Text Summarization For Hindi Language

CSE556: Natural Language Processing

Nikhil Suri	Maanas Gaur	Lakshya Aggrawal	Lakshya Kumar
2021268	2021537	2021535	2021536

April 25, 2024

1 Introduction

1.1 Motivation

Text summarization is an important task in natural language processing (NLP), particularly for languages like Hindi, where the availability of resources such as datasets is limited compared to English. The complexity of Hindi's syntax, coupled with its rich morphological features, poses unique challenges in NLP. The reason for choosing this project is that summarization can aid in information dissemination, making content more accessible and digestible for users.

1.2 Examples

In the realm of news media, automated summaries of news articles can help readers quickly grasp the significant events of the day without delving deep into full articles. In academia, summarization can condense research papers, making the review process more efficient. In business, summarizing reports and documents allows for quick decision-making by highlighting the most essential information.

2 Related Work

2.1 Literature Review

The field of text summarization has seen various approaches, including extractive and abstractive techniques. The common tasks in all these different approaches are: [4]

- Capturing key aspects and storing them
- Scoring them according to importance
- Selecting the appropriate sentences for summary

For Hindi, research has often focused on extractive methods due to the straightforward implementation and lower resource requirements [4]. However, recent advancements have seen a shift towards more sophisticated neural network-based models [6], which can generate more coherent and contextually accurate summaries.

2.2 Gaps in Current Research

Existing models often struggle with the nuances of Hindi syntax and semantics. There is also a notable scarcity of large, annotated datasets for training models in Hindi, which hampers the development of effective summarization tools [6].

3 Dataset Description

	Id	Heading	Summary	Article
0	hindi_2023_train_0	गला दबाकर हत्या की; बॉडी बोरे में भरी, लोकल मा...	Kerala Minor Girl Rape Case - केरल के एर्नाकुल...	केरल के एर्नाकुलम जिले में 5 साल की बच्ची से र...
1	hindi_2023_train_1	तेलंगाना में 18 की जान गई; जम्मू-कश्मीर में बा...	इस साल मानसून सीजन में कई राज्यों में भारी तबा...	मानसून सीजन में हुई भारी बारिश ने कई राज्यों म...
2	hindi_2023_train_2	राजस्थान सरकार बनाएगी कर्ज राहत आयोग, कोर्ट के...	चुनावी साल में राजस्थान सरकार किसानों को लुभान...	चुनावी साल में राजस्थान सरकार किसानों को लुभान...
3	hindi_2023_train_3	3 से 7 अगस्त तक कर सकेंगे अप्लाय, प्राइस बैंड ...	Non-banking lender SBFC Finance's initial publ...	नॉन बैंकिंग फाइनेंस कंपनी 'SBFC फाइनेंस लिमिटेड...
4	hindi_2023_train_4	डाइनिंग टेबल पर कुकर-कड़ाही न सजाएं, चीन के खा...	स्वाद खाने की बुनियाद है। लेकिन अगर खाना सुंदर...	स्वाद खाने की बुनियाद है। लेकिन अगर खाना सुंदर...
...	...	...	...	...

Figure 1: First 5 rows of the Dataset

3.1 Composition

The dataset *hindi_train.csv* from ILSUM-23 contains Hindi articles paired with their respective summaries and headings. Each row represents a unique article-summary pair, structured into specific columns for both article heading and summary.

	Article_Length	Summary_Length
count	21225.000000	21225.000000
mean	2823.756796	206.807256
std	2227.719796	57.134802
min	1.000000	55.000000
25%	1443.000000	165.000000
50%	2059.000000	204.000000
75%	3263.000000	245.000000
max	18016.000000	513.000000

Figure 2: Article summary length stats

3.2 Volume

The dataset includes a significant number of entries (around 20K samples), offering a decent sample size for training the models.

3.3 Sample Texts

Articles: Sample articles cover a range of topics from general news to specific cultural discussions, showcasing the dataset's diversity.

Summaries: The summaries are concise and aim to capture the essential information of the articles, providing a clear target for the models' output.

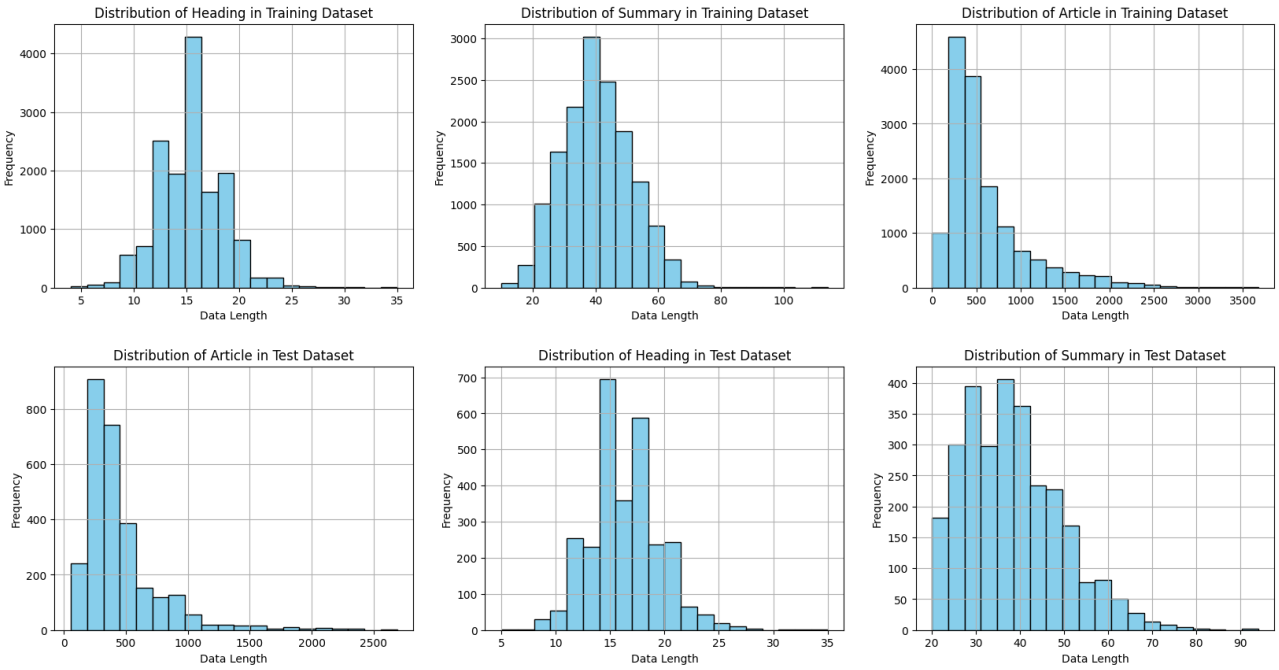

Figure 3: Histogram plots for Columns in the Dataset (ILSUM-22 used for testing)

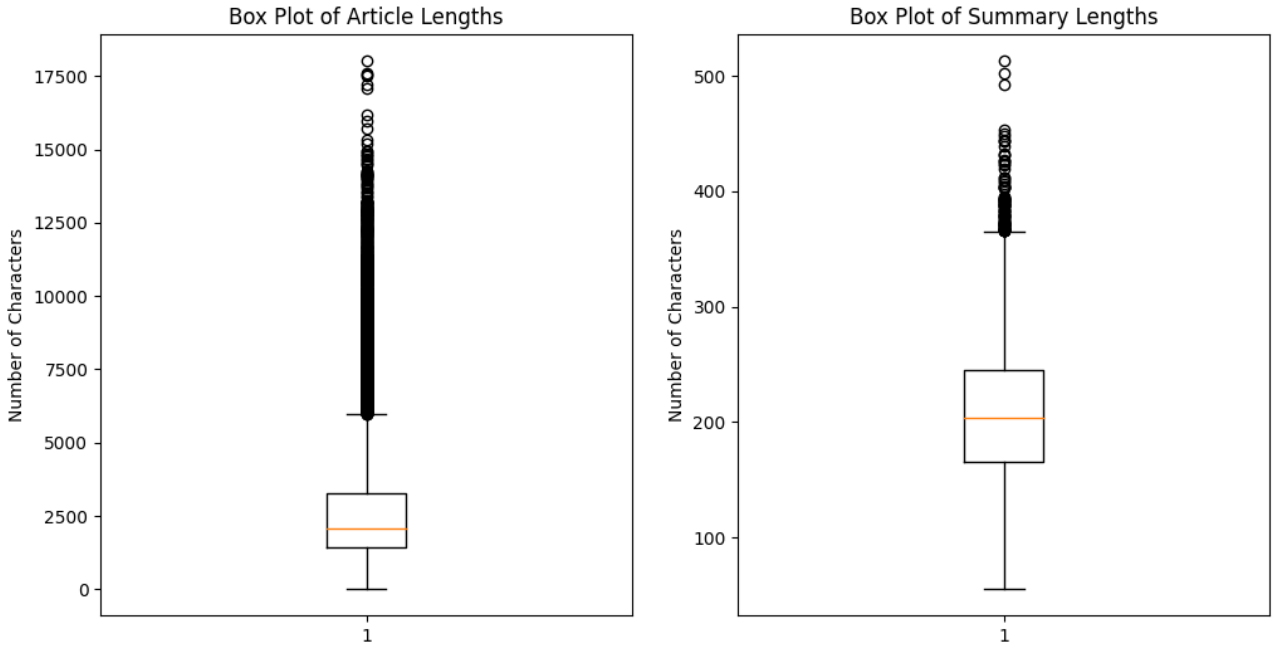

Figure 4: Identifying outliers in text lengths using box plots

4 Methodology

4.1 Pre-processing Steps

Our approach to preprocessing text for the Hindi text summarizer involves several steps to ensure clean and informative summaries. First, we focus on text cleaning. Special characters, punctuation (except those aiding sentence structure), and extra spaces are eliminated. Additionally, to ensure the model prioritizes Hindi content and avoids confusion any English characters that may be present, are removed (Only for IndicBART and Transformer). Finally,

we convert numeric values to their Hindi word representations. This step ensures the model focuses on the semantic meaning conveyed by numbers. By applying these preprocessing steps, we are essentially preparing the Hindi text as input for the models.

4.2 Model Description

This project utilizes three main models: IndicBART, Transformer Model, and MT5. IndicBART is adapted specifically for Indian languages and is fine-tuned for summarization tasks. It is a multilingual seq2seq model, features 6 encoderdecoder layers, 16 attention heads totaling 244M parameters. MT5, being a multilingual model, has also been fine-tuned for the Hindi Summarization task for a direct comparison with IndicBART. It is a multilingual text-to-text model with 12 encoder-decoder layers with 12 attention heads each. The Transformer Model is a custom implementation designed to tackle the specific challenges pertaining to Hindi texts.

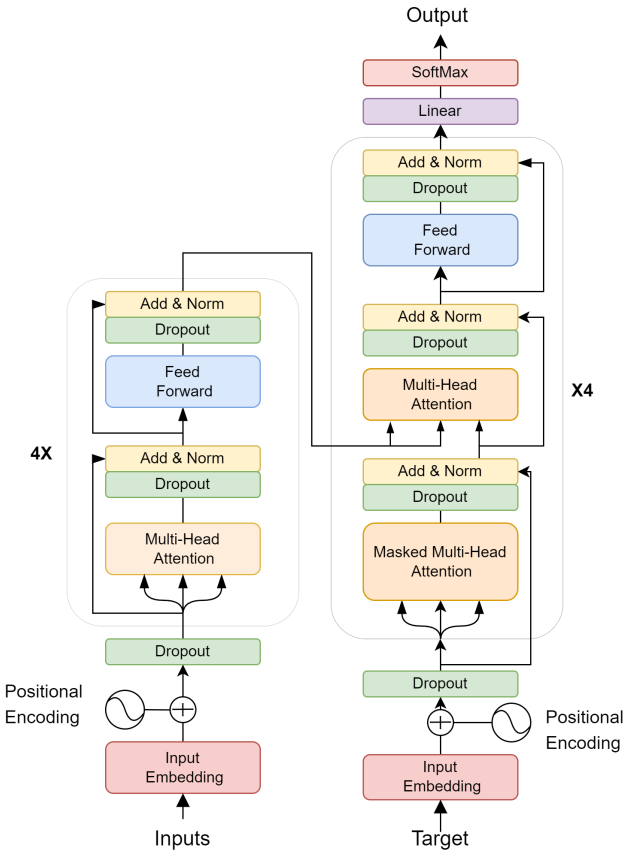

Figure 5: Model architecture of Custom Transformer

4.3 Rationale

These models were chosen to compare the effectiveness of specialized versus generalized approaches in handling the Hindi language, and to explore how different architectures influence performance on the same task.

4.4 Experimental Setup

The models were trained on a standardized hardware setup, with parameters optimized for each model being learned and updated in each iteration/epoch. Performance metrics such as BERTScore and ROUGE scores were used to evaluate model effectiveness.

5 Results/Findings

IndicBART showed the highest performance across the metrics, particularly in handling the linguistic features of Hindi. The Transformer Model, while effective, lagged slightly behind IndicBART, particularly in recall and precision. MT5 also demonstrated robustness in the form of good BERT and Rouge scores.

Table 1: Performance of IndicBART, Transformer, and MT5 on BERT and ROUGE Scores

Metric	IndicBART	Transformer	MT5
BERT Precision	0.771	0.612	0.783
BERT Recall	0.826	0.542	0.765
BERT F1	0.797	0.737	0.773
ROUGE-1 Precision	0.477	0.441	0.525
ROUGE-1 Recall	0.684	0.369	0.444
ROUGE-1 F-Score	0.554	0.393	0.471
ROUGE-2 Precision	0.358	0.162	0.395
ROUGE-2 Recall	0.562	0.149	0.316
ROUGE-2 F-Score	0.429	0.151	0.342
ROUGE-L Precision	0.435	0.356	0.477
ROUGE-L Recall	0.626	0.300	0.405
ROUGE-L F-Score	0.506	0.318	0.429

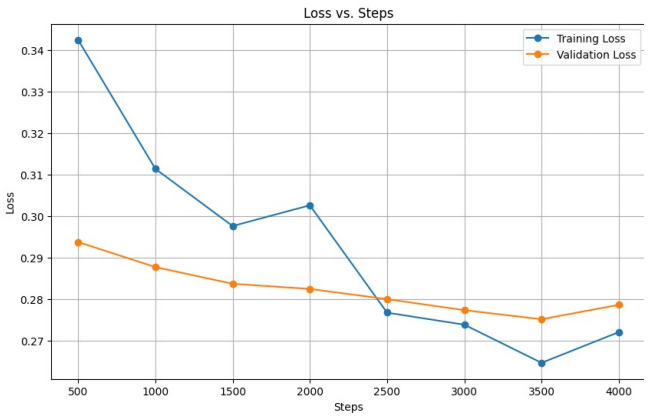

Figure 6: Loss Plot for MT5

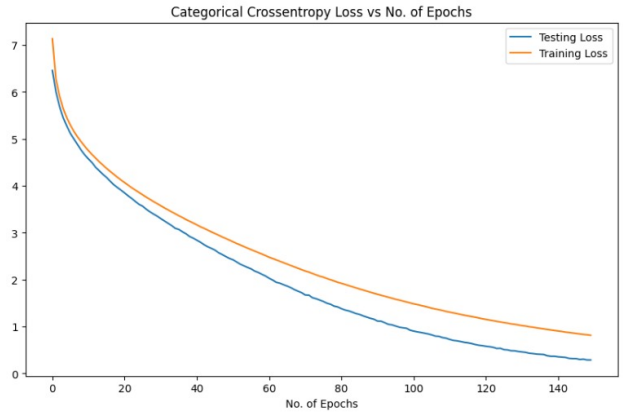

Figure 7: Loss Plot for Custom Transformer

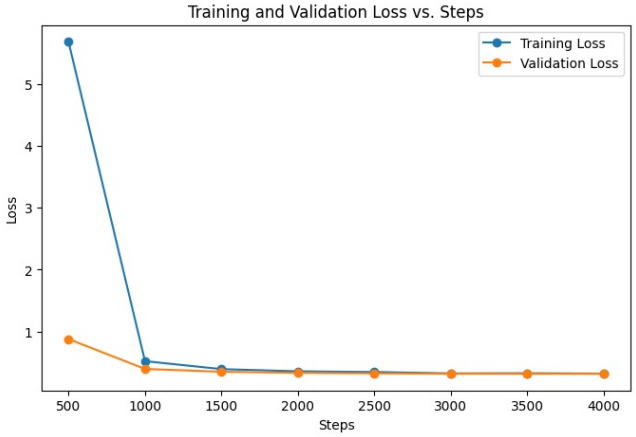

Figure 8: Loss Plot for IndicBART

6 Discussion/Analysis/Observation

Interesting Observations: One notable observation was the impact of training data diversity on model performance. IndicBART, trained specifically on a diverse range of Hindi texts, outperformed others, suggesting the importance of tailored models and datasets for language-specific tasks.

7 Model Loss Analysis

Model	Key Observations	Implications
MT5	• Steep initial training loss decline • Consistent downward trend • Gradual plateau of loss curve • Two spikes in the plot	• Rapid early learning • Effective ongoing learning • Approaching convergence • Stable learning, Slight overfitting
IndicBART	• Sharp initial training loss decline • Loss levels off quickly • Plateaus earlier than MT5	• Characteristic of rapid learning • Max learning capacity reached • Quicker fitting to data
Transformer Model	• Both losses decrease over time • No upward trend in testing loss	• Effective learning, generalization • No overfitting

Table 2: Analysis of Model Loss Plots

Metric	Recall	Precision	F-measure
ROUGE-1	0.739602	0.117524	0.191215
ROUGE-2	0.497628	0.060592	0.099824
ROUGE-L	0.626215	0.100318	0.162549
BERTScore	0.777183	0.628468	0.694204

Figure 9: TF-IDF scores

Category	Content
Article	केरल के एर्नाकुलम जिले में पाँच साल की बच्ची से रेप के बाद गला दबाकर हत्या कर दी गई। आरोपी ने बच्ची का शव बोरे में डालकर डंपिंग ग्राउंड में फेंक दिया था। पुलिस ने आरोपी शख्स को गिरफ्तार कर लिया है। घटना शुक्रवार शाम की है। पुलिस ने शनिवार को मीडिया को इसकी जानकारी दी। फुटेज में बच्ची के साथ नजर आया आरोपी एर्नाकुलम विवेक कुमार ने बताया बच्ची शुक्रवार शाम को किडनैप हुई थी। हमारी टीम ने फुटेज चेक किए जिसमें बच्ची को आरोपी के साथ देखा गया। उसी दिन रात नौ.तीस बजे आरोपी को गिरफ्तार कर लिया गया। उस समय वह नशे की हालत में था और बच्ची उसके साथ नहीं थी। स्थानीय लोगों ने बच्ची को मार्केट के पास आरोपी के साथ देखा था। उन्होंने पुलिस को इसकी जानकारी दी। पुलिस ने पूरे एरिया में सर्चिंग की तो मार्केट के पीछे बच्ची की लाश मिली। इस इलाके में लोग कूड़ा फेंकते थे और कई असामाजिक लोग यहां नशा करने आते थे। बच्ची की बिल्डिंग में ही रहता था आरोपी पुलिस ने बताया आरोपी शख्स बिहार का रहने वाला है और केरल में मजदूरी करता है। बच्ची के माता-पिता भी बिहार के हैं और मजदूरी करते हैं। जिस इमारत में बच्ची रहती थी उसी बिल्डिंग के पहले फ्लोर पर आरोपी रहता था। आरोपी ने पूछताछ में पहले पुलिस को गुमराह करने की कोशिश की। हालांकि शनिवार सुबह अपना जुर्म कबूल कर लिया। कांग्रेस ने की पीड़ित परिवार को मुआवजा देने की मांग घटना को लेकर विपक्षी पार्टी कांग्रेस ने लेफ्ट की सरकार पर हमला बोला है। विधानसभा में विपक्ष के नेता वीडी सतीशन ने कहा कि पुलिस बच्ची को समय रहते नहीं ढूंढ पाई। कांग्रेस प्रदेशाध्यक्ष के सुधाकरण ने सरकार से पीड़ित परिवार को मुआवजा देने की अपील की है।
SUMMARY	केरल के एर्नाकुलम जिले में पाँच साल की बच्ची से रेप के बाद गला दबाकर हत्या कर दी गई। आरोपी ने बच्ची का शव बोरे में डालकर डंपिंग ग्राउंड में फेंक दिया था। फुटेज में बच्ची के साथ नजर आया आरोपी एर्नाकुलम विवेक कुमार ने बताया बच्ची शुक्रवार शाम को किडनैप हुई थी। हमारी टीम ने फुटेज चेक किए जिसमें बच्ची को आरोपी के साथ देखा गया। उसी दिन रात नौ तीस बजे आरोपी को गिरफ्तार कर लिया गया। पुलिस ने पूरे एरिया में सर्चिंग की तो मार्केट के पीछे बच्ची की लाश मिली। इस इलाके में लोग कूड़ा फेंकते थे और कई असामाजिक लोग यहां नशा करने आते थे। हालांकि शनिवार सुबह अपना जुर्म कबूल कर लिया। कांग्रेस ने की पीड़ित परिवार को मुआवजा देने की मांग घटना को लेकर विपक्षी पार्टी कांग्रेस ने लेफ्ट की सरकार पर हमला बोला है। विधानसभा में विपक्ष के नेता वीडी सतीशन ने कहा कि पुलिस बच्ची को समय रहते नहीं ढूंढ पाई। कांग्रेस प्रदेशाध्यक्ष के सुधाकरण ने सरकार से पीड़ित परिवार को मुआवजा देने की अपील की है।

Figure 10: One of the TF-IDF samples

7.1 TF-IDF for Summarization

TF (Term Frequency) measures how frequently a word appears in a single document. The more a word shows up, the higher its TF score. IDF (Inverse Document Frequency) balances out the TF score by considering how common a word is across a collection of documents. If a word appears frequently across many documents, it is less informative for a specific document's summary, thus receiving a lower IDF score.

- **Sentence Scoring**: Each sentence in the document is assigned a score based on the TF-IDF weights of the words it contains. Sentences containing more high-scoring words are considered more informative.
- **Summary Selection**: The top-scoring sentences are selected to form the summary. These sentences are likely to encapsulate the key points and concepts from the original document.

Limited Context: Despite its utility, TF-IDF does not consider the relationships between words or the overall flow of the text. This can result in summaries that lack coherence or fail to capture subtle nuances.

7.2 Qualitative Analysis

Examples of the three model's text summarization performance, with actual outputs compared to the model's predictions:

Category	Details
Actual Summary	दिल्ली पुलिस कमिश्नर एसएन श्रीवास्तव sn shrivastava ने गणतंत्र दिवस republic day पर होने वाली किसानों की ट्रैक्टर रैली farmers tractor rally की तैयारियों का जायज़ा लिया
Output	farmers tractor rally planning update नई दिल्ली दिल्ली पुलिस कमिश्नर एसएन श्रीवास्तव sn shrivastava ने गणतंत्र दिवस republic day पर होने वाली किसानों की तैयारियों का जायज़ा लिया

Figure 11: Qualitative Analysis for MT5

#	Category	Details
Example 1	**Output**	उत्तर प्रदेश में शुक्रवार को कोरोना वायरस से संक्रमण के दो हज़ार चौबिस नए मामले सामने आए हैं इसी के साथ ही सूबे में इस घातक वायरस से संक्रमित होने वाले लोगों की कुल संख्या बढ़कर सात नौ सौ पचहत्तर हो गई है।
	Actual Summary	उत्तर प्रदेश में शुक्रवार को कोरोना वायरस से संक्रमण के दो नौ सौ सरसठ नए मामले सामने आए हैं इसी के साथ सूबे में इस घातक वायरस से संक्रमित होने वाले लोगों की संख्या बढ़कर छः बाईस सात सौ छत्तीस हो गई है।
Example 2	**Output**	बहुजन पार्टी की राष्ट्रीय पार्टी की अध्यक्ष मायावती ने सत्ताईस सितंबर को किसान संगठनों द्वारा बंद बंद का समर्थन किया है।
	Actual Summary	बहुजन समाज पार्टी की अध्यक्ष मायावती ने कल सत्ताईस सितंबर को किसान संगठनों द्वारा बुलाए गए भारत बंद का समर्थन किया है।

Figure 12: Qualitative Analysis for Transformer Model

Category	Details
Actual Summary	भारतीय सिने जगत के पितामह धुंडिराज गोविंद फाल्के के नाम पर इंडस्ट्री में फर्जी पुरस्कार समारोहों को भी बड़ा बाजार है इसके बावजूद कि सीबीएफसी की मेंबर रहीं वाणी त्रिपाठी टिक्कू से लेकर सूचना एवं प्रसारण मंत्रालय
Output	भारतीय सिनेमा जगत के पितामह धुंडिराज गोविंद फाल्के के नाम पर इंडस्ट्री में फर्जी पुरस्कार समारोह का बड़ा बाजार है सीबीएफसी की मेंबर रहीं वाणी त्रिपाठी टिक्कू से लेकर सूचना एवं प्रसारण मंत्रालय के अधिकारी इस पर ऐतराज जाहिर करते रहें हैं हालांकि इस नाम से जो वास्तविक अवॉर्ड है वो एक हज़ार नौ सौ उनहत्तर

Figure 13: Qualitative Analysis for IndicBART

These examples illustrate the model's ability to comprehend and summarize complex Hindi text, indicating its effectiveness.

8 Conclusion and Future Work

8.1 Summary of Findings

- Overall, all models demonstrate a capacity for learning and improving over time, as evidenced by the decreasing loss values. The MT5 model portrays a more gradual learning curve, potentially contributing to better generalization.
- In contrast, the IndicBART model appears to learn more rapidly but also reaches a plateau earlier, which raises considerations of the attainment of its learning limit from the provided data.
- The Transformer Model displays a good balance in learning from the training data and generalizing to the test data, indicative of a well-tuned model.

8.2 Limitations

The limited size and variety of the dataset could have impacted the models' performance.

8.3 Future Work

Future research could explore expanding the dataset, incorporating more complex model architectures, or applying transfer learning from models trained on other languages to enhance performance on Hindi summarization.

References

[1] Mandl, Thomas et al. 2021. "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages." *FIRE 2021.*

[2] Liu, Yinhan et al. 2020. "Multilingual Denoising Pre-training for Neural Machine Translation" *ArXiv, DOI: arXiv:2001.08210.*

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, *Attention is all you need*, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc., 2017.

[4] Agarwal, A., Naik, S. and Sonawane, S., 2022, December. Abstractive Text Summarization for Hindi Language using IndicBART. In Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation, Kolkata, India.

[5] Shantipriya Parida and Petr Motlicek. 2019. "Abstract Text Summarization: A Low Resource Challenge." *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5994–5998, Hong Kong, China. Association for Computational Linguistics.*

[6] Satapara, S., Modha, B., Modha, S. and Mehta, P., 2022. Findings of the first shared task on indian language summarization (ilsum): Approaches, challenges and the path ahead. Working Notes of FIRE, pp.9-13.