# Enhancing Accuracy in Indian Language Summarization

Rishika Jha[1,*,†], Rakshit Ahuja[1,†], Sainik Kumar Mahata[1,†], Monalisa Dey[1,†] and Dipankar Das[2,†]

[1]Institute of Engineering & Management, Kolkata, University of Engineering and Management, Kolkata, India
[2]Jadavpur University, Kolkata, India

**Abstract**

Text summarization is one of the most explored domain within Natural Language Processing and has witnessed significant progress. The ILSUM shared task concentrates on various languages such as such as English, Hindi, Gujarati, and Bengali for their summarization task. Furthermore, it addresses the critical issue of factual inaccuracies in machine-generated text summaries in Hindi and Gujarati languages.The proposed research focuses on the usage of pre-trained sequence-to-sequence model specifically for English language and a machine learning algorithm to detect the occurrence of factual distortions.

**Keywords**

Text-Summarization, Sequence-to-Sequence models, machine learning algorithms, Factual Incorrectness

## 1. Introduction

Summarizing content in Indian languages would be very important for condensing long texts without losing critical information. Since India encompasses a huge number of linguistic diversities, summarization systems which can successfully improve the accessibility of knowledge need to be developed. The T5-small model, a more compact version of the T5 (Text-to-Text Transfer Transformer), would be particularly useful because of its versatility and its operational efficiency.

For this purpose, we have used the data made available by ILSUM 2024, providing an extensive corpus containing different textual data across multiple Indian languages, such as Hindi, Tamil, Telugu, Malayalam, and Bengali. With training on specifically English dataset, the T5-small model provides coherent and contextually accurate summaries across different domains. The applications for this model involve summarizing news articles, educational resources, and legal documents. Using models like T5-small along with the ILSUM 2024 dataset will begin to bridge that information gap between the different linguistic communities in India, opening up access to more knowledge.

In the current era of information dissemination, which is generally made easy by online platforms, accuracy and integrity in its content are paramount. Machine-generated summaries, condensing complex texts into brief and digestible form through efficiency and scalability are beneficial. However, the potential for these systems to have factual inaccuracies is part of a significant challenge. This paper attempts to carry out an in-depth analysis of factual errors in summaries created by machines, using a structured approach to identifying and classifying such errors.

We developed, linear regression based framework that guides participants to recognize and classify factual errors into four broad areas: misrepresentation, incorrect quantities or measures, improper attribution, and invention.

Every class represents a different type of fact inaccuracy, ranging from subtle alterations of the meaning to the mere invention of facts. Misrepresentation involves biased representation that changes the original meaning involved; on the other hand, discrepancies in quantities or measurements refer to changes in numerical data. False attribution refers to the wrong association of statements or actions

*Forum for Information Retrieval Evaluation, December 12-15, 2024, India*
*Corresponding author.
†These authors contributed equally.
rishikajhashiny@gmail.com (R. Jha); rakshitahuja26@gmail.com (R. Ahuja); sainik.mahata@gmail.com (S. K. Mahata); monalisa.dey.21@gmail.com (M. Dey); dipankar.dipnil2005@gmail.com (D. Das)

with certain individuals or entities; on the other hand, fabrication pertains to the creation of completely inaccurate information. This effort represents an important study regarding the dependability of these machine-produced summaries, with a focus on the identification of errors of factality, through juxtaposition with original texts. It has been primarily done to increase their trustworthiness and thereby prove their status as reliable sources of information, especially within such cross-lingual frameworks as those that involve Indian languages. Based on an analysis of the degree of factual inaccuracy, this research aims to improve the quality of machine-generated content while moving towards more accurate and reliable digital communication. Preliminary results show that although the proposed framework is proven to be efficient for specific categories of error detection, there is scope for further improvement of the classification methodology.

## 2. Dataset

### 2.1. Task 1:

The dataset for this task was provided by ILSUM 2024[1] [1, 2, 3, 4, 5, 6, 7]. We utilized the English language dataset for our experiment consisting of train, test and validation datasets. The train set consists of 9,500 articles along with its id, heading and summary , whereas the test set contains 2,500 news articles along with the respective ids and headings. The task was to generate a fixed-length summary on the test data.

### 2.2. Task 2:

The dataset for this task was also provided by ILSUM 2024. We utilized the English language dataset for our experiment that consisting of train, test and validation datasets. The train set consists of 4974 articles along with its id, heading and summary , whereas the test set contains 200 articles from both Hindi and Gujarati dataset which checks and corrects the accuracy of information in articles by using many fields such as:

- **Id:** It is the unique identifier for the entry.
- **Title:** The title to the article.
- **Headlines:** Headlines which summarize the key points.
- **Article:** Text appearing in the article under review.
- **Incorrect_Summary:** A summary that inaccurately presents the content of the article.
- **Correct_Summary:** A summary that clearly reflects the content of the article.
- **Incorrect_Summary_Hindi:** An incorrect summary translated into Hindi.
- **Correct_Summary_Hindi:** Correct Summary in Hindi.
- **Incorrect_Summary_Gujarati:** Incorrect Summary Translated in Gujarati.
- **Correct_Summary_Gujarati:** Correct Summary in Gujarati.

## 3. Related work

Gowhar et al. [8], used sequence-to-sequence T5 model for the summarization task of ILSUM 2023. Ranganathan and Abuka [9], used fine-tuned T5 transformer model for summarization on the UCI and BBC datasets. Lalitha et al. [10], used several abstractive summarization techniques such as T5(Text-to-Text Transfer Transformer), BART (Bidirectional Auto-Regressive Transformer) and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence- to-sequence) for summarization.

Deroy et al. [11] Prompted Zero-Shot Multi-label Classification of Factual Incorrectness in Machine-Generated Summaries, Cao et al. [12] employed the fact triples from open information extraction to

[1]https://ilsum.github.io/ilsum/2024/index.html

enhance summarization models, Goodrich et al. [13] compared extraction systems to assess factual accuracy, and Falke et al. [14] employed natural language inference for the evaluation of summaries but found that existing models were still inadequate. Kryscinski et al. [15] proposed a weakly-supervised fact verification model for factual consistency.

## 4. Task Definition

### 4.1. Task 1:

The objective of this task is to generate a fixed-length summary from news article specifically for English language in an either abstractive or extractive way.

### 4.2. Task 2:

This exercise would seek to establish errors in machine-generated summaries by classifying each summary according to the nature of types of errors identified. Possible categories for factual inaccuracies include:

1. ***Misrepresentation:*** This involves distorting information, which can lead to a misleading or erroneous impression. It may manifest through exaggerating certain aspects, downplaying others, or manipulating facts to support a specific narrative.
   **E.g.,**
   **Original information:** "The company's profits declined by 10% due to increased production costs."
   **Misrepresented summary:** "The company had huge losses because they spent too much on unnecessary expenses."
   **Issue:** The summary implies irresponsibility in spending rather than increased necessary costs.
2. ***Inaccurate Quantities or Measurements:*** Fact inaccuracies may result due to errors in the reporting of specific figures, measurements, or statistics, either through careless mistakes or intentional falsification.
   **E.g.,**
   **Original information:** "The team collected data from 1,200 survey participants."
   **Misrepresented summary:** "The team collected data from over 2,000 people."
   **Issue:** The summary exaggerates the quantity, misrepresenting the study's scale.
3. ***False Attribution:*** False attribution refers to mistaken giving of origin to a statement, concept, or action to the wrong someone or group, thereby distorting the understanding of origin.
   **E.g.,**
   **Original information:** "Dr. Smith's research found that the new drug has a 70% success rate in clinical trials."
   **Summary with false attribution:** "The drug has a 90% success rate according to the FDA."
   **Issue:** The success rate is altered and falsely attributed to a different source (the FDA).
4. ***Fabrication:*** This is the severest form of inaccuracy. It infers inventing data, events, or sources, giving rise to "facts" without any basis in reality.
   **E.g.,**
   **Original information:** "The study did not find any significant difference between the two treatments." **Fabricated summary:** "The study found that one treatment was 50% more effective than the other." **Issue:** The summary introduces a non-existent finding, fabricating results that weren't present in the original study.

By categorizing such errors, the intent is to systematically identify and correct various types of factual mistakes, thus enhancing the understanding of how auto-generated summaries are likely to deviate from factual accuracy.

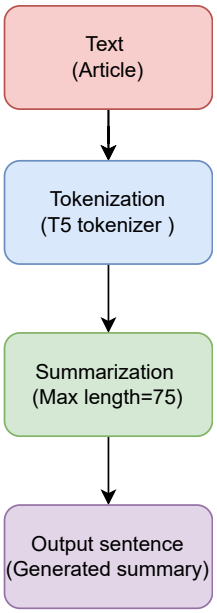


**Figure 1:** Proposed framework for summarization task. (Task 1)

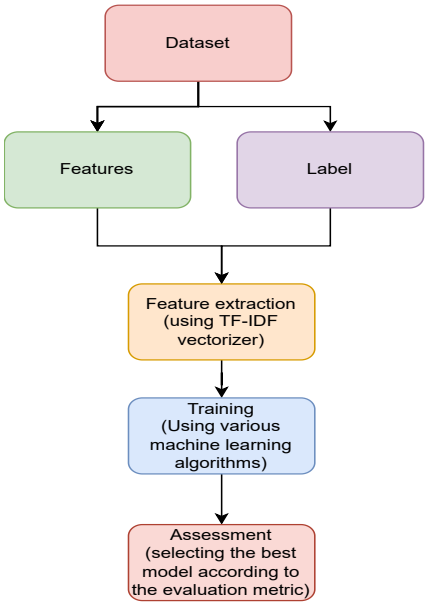


**Figure 2:** Proposed framework for detecting factual incorrect summary. (Task 2)

# 5. Methodology:

## 5.1. Task 1:

### 5.1.1. Model description

We have used pre-trained T5-small model. It is a language model developed by Google AI, part of the T5 (Text-to-Text Transfer Transformer) series. It's a smaller version of the T5 model, with 60 million parameters, making it more efficient and suitable for deployment on smaller devices or in applications with limited resources.it can be used for various NLP tasks, including translation, document summarization, question answering as well as classification tasks. It has been pre-trained on the Colossal Clean Crawled Corpus (C4), which is a mixture of unsupervised and supervised data and performs well on a wide range of languages.

### 5.1.2. Text processing and tokenization

We have used the T5 tokenizer . It converts text into a format that the model can understand and process, as well as converting model outputs back into human-readable text. It is based on the SentencePiece algorithm, which is a data-driven subword tokenizer.

### 5.1.3. Summarization function

We have created a summarization function that takes the "Article" as input and tokenize it using the pretrained T5 tokenizer . The max length of the input text is set to 1,024, which means the input text will be truncated if it exceeds this length. The next step uses the pre-trained model to generate summaries based on the input taking several parameters. The maximum length of the generated summary is set to 75 whereas the minimum length to 40. The length penalty is set to 2 in order to discourage long summaries. The number of beams to use in the beam search algorithm is set to 4 and early stopping to "true". The proposed framework is depicted in Figure 1.

### 5.2. Task 2:

This is a comparison between several machine learning classifiers on a text classification problem. The data is read from a tab-separated file. There are two text columns that will form the features X, and one label column y.

#### 5.2.1. Feature Extraction

The first two columns consist of features, and the last column is the target label. The text in the feature columns is vectorized using TF-IDF (Term Frequency-Inverse Document Frequency). Two independent `TfidfVectorizer` objects transform the text data into numerical representations and are then concatenated to form a complete set of features (X-combined). Data Partitioning: The combined feature matrix (X-combined) and corresponding labels (y) split into the training and testing subsets based on an 80/20 distribution. This approach ensures models are trained on one particular subset but evaluated on a different subset to test their generalization ability.

#### 5.2.2. Training the Model

The training data is used to build different classification models, including Logistic Regression, SVC, Decision Tree, Random Forest, and Naive Bayes that is MultinomialNB. All these classifiers were used from Scikit Learn Package[2].

#### 5.2.3. Assessment

For each classifier, it generates predictions for the test set (X-test). Performance is assessed using several metrics: accuracy, precision (weighted), recall (weighted) and F1-score (macro average). The findings are organized in a dictionary format and subsequently presented for easier comparison of the classifiers' performance based on calculated metrics, thus aiding in finding the best model for the given task and after the successful comparison we found that the logistic regression performs the best among the algorithms. Thus, we made the final submission using this model shown in Figure 2.

## 6. Result

As per the official results for the ILSUM 2024 task, our team 'Iem inturns" was able to achieve notable scores.

### 6.1. Task 1:

Our performance specifically for English dataset in terms of the ROUGE metrics as well as Bert score are shown in Tables 1 and 2.

**Table 1**
Scores using Rouge metrics

| Rank | team-name | language | run name | Rouge-1 | Rouge-2 | Rouge-4 | Rouge-L |
|---|---|---|---|---|---|---|---|
| 4 | Iem inturns | English | run1.csv | 0.3044 | 0.1448 | 0.0843 | 0.2646 |

**Table 2**
Scores using BERT

| Rank | team-name | language | run name | BertScore-Precision | BertScore-Recall | BertScore-F1 |
|---|---|---|---|---|---|---|
| 7 | Iem inturns | English | run1.csv | 0.8482 | 0.8708 | 0.8591 |

[2]https://scikit-learn.org/1.5/supervised_learning.html

### 6.2. Task 2:

Our results for the incorrect summary detection task are shown in Tables 3 and 4.

**Table 3**
Scores for Gujarati data

| Rank | team-name | language | run name | F1-Score |
|---|---|---|---|---|
| 4 | lem inturns | Gujarati | run1.csv | 0.2127 |

**Table 4**
Scores for Hindi data

| Rank | team-name | language | run name | F1-Score |
|---|---|---|---|---|
| 6 | lem inturns | Hindi | run1.csv | 0.2132 |

## 7. Conclusion and Future work

In this study, we addressed the challenge of multi-label error classification for machine-generated summaries, focusing on misrepresentation, fabrication, false attribution, and incorrect quantities. For Task 1, the best results were achieved using the T5-small model. For Task 2, which involved classifying factual inaccuracies, our experiments found that a linear regression model yielded the most promising results. Future work will explore few-shot learning and utilize larger models like GPT-4 to further improve classification accuracy as well as summarization of articles in other languages.

We can additionally implement machine learning algorithms to improve the factual accuracy of neural abstractive summarization systems, with a specific focus on article summarization. Both automated assessments and human evaluations validated that these models, when combined with information extraction (IE) methodologies, notably enhance factual precision. Although the optimization of conventional metrics such as ROUGE led to improved summary quality, the factual scoring framework was insufficiently sensitive to particular errors, including the interchange of numbers or pronouns. In the future, our goal is to enhance fact encoders, optimize the outputs of information extraction, and use complex machine learning methodologies that enhance the reliability of summaries generated by machines.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References


[1] S. Satapara, P. Mehta, S. Modha, A. Hegde, S. HL, D. Ganguly, Overview of the third shared task on indian language summarization (ilsum 2024), in: K. Ghosh, T. Mandl, P. Majumder, D. Ganguly (Eds.), Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation, Gandhinagar, India. December 12-15, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[2] S. Satapara, P. Mehta, S. Modha, A. Hegde, S. HL, D. Ganguly, Key insights from the third ilsum track at fire 2024, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2024, Gandhiinagar, India. December 12-15, 2024, ACM, 2024.

[3] S. Satapara, B. Modha, S. Modha, P. Mehta, FIRE 2022 ILSUM track: Indian language summarization, in: D. Ganguly, S. Gangopadhyay, M. Mitra, P. Majumder (Eds.), Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2022, Kolkata, India, December

9-13, 2022, ACM, 2022, pp. 8–11. URL: https://doi.org/10.1145/3574318.3574328. doi:10.1145/3574318.3574328.

[4] S. Satapara, P. Mehta, S. Modha, D. Ganguly, Indian language summarization at FIRE 2023, in: D. Ganguly, S. Majumdar, B. Mitra, P. Gupta, S. Gangopadhyay, P. Majumder (Eds.), Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Panjim, India, December 15-18, 2023, ACM, 2023, pp. 27–29. URL: https://doi.org/10.1145/3632754.3634662. doi:10.1145/3632754.3634662.

[5] S. Satapara, B. Modha, S. Modha, P. Mehta, Findings of the first shared task on indian language summarization (ILSUM): approaches challenges and the path ahead, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022, volume 3395 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 369–382. URL: https://ceur-ws.org/Vol-3395/T6-1.pdf.

[6] S. Satapara, P. Mehta, S. Modha, D. Ganguly, Key takeaways from the second shared task on indian language summarization (ILSUM 2023), in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023, volume 3681 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 724–733. URL: https://ceur-ws.org/Vol-3681/T8-1.pdf.

[7] S. Satapara, P. Mehta, D. Ganguly, S. Modha, Fighting fire with fire: Adversarial prompting to generate a misinformation detection dataset, CoRR abs/2401.04481 (2024). URL: https://doi.org/10.48550/arXiv.2401.04481. doi:10.48550/ARXIV.2401.04481. arXiv:2401.04481.

[8] S. Gowhar, B. Sharma, A. K. Gupta, A. K. Madasamy, Advancing human-like summarization: Approaches to text summarization., in: FIRE (Working Notes), 2023, pp. 747–754.

[9] J. Ranganathan, G. Abuka, Text summarization using transformer model, in: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2022, pp. 1–5.

[10] E. Lalitha, K. Ramani, D. Shahida, E. V. S. Deepak, M. H. Bindu, D. Shaikshavali, Text summarization of medical documents using abstractive techniques, in: 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, 2023, pp. 939–943.

[11] A. Deroy, S. Maity, S. Ghosh, Prompted zero-shot multi-label classification of factual incorrectness in machine-generated summaries., in: FIRE (Working Notes), 2023, pp. 734–746.

[12] Z. Cao, F. Wei, W. Li, S. Li, Faithful to the original: Fact aware neural abstractive summarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[13] B. Goodrich, V. Rao, P. J. Liu, M. Saleh, Assessing the factual accuracy of generated text, in: proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 166–175.

[14] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, Ranking generated summaries by correctness: An interesting but challenging application for natural language inference, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2214–2220. URL: https://aclanthology.org/P19-1213. doi:10.18653/v1/P19-1213.

[15] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346. URL: https://aclanthology.org/2020.emnlp-main.750. doi:10.18653/v1/2020.emnlp-main.750.