*.

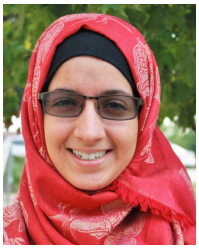
AYHAM ALOMARI received the master's degree in computer science from Yarmouk University, Jordan, in 2010, and the Ph.D. degree in artificial intelligence from the University of Malaya, Malaysia, in 2023. He is currently an Assistant Professor with the Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Jordan. His research interests include artificial intelligence, machine learning, natural language processing, text summarization, machine translation, sentiment analysis, deep learning, transfer learning, pretrained models, and reinforcement learning.

AHMAD SAMI AL-SHAMAYLEH received the master's degree in information systems from The University of Jordan, Jordan, in 2014, and the Ph.D. degree in artificial intelligence from the University of Malaya, Malaysia, in 2020. He is currently an Assistant Professor with the Faculty of Information Technology, Al-Ahliyya Amman University, Jordan. His research interests include artificial intelligence, human–computer interaction, the IoT, Arabic NLP, Arabic sign language recognition, language resources production, the design and evaluation of interactive applications for handicapped people, multimodality, and software engineering.

NORISMA IDRIS received the Ph.D. degree in computer science from the University of Malaya, in 2011. She joined the Faculty of Computer Science and Information Technology, University of Malaya, in 2001, where she is currently an Associate Professor with the Artificial Intelligence (AI) Department. She is also working on a few projects, such as Malay Text Normalizer for Sentiment Analysis with an industry, and Implicit and Explicit Aspect Extraction for Sentiment Analysis under the Research University Grant. For the past five years, she has published more than 15 articles on NLP and AI in various WoS-indexed journals. Her research interests include natural language processing (NLP), where the main focus is on developing efficient algorithms to process texts and to make their information accessible to computer applications, mainly on text normalization and sentiment analysis. She serves as a reviewer for various journals.

AZNUL QALID MD SABRI (Senior Member, IEEE) received the Erasmus Mundus Masters degree in vision and robotics (ViBot) a joint master's degree from the University of Burgundy, France; University of Girona, Spain; and Heriot-Watt University, Edinburgh, U.K., for which he performed a research internship program at the Commonwealth Scientific Research Organization (CSIRO), Brisbane, Australia, with a focus on medical imaging. His Ph.D. thesis (trés honorable) focused on the topic of ''Human Action Recognition,'' under a program jointly offered by a well-known research institution in France, Mines de Douai (a research lab), and the University of Picardie Jules Verne, Amiens, France. He is currently an Associate Professor with the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology (FCSIT), University of Malaya, Malaysia.

IZZAT ALSMADI (Senior Member, IEEE) received the master's and Ph.D. degrees in software engineering from North Dakota State University, in 2006 and 2008, respectively. He is currently an Associate Professor with the Department of Computing and Cyber Security, Texas A&M University-San Antonio. He has more than 100 conferences and journals publications. He is the Lead Author and an Editor of several books, including *The NICE Cyber Security Framework: Cyber Security Intelligence and Analytics* (Springer, 2019), *Practical Information Security: A Competency-Based Education Course* (Springer, 2018), and *Information Fusion for Cyber-Security Analytics–Studies in Computational Intelligence* (Springer, 2016). His research interests include cyber intelligence, cyber security, software security, machine learning, natural language processing, software testing, social networks, and software defined networking.



DANAH OMARY received the master's degree in electrical engineering from the University of North Texas, in 2022, where she is currently pursuing the Ph.D. degree in electrical engineering. Her research interests include mixed reality, VLSI, hardware and software codesign, wireless communications, reconfigurable computing, machine learning, and computer vision.