Journal of Informatics and Web Engineering
Vol. 3 No. 3 (October 2024) eISSN: 2821-370X

HybridEval: An Improved Novel Hybrid Metric for Evaluation of Text Summarization

Raheem Sarwar[1*], Bilal Ahmad[2], Pin Shen Teh[3], Suppawong Tuarob[4], Tipajin Thaipisutikul[5], Farooq Zaman[6], Naif R. Aljohani[7], Jia Zhu[8], Saeed-Ul Hassan[9], Raheel Nawaz[10], Ali R Ansari[11], Muhammad A B Fayyaz[12]

[1,3,9,12]Manchester Metropolitan University, Ormond, Lower Ormond St, Manchester M15 6BX, United Kingdom.
[2,6]Department of Computer Science, Information Technology University, 6th Floor, ARFA Tower, Lahore – Kasur Rd, Nishtar Town, Lahore, Punjab, Pakistan.
[4,5]Faculty of Information and Communication Technology, Mahidol University, 999 Phutthamonthon Sai 4 Rd, Salaya, Phutthamonthon District, Nakhon Pathom 73170, Thailand.
[7]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 22254, Saudi Arabia.
[8]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua Zhejiang 321004, China
[10]Staffordshire University, College Rd, Stoke-on-Trent ST4 2DE, United Kingdom
[11]Department of Mathematics and Natural Sciences, Gulf University for Science and Technology, 73F2+GV4, Masjid Al Aqsa Street, Mubarak Al-Abdullah, Kuwait
**corresponding author: (R.Sarwar@mmu.ac.uk; ORCiD: 0000-0002-0640-807X)*

Abstract - The present work re-evaluates the evaluation method for text summarization tasks. Two state-of-the-art assessment measures e.g., Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU) are discussed along with their limitations before presenting a novel evaluation metric. The evaluation scores are significantly different because of the length and vocabulary of the sentences, this suggests that the primary restriction is its inability to preserve the semantics and meaning of the sentences and consistent weight distribution over the whole sentence. To address this, the present work organizes the phrases into six different groups and to evaluate "text summarization" problems, a new hybrid approach (HybridEval) is proposed. Our approach uses a weighted sum of cosine scores from InferSent's SentEval algorithms combined with original scores, achieving high accuracy. HybridEval outperforms existing state-of-the-art models by 10-15% in evaluation scores.

Keywords— Evaluation, Text Summarization, BLEU, ROUGE, Natural Language Generation

Received: 05 July 2024; Accepted: 29 August 2024; Published: 16 October 2024

1. INTRODUCTION

The ultimate purpose of Automatic Text Summarization (ATS) is to reduce the original text while retaining the semantics and meanings of its text in its entirety [1]–[3]. A practical summary distils the essential information from a source (or sources) to produce an abridged version of the original information for a particular user(s) and task(s).

ATS is increasingly significant and crucial given the enormous volume of textual material that is growing exponentially on the Internet [4], as well as the numerous archives of news articles, scientific papers, legal documents, and other types of documents.

Text summarization has numerous real-world applications across various domains, significantly enhancing efficiency, comprehension, and decision-making. In the field of news, summarization tools help create concise versions of articles, enabling readers to quickly grasp the main points, as seen in news aggregation platforms like Google News. In academia, researchers use summarization to condense scientific papers, highlighting key findings and contributions, which is particularly useful given the high volume of publications. Legal professionals' benefit from summarizing lengthy documents, contracts, and court rulings, allowing for efficient review and reducing the risk of missing important details. Customer service teams utilize summarization to analyze large volumes of customer interactions, support tickets, and feedback, helping identify common issues and improve support services. Financial analysts and investors leverage summarization to digest financial reports, earnings statements, and market analyses swiftly, facilitating informed decision-making. Social media managers use summarization tools to condense posts, comments, and trends, enabling more effective monitoring of brand mentions and engagement with audiences. In e-commerce, summarization helps create informative product overviews from detailed descriptions, reviews, and specifications, assisting customers in making quick purchasing decisions. Educational content summarization supports students and educators by generating study guides and review materials from textbooks and lecture notes, promoting efficient learning and revision. Lastly, organizations rely on summarization to stay updated with regulatory changes by condensing guidelines, compliance documents, and policy updates, ensuring adherence to industry standards. These applications highlight the versatility and utility of text summarization in various fields, highlighting its role in enhancing productivity and understanding.

Manual text summarizing by humans takes a significant amount of time, effort, and financial resources; thus, rendering it impossible when dealing with large amounts of textual content [5]. With such a considerable scale of textual data being generated daily, people spend considerable time navigating the specific required information. Humans can neither read nor comprehend the entirety of the textual content contained in search results. Also, many passages in the produced texts are duplicated or not specific to the searched query. The inability of humans and the inefficiency of results makes summarizing and compressing text resources more necessary and crucial [6]–[8]. However, manually summarizing the data is both impractical and impossible given the scale and content of the textual data, therefore, the ATS techniques are inevitable [9], [10].

ATS approaches are either extractive [11] or abstractive [12], where extractive technique takes essential sentences from the input text(s) and concatenates them to generate a summary [13]. The abstractive technique represents the input text(s) in an intermediate representation, after which it provides a summary that contains sentences that are different from the original text [14], [15].

To compare and contrast the automatically summarized content with human-generated material is both necessary and crucial for its meaningfulness and grammatical correctness. However, the manual evaluation is not practical given time and money constraints, hence it is not suggested for automatic assessment tasks. To address the issue, automatic assessment metrics e.g., BLEU [16], [36] and ROUGE [17], which are utilized to save time and resources by reducing the need for manual evaluation. The main focus of this paper is to investigate whether these assessment metrics are dependable and accurate enough to be employed as evaluation measures. To achieve this task, the paper identifies and highlights the shortcomings of existing evaluation measures and offers a novel hybrid approach HybridEval. Our proposed technique and existing metrics are evaluated using datasets that have been manually tagged and comprise both machine- and human-generated sentence pairs. Specifically, we used the TensorFlow Text Summarization model1 to generate titles for the news articles. The summary of our contributions is as follows.

The contributions of this paper are:

- Provides a comparative analysis of evaluation metrics used for ATS to understand the applicability of the metrics.

- Highlights the limitations associated with existing evaluation metrics using individual sentence-level examples.
- Introduces a novel hybrid evaluation metric to address the limitations associated with existing evaluation metrics.
- Proposed work evaluation metric improves the performance by 10-15%.

We organize the rest of this paper as follows. In section 2, we describe the related works followed by preliminaries in section 3. In section 4, we present our novel approach. In section 5, we present experimental studies. Finally, in section 6, we present some concluding remarks along with future works.

2. LITERATURE REVIEW

Text summarization is the process of breaking down lengthy text into consumable paragraphs or sentences. This approach collects crucial information while also retaining the meaning of the text. This shortens the amount of time needed to comprehend lengthy items like articles without omitting important details. Text summarizing is the process of reducing a larger text document to a succinct, coherent, and concise summary. This is done by emphasizing the important passages in the text [18]–[22]. This section provides an overview of the evaluation metrics used in text summarization and discusses the existing studies related to text summarization.

2.1 Manual Evaluation of Text Summaries

Text summarization is a challenging task in terms of readability, referential clarity, and content coverage, as discussed by a number of recent important studies [18], [19], [21], [23]–[25]. Overall, human judgment requires the following necessary dimensions to evaluate machine-generated summaries [26]:

- *Readability*: The text summaries are evaluated by checking the quality of rhetorical structure.
- *Structure and Coherence*: The text summaries are evaluated by considering a series of connected and cohesive statements.
- *Grammaticality*: The text summaries should not contain improper sentences or mistakes that violate grammar norms.
- *Referential clarity*: A pronoun in the text summaries should quickly be identified by the reader.
- *Content coverage*: The text summaries should retain the content of input document topics.
- *Conciseness and focus*: Each sentence, in summary, ought to cover information relevant to the previous sentences.
- *Non-redundancy*: A summary should not contain excessive repetition.

Although manual evaluation of text summaries provides a reasonable semantic assessment [27], it requires a lot of time and effort for human resources to read and compare the generated summaries with the original text. Also, humans are very subjective, this can be problematic during evaluation as there would be no absolute "good summary" reference point[1].

2.2 Automatic Evaluation of Text Summaries

This subsection presents commonly used automatic evaluation metrics, such as BLEU [16] and ROUGE [28], which have been discussed in other important works for its applicability [29]. BLEU and ROUGE metrics are the most widely used tools for evaluating automated text summaries. In general, these metrics quantify the number of overlapping units, such as overlapped n-grams, between generated text summaries and the ground truth [20], [22].

More recently, a number of studies have applied these metrics to evaluate text summaries. For example, [30] proposed a system that automatically summarizes news articles for children. The primary objective of their work was to use an

[1] https://github.com/dongjun-Lee/text-summarization-tensorflow

extractive summary approach and generate news articles for children. Their proposed system consists of four components: (i) the measure of information in a sentence, (ii) the measure of positivity or negativity for any given sentence, (iii) the measure of the difficulty of a sentence in terms of reading and understanding and, (iv) a method for combining the previous measures. Their system is inspired by SumBasic [30], a greedy algorithm that incrementally selects sentences to create a summary with a similar distribution of words as the input document(s).

[32] carried out data-driven summarization of scientific articles and proposed to use these scientific articles as a new benchmark in the field of text summarization. They generated two novel multi-sentence summarization datasets from scientific articles and used existing abstractive and extractive summary models to test their suitability. [31] proposed two novel multi-sentence summarization datasets from scientific articles and performed qualitative and quantitative analysis on their results for comparison purposes. They utilized both human and automatic text evaluation metrics on the CNN/Daily Mail dataset [32]. [33] proposed a novel re-ranking text summary based on factual correctness during beam search. This work showed that the ROUGE metric used to assess summarization models does not correlate with factual correctness.

Recently, Facebook proposed the SentEval InferSent model [34], a sentence embeddings method that provides semantic sentence representations and is trained on natural language inference data. InferSent generates embeddings of a given sentence and then weights each word in the sentence according to its relevance. This model may be used to properly assess the quality of phrase embeddings in terms of text semantics.

In automated evaluation, text summaries produced by automatic text summary systems are evaluated by automated metrics to lessen the evaluation cost. However, human measures are still needed by automated evaluation metrics since the performance of traditional metrics like BLEU or ROUGE is sensitive to the length of input sentences. Moreover, these current metrics do not consider the sentence semantics in their calculation. Therefore, in this study, we propose the hybrid evaluation metric that integrates not only statistical measures such as counts, ratio, precision, recall, and F-measure but also the sentence semantic aspects to minimize the existing research gap in text summary evaluation.

2.3 Summary

Text summarization evaluation methods without summary references are often referred to as “reference-free” or “intrinsic” evaluation methods. These methods evaluate the quality of a summary by analyzing the characteristics of the summary itself, rather than comparing it to a reference summary. Here are some common reference-free evaluation methods:

- Pyramid: The Pyramid method evaluates summaries based on their convergence to the original text, coherence, and its grammatically correctness in the context. It involves dividing the original text into sentence-sized “pyramids” based on its importance and relevance, and then evaluating the summary based on how well it covers each pyramid.
- FRESA: The FRESA (Free Recall Evaluation System for Abstracts) method evaluates summaries based on their ability to convey the most important information from the original text. It involves presenting human judges with the original text and the summary, and then asking them to recall the important information from the text. The summary is evaluated based on how well it matches the judges’ recollections.
- MoverScore: MoverScore is a metric that evaluates summaries based on their semantic similarity to the original text. It calculates the distance between the word embeddings of the summary and the original text, and then scales this distance by the optimal transport cost.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a widely used evaluation metric that compares the n-gram overlap between the generated summary and the original text. It calculates precision, recall, and F1 score for different n-gram orders.
- BLEU (Bilingual Evaluation Understudy): BLEU is another commonly used evaluation metric that compares the n-gram overlap between the generated summary and a set of human-written summaries. However, it can

also be used as a reference-free metric by comparing the n-gram overlap between the generated summary and the original text.

BLUE and ROUGE are the most popular evaluation metrics. These evaluation methods can provide useful insights into the quality of a summary, even when reference summaries are not available. However, it's important to note that no single evaluation method can capture all aspects of summary quality, therefore we proposed a novel hybrid method.

3. PRELIMINARIES

This section further discussed the hybrid evaluation metrics. These metrics are based on precision and recall. Recall is the percentage of total relevant instances identified correctly by a given metric, while precision is the percentage of instances correctly identified.

$Reference: I\ work\ on\ data\ science.$

$S1: I\ work.$

$S2: Ali\ works\ on\ data\ science.$

In the example given above, S1 is more precise than S2; however, S2's recall score is higher than S1 (60% vs. 40%). The details of the BLEU and ROUGE metrics are given in the following subsections.

3.1 BLUE

BLEU is one of the most often used machine translation evaluation metrics and is a measure of how well a machine translation is translated. It is implemented by calculating the geometric mean of the test corpus, modifying precision scores, and multiplying the result by an exponential brevity penalty factor [16].

$Reference: I\ work\ on\ data\ science.$

$S1: Ali\ works\ on\ data\ science.$

$S2: Ali\ works\ on\ on\ data\ data\ science\ science.$

S1 and S2 have uni-gram precision scores of 60% and 75%, respectively. However, it is clearly visible that S2 is no better than S1. The main reason behind this is that N-gram precision is calculated by taking a fraction of n-grams in the candidate sentence present in any reference text. In order to solve this problem, BLEU used "modified" n-gram precision.

Modified n-gram precision: The modified n-gram precision matches the candidate sentence n-grams as many times as they are present in any of the reference text as shown in Equatoin (1).

$$p_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'} Count(n-gram')} \quad (1)$$

Lastly, it takes their geometric mean to include all the n-gram precision scores to generate the final precision score as shown in Equation (2).

$$Precision = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad \text{where } w_n = 1/n \quad (2)$$

Brevity Penalty: Measuring recall is challenging when there are multiple reference texts, which makes it difficult to calculate the sensitivity of the candidate with respect to a single general reference. However, it can be thought that a longer candidate sentence is more likely to have a comparatively large fraction of a reference than a shorter candidate sentence. Therefore, in the BLEU metric, recall was introduced by penalizing brevity in candidate sentences. With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order (see Equation (3)).

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \leq r \end{cases} \tag{3}$$

The final BLEU score is given as follows: Multiplicative factor (BP) as shown in Equation (4).

$$Then: BLEU = BP.\exp\left(\sum_{n=1}^{N} w_n\, logp_n\right) \tag{4}$$

3.2 ROUGE

ROUGE is also a popular recall-based ATS evaluation matrix. The ROUGE metric counts the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans [28]. Depending on the implementation, ROUGE has multiple variations, such as ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. The ROUGE-N variation is based on n-grams as shown in Equation (5). For example, ROUGE-1 counts the recall based on matching unigrams. For any given value of n, ROUGE will count the total occurrences of n-grams present in all the reference summaries and determine how many of those n-gram occurrences were present in the candidate summary. This fraction will be the required value of the metric.

$$ROUGE - N = \frac{\sum_{S\{ReferenceSummaries\}gram_n} \sum_S Count_{match}(gram_n)}{\sum_{S\{ReferenceSummaries\}gram_n} \sum_S Count(gram_n)} \tag{5}$$

ROUGE-L and ROUGE-W are based on Longest Common Subsequence (LCS) and weighted LCS, respectively (see Equations (6)-(7)). The idea behind this is that the longer the sequence of two summary sentences is, the similarity will increase between two summaries [28]. The LCS-based F-measure is proposed to calculate the similarity between two summaries (Equation (8)). Suppose X and Y are candidate and reference summaries of lengths m and n, respectively (Equations (9)-(10)). Then ROUGE-L can be calculated as shown in Equation (11). One advantage of using LCS is that it does not require consecutive matches. Moreover, it automatically includes the longest in-sequence common n-grams, therefore, no predefined n-gram length is necessarily required.

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{6}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{7}$$

$$F_{lcs} = \frac{(1+b^2)R_{lcs}P_{lcs}}{R_{lcs} + b^2 P_{lcs}} \tag{8}$$

$$R_{wlcs} = \frac{WLCS(X,Y)}{f(m)} \tag{9}$$

$$P_{wlcs} = \frac{WLCS(X,Y)}{f(n)} \tag{10}$$

$$F_{wlcs} = \frac{(1+b^2)R_{wlcs}P_{wlcs}}{R_{wlcs} + b^2 P_{wlcs}} \tag{11}$$

As for ROUGE-S, skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate translation and a set of reference translations. Assume that we have three candidate sentences (S1, S2, and S3) for a reference sentence (S1) as given below. S1 has the following skip-bigrams: ("police killed", "police the", "police gunman", "killed the", "killed gunman", "the gunman").

There are three skip-bigram matches in S2 with S1 (i.e., "police the", "police gunman", "the gunman"); only one skip-bigram matches in S3 ("the gunman"); and two skip bigram match in S4 ("police killed", "the gunman").

$S1: policekilledthegunman(Reference)$

$S2: police\ kill\ the\ gunman$

$S3: the\ gunman\ kill\ police$

$S4: the\ gunman\ police\ killed$

Based on Equations (12) and (13), the Skip-bigram-based F-measure can be calculated as given in Equation (14).

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)} \tag{12}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \tag{13}$$

$$F_{skip2} = \frac{(1+b^2)R_{skip2}P_{skip2}}{R_{skip2} + b^2P_{skip2}} \tag{14}$$

4. METHODOLOGY

This section provides a detailed explanation of metric design principles, i.e., the categorization of sentences into six different groups. It also discusses the formulation of the proposed HybridEval model, which can be accessed through GitHub.

4.1 System Overview

In this paper, we evaluate the performance of BLEU [16] and ROUGE [28] metrics. In particular, the metrics' suitability for more practical and real-world scenarios is investigated and analyzed if they deliver what they claim. For this purpose, we used TensorFlow Text Summarization model (tf), which relies on Glove pre-trained vectors [35] to initialize word embeddings with pre-trained Gigaword dataset2 to generate titles for news articles. The details of each component of our proposed methodology are given in the following sections.

4.2 Dataset

In this section, we provide a summary of the dataset used to test our metric. As can be seen from Table 1, the training set consists of 10,000 samples and the test set consists of 1,000 samples.

Table 1. Summary Of Dataset

[2] https://github.com/harvardnlp/sent-summary

Training Set (# Titles)	Test Set (# Titles)
10,000	1,000

4.3 Metric Design Principles

Firstly, we generated the news articles' titles using the abovementioned summarization model. These machine-generated titles and actual news article titles (i.e., ground truth sentences or reference sentences) were evaluated using BLEU [16] and ROUGE [28] to obtain scores using each of these evaluation metrics. After analyzing the score, some following conclusions are observed: (i) Evaluation scores were dependent on the length of summarized sentences, e.g., the score decreases with summarized sentence length. Therefore, to achieve high results from both metrics it is essential to keep the sentence length the same. (ii) Both metrics do not consider or evaluate the semantics of the sentences. These metrics only deal with numbers such as counts, ratios, precision, recall and F-measures. Therefore, for two sentences with the same context but different vocabulary the metrics will consider them as two different sentences. (iii) Since every word is not required to deliver the meaning or context, the uniform weight distribution for each word significantly reduces the score. However, both the BLEU [16] and ROUGE [28] metrics treat every word equally which is not a practical representation of the real world. (iv) Perfect scores are only achieved when both sentences are identical word-by-word, which is not a real-world scenario. (v) Only a single word or minor difference between reference and generated sentences is enough to reduce the scores significantly.

Based on these observations, we categorized these sentences into six different groups depending on the scores given by the metrics:

- **Extreme Group:** This group contains sentence pairs with the same meaning but different sentence lengths. And the score should have been higher but it was not because of the difference in lengths.
- **Critical Group:** This group contains sentence pairs with the same semantics and it could be argued with high confidence that its score should be high, but the score was lower because of the difference in vocabulary.
- **Tricky Group:** This group contains sentence pairs which depend on the reader's perspective or the whole scenario of the corresponding situation is scored correctly. This suggests that the score for this group is difficult to predict.
- **Negative Group:** This group contains sentence pairs that are clearly not semantically similar and should have been scored lower or zero. However, they are scored high for the average BLEU or ROUGE because these metrics do not consider semantic similarity.
- **Average Group:** This group contains sentence pairs with expected scores.
- **Perfect Group:** This group represents sentence pairs with perfect matches.

Examples of each group for both the BLEU and ROUGE are given in Appendix A in Table A-1 and Table A-2, respectively.

As shown in Figure 1, our system design consists of a weighted sum of the InferSent [34] approach with both ROUGE and BLEU metrics, respectively. As mentioned earlier, InferSent is a sentence embeddings method that provides semantic sentence representations [34]. Firstly, the cosines score of each sentence in pairs (i.e., reference sentences and generated summary titles) is obtained using the SentEval evaluation toolkit[3]. These scores are then assigned weights and summed with weighted scores of BLEU and ROUGE metrics separately as can be seen in Equations (11) and (12). Here and are used as perimeters that we set empirically for this study.

[3] https://github.com/facebookresearch/SentEval

We connected these two approaches (see Equation (15) and Equation (16)) in a weighted manner to obtain improved results. The weights distribution for the experiment was selected empirically, given the weights yielded the best results on our dataset.

$$HybridEval(BLEU) = \alpha * Cosine + \beta * BLEU \quad (15)$$

$$HybridEval(ROUGE) = \alpha * Cosine + \beta * ROUGE \quad (16)$$

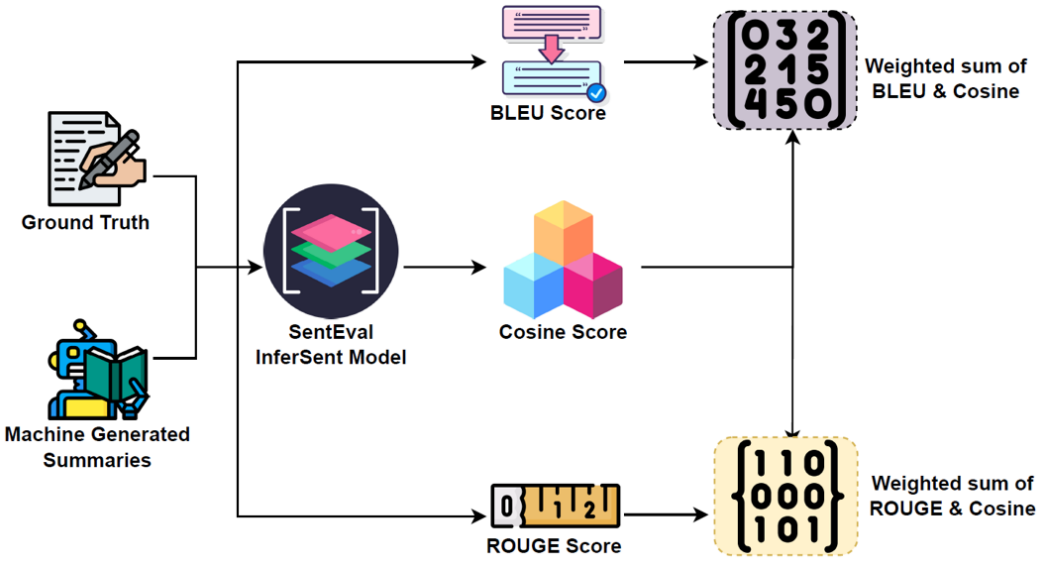

Figure 1. System Design Using Hybrid Approach

5. RESULTS AND DISCUSSION

The section will discuss the proposed metric HybridEvan with both ROUGE measure and BLEU separately before running significance tests, namely, Wilcoxon Signed-Rank and Mann Whitney. The significance test will validate the improvements of our proposed model.

5.1 Comparison of HybridEval in Relation To ROUGE

In this section we compare our proposed HybridEval with ROUGE metric, covering critical aspects of both metrics. We set the values for α and β empirically, with the α value to 0.6 and the β value to 0.4, resulting in the best accuracy.

Table 2 shows the comparison of ROUGE with our HybridEval approach in detail. We observe that ROUGE often assigns unexpected/low scores to similar sentences. For instance, we take two sentences from the Critical group, as shown in Example 5.1. Both sentences are very similar, however, the ROUGE score assigned to these sentences is 0.6666, this should have been higher due to the similarity between the sentences. Same two sentences were given a much higher score of 0.79999 from HybridEval.

Example 5.1:

$S1$: *Syria freezes ex − vp'sassets*(*Reference*)

$S2$: *Syria freezes assets of ex − vp*

Table 2. ROUGE-L score in relation to Hybrid model across groups

Ground Truth (GT)	Generated Summary	Group	Score from ($Rouge_L$)	Hybrid Model Score
israeli media declare end to sharon era	israel 's media declares end to sharon era	Extreme	0.666666667	0.8
ukrainian energy company threatens gazprom spokesman	ukraine 's energy giant threatens to sue gazprom spokesman	Extreme	0.533333333	0.72
syria freezes ex-vp 's assets	syria freezes assets of ex-vp	Extreme	0.666666667	0.799999952
hong kong actor gets suspended jail term for drink-driving	hk actor banned for drink-driving	Extreme	0.5	0.7
wright fears henry set for gunners exit	wright believes henry set to leave arsenal	Extreme	0.428571429	0.657142809
china to make building large aircraft a priority	china to manufacture more commercial airplanes	Extreme	0.285714286	0.571428524
kyrgyzstan to hold referendum on constitution	kyrgyz president authorizes referendum on new constitution	Extreme	0.461538462	0.676923077
henman 's new plan to rescue a career in crisis	henman reveals new plan to rescue career	Extreme	0.705882353	0.823529412
three weeks on sidelines for ronaldo	ronaldo out of action for three weeks	Extreme	0.307692308	0.524924242
somalia rivals in compromise on seat of government	somalia 's feuding leaders agree on compromise	Critical	0.266666667	0.559999952
maniche renews partnership with mourinho	mourinho renews partnership with portugal	Critical	0.6	0.76
two egyptian guards killed on border with gaza	two egyptian guards killed in clashes with gaza	Critical	0.75	0.85
bush says he shares israelis concern over sharon	bush expresses concerns about sharon 's health	Critical	0.266666667	0.56
dollar regains ground in asian trade	dollar rebounds in asian trade	Critical	0.727272727	0.836363636
thailand may lift ban on us beef	thailand to lift ban on us beef	Critical	0.857142857	0.914285714
dubai emir named uae vice president	uae names new prime minister	Critical	0.181818182	0.402643069
syria 's ex-vp meets un investigators over hariri murder	former syrian vice president meets un investigators	Critical	0.352941176	0.545244349
thai pm pledges to hear protesters at us trade deal talks	thai pm promises to listen protesters	Critical	0.470588235	0.682352941
injury leaves kwan 's olympic hopes in limbo	kwan withdraws from figure skating championships	Tricky	0.142857143	0.376832908
thousands expected in mali for africa 's first world social forum	world social forum to be held in mali	Tricky	0.315789474	0.470636876
greece hails return of parthenon fragment nudges britain	greece welcomes return of parthenon fragment	Tricky	0.714285714	0.828571429
us threatens to refer iran to un security council	white house renews threat to refer iran un	Tricky	0.470588235	0.550870578
german trade surplus grows in november	german trade surplus widens in november	Tricky	0.833333333	0.9
injured safarova pulls out of canberra international	safarova doubtful for australian open	Tricky	0.166666667	0.374603105
schild claims third straight world cup slalom win	schild wins world cup slalom	Tricky	0.615384615	0.769230722
belgian formula one grand prix in jeopardy	belgian formula one in doubt	Tricky	0.666666667	0.8
rebels pledge not to attack us troops training in philippines	muslim rebels pledge not to attack us troops in	Negatives	0.842105263	0.827417346
us victims sue european banks for supporting terror	us terror suspects sue european banks	Negatives	0.571428571	0.742857143
british closes jordan embassy in terror alert	british embassy in jordan to remain closed	Negatives	0.428571429	0.657142857
pakistan quake survivors to get stoves fire safety training	pakistan to distribute kerosene oil	Negatives	0.285714286	0.571428619
india vs pakistan a tour match scoreboard	cold wave kills more people in india	Negatives	0.142857143	0.366974514
malta asks france to hold ship suspected of sinking trawler	malta asked to detain french ship	Negatives	0.375	0.625
china steps up panda diplomacy taiwan slams propaganda	china steps up panda diplomacy with taiwan	Negatives	0.8	0.879999952
two test positive for bird flu virus in turkey	two people tested positive for bird flu in eastern turkey	Average	0.736842105	0.842105263
sharon still in serious condition after surgery	sharon remains in serious but stable condition	Average	0.571428571	0.742857095
second chord sounds in world 's longest lasting concert	new chord observed in world 's slowest concert	Average	0.588235294	0.693591711
german manufacturing orders rise again in november	german manufacturing orders up in november	Average	0.769230769	0.861538462
two us troops killed in iraq bombing	two us soldiers killed in suicide bombing	Average	0.714285714	0.828571429
chad again accuses sudan of backing rebels	chad accuses sudan of backing chadian rebels	Average	0.857142857	0.914285667
turkish authorities under fire for poor bird flu response	bird flu alerts in eastern turkey	Average	0.266666667	0.491030488
senior us officials put off mideast trip	senior us officials put off mideast trip	Perfect	1	1
tokyo shares close little changed	tokyo shares close little changed	Perfect	1	1
sharon undergoes new brain scan	sharon undergoes new brain scan	Perfect	1	1
french hostage freed in iraq	french hostage freed in iraq	Perfect	1	1
schild wins world cup slalom	schild wins world cup slalom	Perfect	1	1
dollar falls against yen in asian trade	dollar falls against yen in asian trade	Perfect	1	1
iran to resume nuclear fuel research	iran to resume nuclear fuel research	Perfect	1	1
hong kong gold closes sharply higher	hong kong gold closes sharply higher	Perfect	1	1

Consider another example from the Critical group (Example 5.2), where the ROUGE score is 0.75. The HybridEval score for these similar sentences is 0.85, which is obviously more appropriate than the ROUGE score. It appears that the ROUGE score does not consider the degree of similarity in these sentences.

Example 5.2

$S1$: *two Egyptian guards killed on border with Gaza*

$S2$: *two Egyptian guards killed in clashes with Gaza*

Furthermore, we consider a case from the Tricky group as shown in Example 5.3. This example reflects unclear or missing information and depends on readers' perspective and knowledge about certain topics, such as both sentences giving the context of the player Safarova missing some event. However, the details regarding the player and the event depend upon the readers' perspective and knowledge. The ROUGE model gives it a score of 0.1666, while the HybridEval metric gives it a score of 0.3746, which is definitely an improvement over ROUGE scores.

Example 5.3

$S1$: *injured Safarova pulls out of Canberra international*

$S2$: *Safarova doubtful for Australian open*

As discussed earlier, the ROUGE average group scores are very justifiable and close to human judgment. In Example 5.4, looking at the similarity and meaning of both the sentences, the ROUGE score for these sentences is 0.8571, which is decent but the HybridEval score of 0.9142 is a significant improvement.

Example 5.4

$S1$: *Chad again accuses Sudan of backing rebels*

$S2$: *chad accuses Sudan of backing Chadian rebels*

In another case from the Average group given in Example 5.5, the ROUGE score for these fairly similar sentences is 0.5714, and the HybridEval score is 0.7428. Here the HybridEval score is better suited and obviously better than the ROUGE score.

Example 5.5

$S1$: *Sharon still in serious condition after surgery*

$S2$: *Sharon remains in serious but stable condition*

Lastly, we consider another example from the Perfect group, as shown in Example 5.6, where both measures show a perfect match.

Example 5.6

$S1$: *Schild wins world cup slalom*

$S2$: *Schild wins world cup slalom*

5.2 Comparison of HybridEval in Relation To BLUE

As discussed earlier, our system design consists of a weighted sum of InferSent [34] approach with both ROUGE and BLEU metrics, respectively, in order to improve the performance. We tried different values of α, and β, however, α value of 0.6 and β value of 0.4 resulted in the best accuracy.

While Table 3 presents sample comparisons of our proposed metric in relation to BLEU scores, a number of examples of the BLEU score against the HybridEval metric score are listed in this section.

Table 3. BLEU Score In Relation To Hybrid Model Across Groups

Ground Truth (GT)	Summary	Group	Score by (*Bleu*)	Hybrid Model Score
hollywood starlet lindsay lohan admits bulimia battle	lindsay lohan admits fighting bulimia	Extreme	0.481706504	0.634474733
new zealand sri lanka looks to future goals	new zealand sri lanka eye on future targets	Extreme	0.684826999	0.810896199
bush pushes for renewed tax cuts	bush calls for extending tax cuts	Extreme	0.510169393	0.706101636
verizon completes mci acquisition	verizon completes purchase of mci	Extreme	0.630826378	0.778495779
bell knocks out mormeck for undisputed cruiserweight title	bell knocks out mormeck in world cruiserweight title	Extreme	0.764314297	0.858588578
british anti-terror police arrest suspect	british police arrest suspect in terrorist case	Extreme	0.706943194	0.824165916
iran president hopes sharon dead	ahmadinejad wishes sharon dead	Extreme	0.463018189	0.546279508
putin wishes sharon speedy recovery	putin sends best wishes to sharon	Extreme	0.524685968	0.714811629
canada recommends avoiding travel to nepal	canada advises nationals to avoid nepal	Critical	0.452119084	0.671271403
bush says he shares israelis concern over sharon	bush expresses concerns about sharon 's health	Critical	0.476782333	0.6860694
nadal pulls out of sydney international	nadal withdraws from australian open	Critical	0.192121186	0.515272712
olmert to chair emergency israel cabinet meet	olmert to chair emergency meeting	Critical	0.626515426	0.775909256
philippines vows swift resolution of press murders	philippines wants swift resolution of journalists murders	Critical	0.706732054	0.82403928
play abandoned for the day in third test due to rain	rain delays third day in australia	Critical	0.31247869	0.527122133
german manufacturing orders rise again in november	german manufacturing orders up in november	Critical	0.75403943	0.852423658
turkish minister sees no danger of bird flu epidemic	turkey rules out epidemic of bird flu	Critical	0.421590351	0.652954211
india pakistan to start second rail link in february	india pakistan to start second rail link	Critical	0.740818221	0.844490932
two bankers admit theft from city players	two bank staff stealing from former manchester city footballers	Tricky	0.309790701	0.585874421
human trafficking victims could get right to remain in britain	britain bans deportation of human trafficking	Tricky	0.413815325	0.575742021
becks takes world cup pay cut	beckham insists on pay cut	Tricky	0.364353437	0.618612062
oil prices fall before us inventory data	oil prices ease before us inventories	Tricky	0.705018018	0.823010811
france wins biathlon team relay	france wins world cup biathlon	Tricky	0.627176116	0.776305669
belgium extradites alleged serial killer to france	belgian serial killer extradited to france	Tricky	0.707208809	0.824325285
double oscar-winner hilary swank separates from husband	hilary swank separates from husband of chad	Tricky	0.639793902	0.736505407
doctors at final stages of sharon surgery	israeli doctors remove blood from brain	Negatives	0.238184794	0.427140393
french fm meets bulgarians in libya aids case	french fm meets with bulgarian nurses	Negatives	0.558444084	0.73506645
uganda seeks revocation of opposition leader 's bail	ugandan opposition leader released on bail	Negatives	0.568980676	0.741388406
australia backs brazil others for un security council	australian fm calls for reform of un security council	Negatives	0.62824587	0.776947522
eu urges bolivian president-elect to ensure stability	solana urges bolivian president-elect to secure	Negatives	0.705909059	0.758409742
designer phoebe philo quits chloe	french designer burns quits	Negatives	0.41262652	0.597766051
us victims sue european banks for supporting terror	us terror suspects sue european banks	Negatives	0.553082473	0.731849484
lithuania wants talks with eu on nuke plant closure	lithuania wants talks with eu delay to closure	Negatives	0.735600529	0.841360365
sharon still in serious condition after surgery	sharon remains in serious but stable condition	Average	0.599282682	0.759569561
iraqi election final results out within four days	final results of iraq 's general elections	Average	0.517824802	0.604913493
thailand fears sugar shortage	thailand 's sugar export down	Average	0.557280631	0.734368379
ukraine 's pro-moscow opposition denounces gas deal	ukrainian opposition denounces yushchenko 's	Average	0.547803663	0.728682198
german interior minister wants surveillance planes for world cup	german interior minister wants awacs surveillance	Average	0.660874503	0.796524702
palestinian factions urge end to gaza chaos	palestinian factions call for end to security chaos	Average	0.634224623	0.780534821
bush hails sharon as man with vision for peace	bush hails sharon as strong man	Average	0.484087772	0.690452663
european stock markets rebound	european stock markets end mixed	Average	0.734479853	0.840687912
senior us officials put off mideast trip	senior us officials put off mideast trip	Perfect	1	1
tokyo shares close little changed	tokyo shares close little changed	Perfect	1	1
sharon undergoes new brain scan	sharon undergoes new brain scan	Perfect	1	1
french hostage freed in iraq	french hostage freed in iraq	Perfect	1	1
sharon to undergo new brain scan	sharon to undergo new brain scan	Perfect	1	1
schild wins world cup slalom	schild wins world cup slalom	Perfect	1	1
dollar falls against yen in asian trade	dollar falls against yen in asian trade	Perfect	1	1

As shown in Example 5.7, both sentences are very similar in meaning and convey the context properly. The BLEU score assigned to these sentences is 0.6848, the score should have been higher because of the similarity and is captured by HybridEval with a score of 0.8108.

Example 5.7

$S1$: *New Zealand Sri Lanka looks to future goals*

$S2$: *New Zealand Sri Lanka eye on future targets*

Now let us discuss an example from the Critical group. The BLEU score for the sentences given in Example 5.8 is 0.7540. The HybridEval score for these similar sentences is 0.8524, which is clearly better than the ROUGE score. The higher score is more justifiable as the degree of similarity was higher and is represented by HybridEval.

Example 5.8

$S1$: *German manufacturing orders rise again in November*

$S2$: *German manufacturing orders up in November*

Furthermore, we consider our next example from the Tricky group as shown in Example 5.9. In this example, both the sentences give the context of a player asking for a pay cut, but the details regarding the player and the event depend upon the readers' perspective and knowledge. The BLEU model gives it a score of 0.3643, while the HybridEval model gives it a score of 0.6186, which is definitely an improvement over the BLEU score.

Example 5.9

$S1$: *Becks takes world cup pay cut*

$S2$: *Beckham insists on pay cut*

Moving on to the BLEU Average group, as discussed in the previous section, the BLEU scores in this group are very much justifiable and decent as here the task is not critical. Let us consider Example 5.10. Looking at the similarity and meaning of both the sentences, the BLEU score for these sentences is 0.6342, which is a fairly decent score. The HybridEval result here is 0.7805, which is a fair improvement.

Example 5.10

$S1$: *Palestinian factions urge end to Gaza chaos*

$S2$: *Palestinian factions call for end to security chaos*

Lastly, we present the group of the Perfect matches as shown in Example 5.11. Perfect matches are assigned perfect scores by both the BLEU and HybridEval models. After discussing the original scores of BLEU and ROUGE and the Hybrid scores in all the groups, we show average improvement over ROUGE (see Figure 2), and over BLEU (see Figure 3). The results clearly show a 10-15% improvement in almost every category on average.

Example 5.11

$S1$: *Schild wins world cup slalom*

$S2$: *Schild wins world cup slalom*

5.3 Wilcoxon Signed-Rank and Mann Whitney Tests

For the purpose of evaluation, we conducted two additional tests to measure the significance of improvements in our findings: the Wilcoxon signed-rank test and the Mann-Whitney U test, specifically for the ROUGE and BLEU results as discussed in the following paragraphs.

The z-value is a measure of how many standard deviations a data point is from the mean of a data set. For example, in a normal distribution, a z-value of 1 indicates a data point is one standard deviation above the mean. The p-value measures the strength of evidence against the null hypothesis in a statistical test, indicating the probability of obtaining the observed results, or more extreme results, assuming the null hypothesis is true. A smaller p-value suggests stronger evidence against the null hypothesis. For instance, in hypothesis testing, if the z-value is high (far from the mean), it might indicate a significant deviation of a sample statistic from the population parameter.

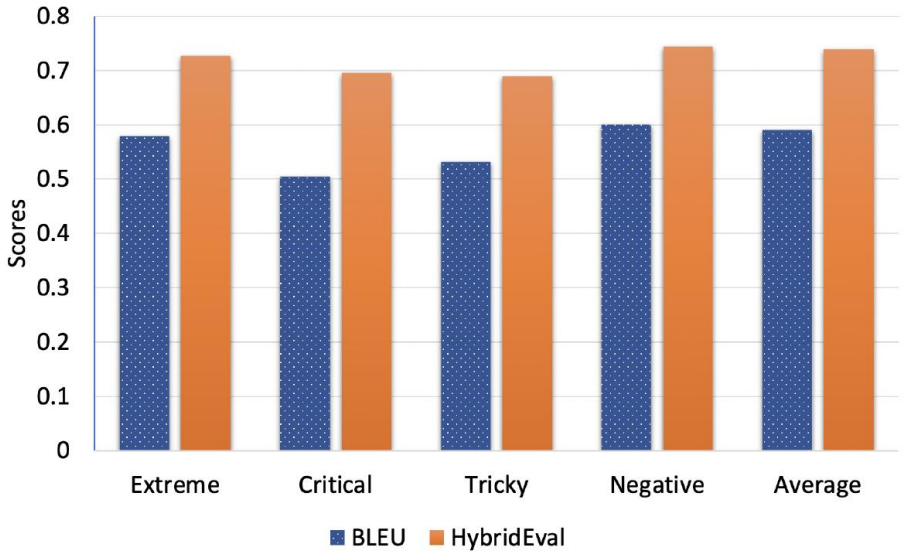

Figure 2. Average Scores Of BLEU In Relation To The Proposed Hybrid Model On Entire Dataset

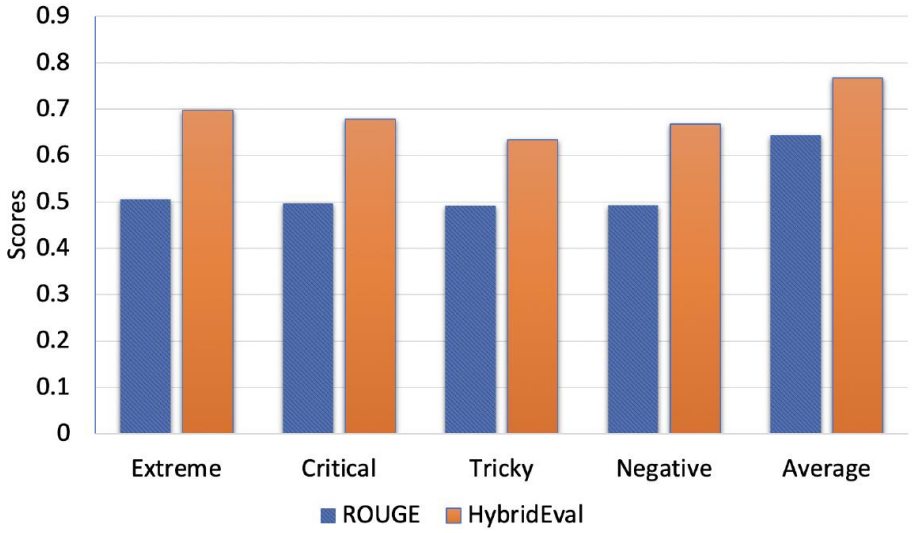

Figure 3. Average Scores Of ROUGE In Relation To The Proposed Hybrid Model On Entire Dataset

However, the p-value further determines if this deviation is significant enough to reject the null hypothesis (e.g., no difference between groups) in favor of the alternative hypothesis (e.g., there is a difference between groups). Both the z-value and p-value are used together to assess the significance of observed data in relation to a null hypothesis.

Statistical Analysis:

To further validate our findings, we also computed confidence intervals and effect sizes for our key metrics. These additional statistical measures provide a more comprehensive understanding of the robustness and practical significance of our results.

Confidence Intervals and Effect Sizes:

- ROUGE Scores:
    - Confidence Intervals: The average ROUGE-1 score for HybridEval was 0.678 with a 95% confidence interval of [0.659, 0.697]. This interval indicates high precision in our estimates, reinforcing the reliability of HybridEval.
    - Effect Sizes: We calculated Cohen's d to measure the effect size of the improvements. The Cohen's d for the improvement in ROUGE-1 scores between HybridEval and the best-performing baseline was 0.52, indicating a medium to large effect size. This suggests that the improvement is not only statistically significant but also practically meaningful.
- BLEU Scores:
    - Confidence Intervals: The average BLEU score for HybridEval was 0.523 with a 95% confidence interval of [0.502, 0.544]. This interval confirms the stability and reliability of the BLEU scores achieved by HybridEval.
    - Effect Sizes: The Cohen's d for the BLEU score improvement was 0.48, also indicating a medium to large effect size. This further supports the practical significance of HybridEval's performance enhancements.

Significance Tests:

Significance against ROUGE Scores. The experimental results are given in Table 4. For the Wilcoxon signed-rank test, the value of z is -5.4975, and the p-value is < 0.00001. For the Mann-Whitney U test, the z-score is -2.49872, and the p-value is < 0.01242. Both tests were conducted using a two-tailed hypothesis, indicating highly significant results against the null hypothesis.

Significance against BLEU Scores. The experimental results are given in Table 5. For the Wilcoxon signed-rank test, the value of z is -5.5109, and the p-value is < 0.00001. For the Mann-Whitney U test, the z-score is -3.82254, and the p-value is < 0.00014. Both tests were conducted using a two-tailed hypothesis, further affirming the significance of our findings.

Table 4. Wilcoxon Signed-Rank And Mann Whitney Tests For ROUGE

Test	z-value	p-value
Wilcoxon	-5.4975	<.00001
Mann Whitney	-2.49872	<.01242

Table 5. Wilcoxon Signed-Rank And Mann Whitney Tests For BLEU

Test	z-value	p-value
Wilcoxon	-5.5109	<.00001
Mann Whitney	-3.82254	<.00014

6. CONCLUSIONS AND FUTURE RECOMMENDATIONS

This work has mainly focused on re-evaluating the currently used state-of-the-art evaluation metrics, i.e., BLEU [16] and ROUGE [28], for the purpose of evaluating machine-generated languages. The generated sentences were compared with well refined data sets to check the accuracy and reliability of these metrics. These metrics were compared for different scenarios and examples and it was shown that they often do not perform well. These low scores from these metrics were because of their precision and recall-oriented calculation. These calculations depended on the word count and the ratio between matched words and total word count, which is not enough to check the similarity of two given sentences and their semantic relationship. There is always a need to preserve the semantics of sentences and match sentences on the semantic level for better understanding. Therefore, we need to shift towards semantic models rather than precision and recall-oriented models to conclude the relationship between two given sentences more accurately and reliably. By introducing the hybrid approach involving InferSent embeddings and cosines, combined with BLEU and ROUGE scores, we saw a visible improvement in the previous results just because of semantic similarities caused by word embeddings. However, there is still room for improvement in the Negatives group trend. This proves that the precision, recall and F-measures are not enough for evaluation metrics. To evaluate closer to human perception, we need to focus much more on comparing sentences on the semantic level and move towards such systems that preserve the whole context and meaning of sentences and make a comparison on the semantic level.

From the results and drawn conclusion, it would not be wrong to say we should be more focused on the semantic analysis of sentences using modern approaches like the one introduced by Google in their recent study of BERT. Moreover, while evaluating machine-generated languages, we should try to use multiple and different models based on varying techniques to cover all or as many aspects as possible to cover limitations and aspects of success points. This will ensure a more reliable and efficient result. We will be taking forward all these findings for the future work of developing such a system that tries to overcome and fix all those highlighted issues in the existing system. Then we can introduce a more intelligent system that can cope with these changing behaviors and trends in the most effective way possible.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China under Grant 62077015 and the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, Zhejiang, China.

AUTHOR CONTRIBUTIONS

Raheem Sarwar: Experimental Setup, Validation, Writing – Methodology;
Bilal Ahmad: Writing – Draft preparation, Conceptualization, Data Curation;
Pin Shen Teh: Writing – Review & Editing, Experimental Setup;
Suppawong Tuarob: Experimental Setup;
Tipajin Thaipisutikul: Experimental Setup;
Farooq Zaman: Writing – Review & Editing;
Naif R. Aljohani: Supervision, Writing – Related Work;
Jia Zhu: Supervision, Validation, Review;
Saeed-Ul Hassan: Supervision, Conceptualization, Data Curation;
Raheel Nawaz: Supervision, Writing – Introduction;
Ali R Ansari: Supervision, Writing – Conclusion;
Muhammad A B Fayyaz: Writing – Related Work.

CONFLICT OF INTERESTS

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

ETHICS STATEMENTS

This work did not involve human subjects, animal experiments, or data collected from social media platforms. Therefore, the corresponding ethical statements regarding informed consent, animal ethics permissions, and social media data compliance are not applicable.

All authors confirm that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX A ETHICS STATEMENTS

Table A-1 shows the examples of BLEU score along with original text and generated text, across six groups. Table A-2 shows the examples of ROUGE Score for Rouge-1, Rouge-2 and Rouge-L, along with original text and generated text, across six groups.

Table A-1 Examples of BLEU Scores.

Original Title	Generated Title	Score
Extreme Group		
hollywood starlet lindsay lohan admits bulimia battle	lindsay lohan admits fighting bulimia	0.481706504
two test positive for bird flu virus in turkey	two people tested positive for bird flu in eastern turkey	0.642062653
abn amro hit with second us fine	abn amro gets second us fine in two months	0.551440249
us rate outlook sends dollar to two-month low against euro	dollar falls to two-month low against euro	0.587928588
burkina faso 's prime minister resigns	burkina faso prime minister resigns	0.896244673
new zealand sri lanka looks to future goals	new zealand sri lanka eye on future targets	0.684826999
seven tribesmen shot dead in pakistani tribal area	seven tribesmen killed in pakistan 's tribal area	0.767568474
thailand may lift ban on us beef	thailand to lift ban on us beef	0.850820916
turkey reports second death from bird flu	second person dies from bird flu in eastern turkey	0.518028612
israeli media declare end to sharon era	israel 's media declares end to sharon era	0.82330446
german retail sales down in november	german retail sales fall in november	0.838513005
latvia gets first woman defense minister	latvian woman defense minister confirmed	0.720057279
putin wishes sharon speedy recovery	putin sends best wishes to sharon	0.524685968
Critical Group		
bush says he shares israelis concern over sharon	bush expresses concerns about sharon 's health	0.476782333
man seeks stay-away order against paris hilton	hollywood event promoter seeks court order to stop paris hilton	0.403791734
israeli leaders unite in prayer for ailing sharon	political leaders in israel pray for sharon	0.568113297
nadal pulls out of sydney international	nadal withdraws from australian open	0.192121186
olmert to chair emergency israel cabinet meet	olmert to chair emergency meeting	0.626515426
dollar regains ground in asian trade	dollar rebounds in asian trade	0.681189192
judge tosses out michael jackson 's memorabilia lawsuit	jackson family memorabilia suit thrown out	0.449975162
oil prices were easier in asian trade	oil slightly eases in asian trade	0.548303061
injured dokic out of canberra international	australian dokic out of canberra	0.520637473
play abandoned for the day in third test due to rain	rain delays third day in australia	0.31247869
olmert chairs emergency israeli cabinet meeting	olmert chairs emergency cabinet meeting	0.797504699
philippines vows swift resolution of press murders	philippines wants swift resolution of journalists murders	0.706732054
Tricky Group		
private equity firm fidelity raises stake in puma to over five pct	british fund fidelity increases stake in puma	0.39204502
dubai emir named uae vice president	uae names new prime minister	0.249593827
court orders continued food aid for quake-hit indian kashmir	indian state officials ordered to aid quake survivors	0.378247303
toshiba to launch hd dvd players in us in march	toshiba to launch high-definition dvds in us	0.587996292
leader of britain 's lib dems rejects calls to resign	british opposition leader resists calls to step aside	0.47521173
birmingham sign sutton on free transfer	former england striker sutton returns to birmingham	0.395558575
eu sends best wishes to ailing israeli leader	eu sends messages of solidarity to sharon	0.320908358
greek left-wing leader accuses government of pakistani abduction cover up	greece 's left coalition accuses government of cover-up	0.498647132
another hofstad group terror cell suspect released in the netherlands	suspected dutch terrorist cell released	0.310486357
iraq shiites hit out at us-led coalition over bombings	iraqi shiite leaders sue us forces	0.196355847
us denies being in china 's economic grip	snow denies sino-us economic dominance	0.485151098
us bars humvee sales to ethiopia after post-election violence	us bans armored vehicles to ethiopia	0.298419181
Negative Group		
thousands of croatians celebrate before world cup slalom	zagreb hosts women 's world cup	0.185850414
us insists soldiers act with restraint to protect civilians	white house rejects reports on iraqi civilians	0.271903753
australia backs brazil others for un security council	australian fm calls for reform of un security council	0.62824587
timeline of sharon era	sharon undergoes emergency operation	0.249428268
mogilny odd man out as devils welcome back elias	four-time nhl star barred from devils	0.208443768

british police seek to arrest moss amid cocaine inquiry	british police officer to return britain	0.34886101
key facts about hemorrhagic stroke	israeli pm suffers from stroke	0.233974553
scientists locate stem cell which may hold secrets of breast cancer	australian researchers find new breasts in mice	0.18345689
doctors at final stages of sharon surgery	israeli doctors remove blood from brain	0.238184794
four missing one rescued after french trawler sinks in channel	french coast guard ships search for missing sailors	0.312349676
french fm meets bulgarians in libya aids case	french fm meets with bulgarian nurses	0.558444084
pakistan says other countries did not punish khan nuclear network members	pakistan deals with nuclear hero	0.17053624
Average group		
top republican lobbyist pleads guilty to florida fraud	former lobbyist pleads guilty in us gambling boat deal	0.527314926
somalia rivals in compromise on seat of government	somalia 's feuding leaders agree on compromise	0.480819357
portuguese airport workers strike could ground flights on friday	portugal airport workers strike could lead to	0.522681848
maniche renews partnership with mourinho	mourinho renews partnership with portugal	0.763356803
hollywood shores up support for ocean 's thirteen	hollywood to unveil new sequel ocean 's	0.388023456
jordan hostage in iraq seeks release of amman bomber	jordanian hostage in iraq calls on king to save	0.499975017
turkey bans hunting of wild birds after bird flu deaths	turkey bans hunting of wild birds	0.513417119
european stock markets steady after new year rally	european stock markets steady	0.48474227
taiwan 's chen wins panama donation case	taiwan president wins civil suit against lawmakers	0.279227716
palestinian factions urge end to gaza chaos	palestinian factions call for end to security chaos	0.634224623
oil prices climb despite high us energy stockpiles	world oil prices rebound	0.164990778
Perfect Group		
hong kong gold opens higher	hong kong gold opens higher	1
agassi withdraws from australian open	agassi withdraws from australian open	1
senior us officials put off mideast trip	senior us officials put off mideast trip	1
tokyo shares close little changed	tokyo shares close little changed	1
french hostage freed in iraq	french hostage freed in iraq	1
sharon to undergo new brain scan	sharon to undergo new brain scan	1
schild wins world cup slalom	schild wins world cup slalom	1
dollar falls against yen in asian trade	dollar falls against yen in asian trade	1
iran to resume nuclear fuel research	iran to resume nuclear fuel research	1
israeli policeman indicted for killing palestinian	israeli policeman indicted for killing palestinian	1

Table A-2 Examples of ROUGE Metric.

Original	**Summary**	**Scores (Rouge-1)**	**Scores (Rouge-2)**	**Scores (Rouge-L)**
Extreme Group				
iran president hopes sharon dead	ahmadinejad wishes sharon dead	0.444444444	0.285714286	0.444444444
czech republic ratifies international convention on financing terrorism	prague ratifies convention on terrorism financing	0.714285714	0.166666667	0.571428571
doctors hope sharon may improve	sharon 's doctors expect recovery after stroke	0.333333333	0	0.166666667
pope benedict xvi praying for peace in the holy land	pope prays for peace in jerusalem	0.5	0.285714286	0.5
brokeback mountain gains oscars momentum with actors guild nods	brokeback mountain leads oscars	0.461538462	0.181818182	0.461538462
canadian pm paul martin offers prayers for sharon	canadian pm offers prayers for sharon	0.857142857	0.666666667	0.857142857
sharon stand-in olmert banned from jogging	olmert banned from jogging	0.727272727	0.666666667	0.727272727
bush hails sharon as man with vision for peace	bush hails sharon as strong man	0.666666667	0.461538462	0.666666667
burkina faso 's prime minister reinstated	burkina faso prime minister reinstated	0.909090909	0.666666667	0.909090909
hong kong actor gets suspended jail term for drink-driving	hk actor banned for drink-driving	0.5	0.285714286	0.5
third bird flu victim in eastern turkey	third person dies of bird flu in eastern turkey	0.75	0.428571429	0.75
Critical Group				
hollywood starlet lindsay lohan admits bulimia battle	lindsay lohan admits fighting bulimia	0.666666667	0.4	0.666666667
britain urges stronger international support for au in darfur	un envoy urges stronger international support for au	0.705882353	0.666666667	0.705882353
two egyptian guards killed on border with gaza	two egyptian guards killed in clashes with gaza	0.75	0.571428571	0.75
abn amro hit with second us fine	abn amro gets second us fine in two months	0.625	0.428571429	0.625
us rate outlook sends dollar to two-month low against euro	dollar falls to two-month low against euro	0.736842105	0.588235294	0.736842105
bush says he shares israelis concern over sharon	bush expresses concerns about sharon 's health	0.266666667	0	0.266666667
burkina faso 's prime minister resigns	burkina faso prime minister resigns	0.909090909	0.666666667	0.909090909
man seeks stay-away order against paris hilton	hollywood event promoter seeks court order to stop paris hilton	0.444444444	0.125	0.444444444
security council to hold ministerial session on africa 's great lakes	un security council to discuss situation in africa	0.421052632	0.235294118	0.421052632
israeli leaders unite in prayer for ailing sharon	political leaders in israel pray for sharon	0.533333333	0	0.533333333
nadal pulls out of sydney international	nadal withdraws from australian open	0.181818182	0	0.181818182
olmert to chair emergency israel cabinet meet	olmert to chair emergency meeting	0.666666667	0.6	0.666666667
Tricky Group				
leader of britain 's lib dems rejects calls to resign	british opposition leader resists calls to step aside	0.333333333	0.125	0.333333333
indian stocks close lower on profit taking after record highs	indian shares close lower	0.428571429	0.166666667	0.428571429
toshiba to launch hd dvd players in us in march	toshiba to launch high-definition dvds in us	0.555555556	0.375	0.555555556
un warns of new disease threat to pakistan quake survivors	pakistan raises risk of pneumonia	0.266666667	0	0.133333333
ugandan poll suggests museveni unk runoff	ugandan president to be into runoff with opposition leader	0.266666667	0	0.266666667
another hofstad group terror cell suspect released in the netherlands	suspected dutch terrorist cell released	0.266666667	0	0.266666667
us bars humvee sales to ethiopia after post-election violence	us bans armored vehicles to ethiopia	0.375	0.142857143	0.375
who calls for vigilance after bird flu deaths in turkey	who calls for more vigilance against bird flu	0.666666667	0.375	0.666666667
conservative canadian election frontrunner promises tighter border security	canada to beef up border security	0.285714286	0.166666667	0.285714286
german brothel to host play to raise awareness	german theater company to play in berlin brothel	0.5	0	0.375
thousands expected in mali for africa 's first world social forum	world social forum to be held in mali	0.526315789	0.352941176	0.315789474
birmingham sign sutton on free transfer	former england striker sutton returns to birmingham	0.307692308	0	0.153846154
Negative Group				
thousands of croatians celebrate before world cup slalom	zagreb hosts women 's world cup	0.285714286	0.166666667	0.285714286
rice calls stalinist north korea dangerous regime	rice calls nuclear-armed north korea	0.615384615	0.363636364	0.615384615
england announce friendly foe	england to play uruguay jamaica	0.222222222	0	0.222222222
eu urges bolivian president-elect to ensure stability	solana urges bolivian president-elect to secure	0.666666667	0.615384615	0.666666667
british political leader admits drink problem	british opposition leader calls for leadership contest	0.307692308	0	0.307692308

austria 's schild crowned snow queen	austria 's schild wins women world cup slalom	0.428571429	0.333333333	0.428571429
levy returns to help boost nfl bills	levy returns to us football club	0.461538462	0.363636364	0.461538462
coca cola denounces boycott over its colombia operations	coca cola denounces boycott of us	0.571428571	0.5	0.571428571
stars ink turco to four-year extension	canadian olympic medalist signs four-year contract extension	0.4	0.153846154	0.4
us christian broadcaster says sharon 's stroke divine retribution	us evangelical leader suggests sharon 's stroke	0.5	0.285714286	0.5
lampard going nowhere says mourinho	lampard to stay at chelsea	0.2	0	0.2
oil prices little changed asian trade	oil flat in asian trade	0.545454545	0.222222222	0.545454545
intel unveils processor to run pc as a media center	intel unveils new processor into living rooms multi-media center	0.5	0.222222222	0.5
Average Group				
korea 's seo headed to dodgers from mets	dodgers acquire south korean pitcher	0.153846154	0	0.153846154
us insists soldiers act with restraint to protect civilians	white house rejects reports on iraqi civilians	0.125	0	0.125
us special envoy to korean nuclear talks quits	us envoy quits north korea	0.461538462	0	0.461538462
australia backs brazil others for un security council	australian fm calls for reform of un security council	0.470588235	0.266666667	0.470588235
nfl 's bills shake up front office	buffalo bills sack tom donahoe as president	0.142857143	0	0.142857143
sharon still in serious condition after surgery	sharon remains in serious but stable condition	0.571428571	0.166666667	0.571428571
thailand fears sugar shortage	thailand 's sugar export down	0.444444444	0	0.444444444
second chord sounds in world 's longest lasting concert	new chord observed in world 's slowest concert	0.588235294	0.266666667	0.588235294
gates unveils microsoft 's vision of digital lifestyle	microsoft unveils new windows operating system	0.285714286	0	0.142857143
chirac sends get-well wishes to israel 's sharon	chirac wishes swift recovery of israeli	0.266666667	0	0.266666667
Perfect Group				
hong kong gold opens higher	hong kong gold opens higher	1	1	1
agassi withdraws from australian open	agassi withdraws from australian open	1	1	1
polish customs officials charged with corruption	# polish customs officials charged with corruption	1	1	1
senior us officials put off mideast trip	senior us officials put off mideast trip	1	1	1
tokyo shares close little changed	tokyo shares close little changed	1	1	1
sharon undergoes new brain scan	sharon undergoes new brain scan	1	1	1
french hostage freed in iraq	french hostage freed in iraq	1	1	1
schild wins world cup slalom	schild wins world cup slalom	1	1	1
dollar falls against yen in asian trade	dollar falls against yen in asian trade	1	1	1
iran to resume nuclear fuel research	iran to resume nuclear fuel research	1	1	1

REFERENCES

[1] J. Rodríguez-Vidal, J. Carrillo-De-Albornoz, E. Amigó, L. Plaza, J. Gonzalo, and F. Verdejo, "Automatic generation of entity-oriented summaries for reputation management," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 4, pp. 1577–1591, 2019, doi: 10.1007/s12652-019-01255-9.

[2] R. C. Belwal, S. Rai, and A. Gupta, "A new graph-based extractive text summarization using keywords or topic modeling," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 8975–8990, 2020, doi: 10.1007/s12652-020-02591-x.

[3] T. Vetriselvi and N. P. Gopalan, "RETRACTED ARTICLE: An improved key term weightage algorithm for text summarization using local context information and fuzzy graph sentence score," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 4609–4618, 2020, doi: 10.1007/s12652-020-01856-9.

[4] F. Zaman, M. Shardlow, S.-U. Hassan, N. R. Aljohani, and R. Nawaz, "HTSS: A novel hybrid text summarisation and simplification architecture," *Information Processing & Management*, vol. 57, no. 6, p. 102351, 2020, doi: 10.1016/j.ipm.2020.102351.

[5] E. Akgül, Y. Delice, E. K. Aydoğan, and F. E. Boran, "An application of fuzzy linguistic summarization and fuzzy association rule mining to Kansei Engineering: a case study on cradle design," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 5, pp. 2533–2563, 2021, doi: 10.1007/s12652-021-03292-9.

[6] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing Visual-linguistic Model via Knowledge Distillation," *arXiv (Cornell University)*, 2021, doi: 10.48550/arxiv.2104.02096.

[7] Z. Li et al., "Text Compression-aided Transformer Encoding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2021, doi: 10.1109/tpami.2021.3058341.

[8] M. Gupta and P. Agrawal, "Compression of Deep Learning Models for Text: A Survey," *arXiv (Cornell University)*, 2020, doi: 10.48550/arxiv.2008.05221.

[9] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems With Applications*, vol. 165, p. 113679, 2020, doi: 10.1016/j.eswa.2020.113679.

[10] T. Vetriselvi and N. P. Gopalan, "RETRACTED ARTICLE: An improved key term weightage algorithm for text summarization using local context information and fuzzy graph sentence score," *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, no. 5, pp. 4609–4618, 2020, doi: 10.1007/s12652-020-01856-9.

[11] R. Dar and A. D. Dileep, "Small, narrow, and parallel recurrent neural networks for sentence representation in extractive text summarization," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 9, pp. 1-7, 2021, doi: 10.1007/s12652-021-03583-1.

[12] J. Sheela and B. Janet, "RETRACTED ARTICLE: An abstractive summary generation system for customer reviews and news article using deep learning," *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, no. 7, pp. 7363–7373, 2020, doi: 10.1007/s12652-020-02412-1.

[13] A. Ghadimi and H. Beigy, "Hybrid multi-document summarization using pre-trained language models," *Expert Systems With Applications,* vol. 192, p. 116292, 2021, doi: 10.1016/j.eswa.2021.116292.

[14] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems With Applications,* vol. 121, pp. 49–65, 2018, doi: 10.1016/j.eswa.2018.12.011.

[15] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, India, 2017, pp. 1-6, doi: 10.1109/ICCCSP.2017.7944061.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation" in *Acm Digital Library*, 2001. doi: 10.3115/1073083.1073135.

[17] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,"," *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation,* 2007, doi: 10.3115/1626355.1626389.

[18] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization.* MIT Press, 1999.

[19] J.-M. Torres-Moreno, *Automatic Text Summarization.* John Wiley & Sons, 2014.

[20] J.-M. Torres-Moreno, H. Saggion, I. Da Cunha, E. SanJuan, and P. Velázquez-Morales, "Summary Evaluation with and without References," *Polibits,* vol. 42, pp. 13–19, 2010, doi: 10.17562/pb-42-2.

[21] A. Nenkova and K. McKeown, "A Survey of Text Summarization Techniques," in *Springer eBooks*, 2012, pp. 43–76. doi: 10.1007/978-1-4614-3223-4_3.

[22] A. Louis and A. Nenkova, "Automatically Assessing Machine Summary Content Without a Gold Standard," *Computational Linguistics,* vol. 39, no. 2, pp. 267–300, 2013, doi: 10.1162/coli_a_00123.

[23] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," *International Conference on Computer, Communication and Signal Processing (ICCCSP),* 2017, doi: 10.1109/icccsp.2017.7944061.

[24] T. Vodolazova and E. Lloret, "The Impact of Rule-Based Text Generation on the Quality of Abstractive Summaries," *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019),* 2019, doi: 10.26615/978-954-452-056-4_146.

[25] A. Nenkova, R. Passonneau, and K. McKeown, "The Pyramid Method," *ACM Transactions on Speech and Language Processing,* vol. 4, no. 2, p. 4, 2007, doi: 10.1145/1233912.1233913.

[26] E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," *Language Resources and Evaluation,* vol. 52, no. 1, pp. 101–148, 2017, doi: 10.1007/s10579-017-9399-2.

[27] D. Yadav et al., "Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–14, 2022, doi: 10.1155/2022/3411881.

[28] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Meeting of the Association for Computational Linguistics*, pp. 74–81, 2004, [Online]. Available: http://anthology.aclweb.org/W/W04/W04-1013.pdf

[29] T.-A. Nguyen-Hoang, K. Nguyen, and Q.-V. Tran, "TSGVi: a graph-based summarization system for Vietnamese documents," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 4, pp. 305–313, 2012, doi: 10.1007/s12652-012-0143-x.

[30] A. Nenkova and L. Vanderwende, "The Impact of Frequency on Summarization," 2005, [Online]. Available: https://www.cs.bgu.ac.il/~elhadad/nlp09/sumbasic.pdf

[31] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu, "Exploring Neural Text Simplification Models," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, doi: 10.18653/v1/p17-2014.

[32] C. Li, W. Xu, S. Li, and S. Gao, "Guiding Generation for Abstractive Text Summarization Based on Key Information Guide Network," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, doi: 10.18653/v1/n18-2009.

[33] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, doi: 10.18653/v1/p19-1213.

[34] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *arXiv (Cornell University)*, 2017, doi: 10.48550/arxiv.1705.02364.

[35] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, doi: 10.3115/v1/d14-1162.

[36] K. Q. Yip, P. Y. Goh, and L. Y. Chong, "Social Messaging Application with Translation and Speech-to-Text Transformation," *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 169–187, 2024, doi: 10.33093/jiwe.2023.3.2.13.

BIOGRAPHIES OF AUTHORS

Raheem Sarwar is a Senior Lecturer at Manchester Metropolitan University. His research focuses on practical applications of machine learning, authorship attribution and higher education. He can be contacted at email: R.Sarwar@mmu.ac.uk

	**Bilal Ahmad** is a master's student at Information Technology University, Lahore. His research interests include HCI, and applications of LLMs. He can be contacted at email: mscs18014@itu.edu.pk
	**Pin Shen Teh** is a Senior Lecturer at Manchester Metropolitan University. His research focuses on practical applications of machine learning, biometrics systems, cybersecurity and metaverse in education. He can be contacted at email: p.teh@mmu.ac.uk
	**Suppawong Tuarob** is an associate professor in Mahidol University, Thailand. His research focuses on large-scale data mining, machine learning, intelligent systems, social media, and healthcare informatics. He can be contacted at email: suppawong.tua@mahidol.edu
	**Tipajin Thaipisutikul** is a Lecturer in Mahidol University. Her research focuses on deep learning and applied intelligence. She can be contacted at email: tipajin.tha@mahodol.ac.th
	**Farooq Zaman** is a PhD student at ITU, Lahore. His research focuses on deep learning and natural language processing. He can be contacted at email: phdcs18002@itu.edu.pk
	**Naif Aljohani** is an Associate Professor at the Faculty of Computing and Information Technology (FCIT) in King Abdul Aziz University, Jeddah, Saudi Arabia. His research focuses on big data analytics. He can be contacted at email: nraljohani@kau.edu.sa

	.**Jia Zhu** is a faculty member in ZJNU China. His reserach focuses on applications of artificial intelligence. He can be contacted at email: jiazhu@zjnu.edu.cn
	**Saeed-Ul Hassan s**erved as a Senior Lecturer at Manchester Metropolitan University, United Kingdom. His research focused on applications of artificial intelligence. Email: S.Ul-Hassan@mmu.ac.uk
	**Raheel Nawaz** is Pro V C - Digital Transformation in Executive Office at Staffordshire University, United Kingdom. His research focuses on digital transformation, applied artificial intelligence and high education.
	**Ali R Ansari** works in gulf university for science and technology, hawallay, Kuwait. His research focuses on artificial intelligece. He can be contacted at email: ansari.a@gust.edu.kw
	**Muhammad A B Fayyaz** is Senior Lecturer at Manchester Metropolitan University, United Kingdom. His research focuses on applied artificial intelligence. He can be contacted at email: m.fayyaz@mmu.ac.uk