Text Summarization of Hindi News Articles

Dimpy, Ritwik, Aamir
Group-16
Course name: CSE641

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

Many text summarization methods exists for English language. But Hindi neither has an extensive standard dataset nor state-of-art method to generate automated summaries.

Problem Statement

> To develop a model using deep learning techniques that can generate automated summaries from Hindi news articles.

Dataset

A novel dataset created by Gaurav Arora is used for the training and evaluation purposes. The dataset is available on Kaggle.

Evaluation Metric

ROUGE metric is used to assess the quality of the generated summaries

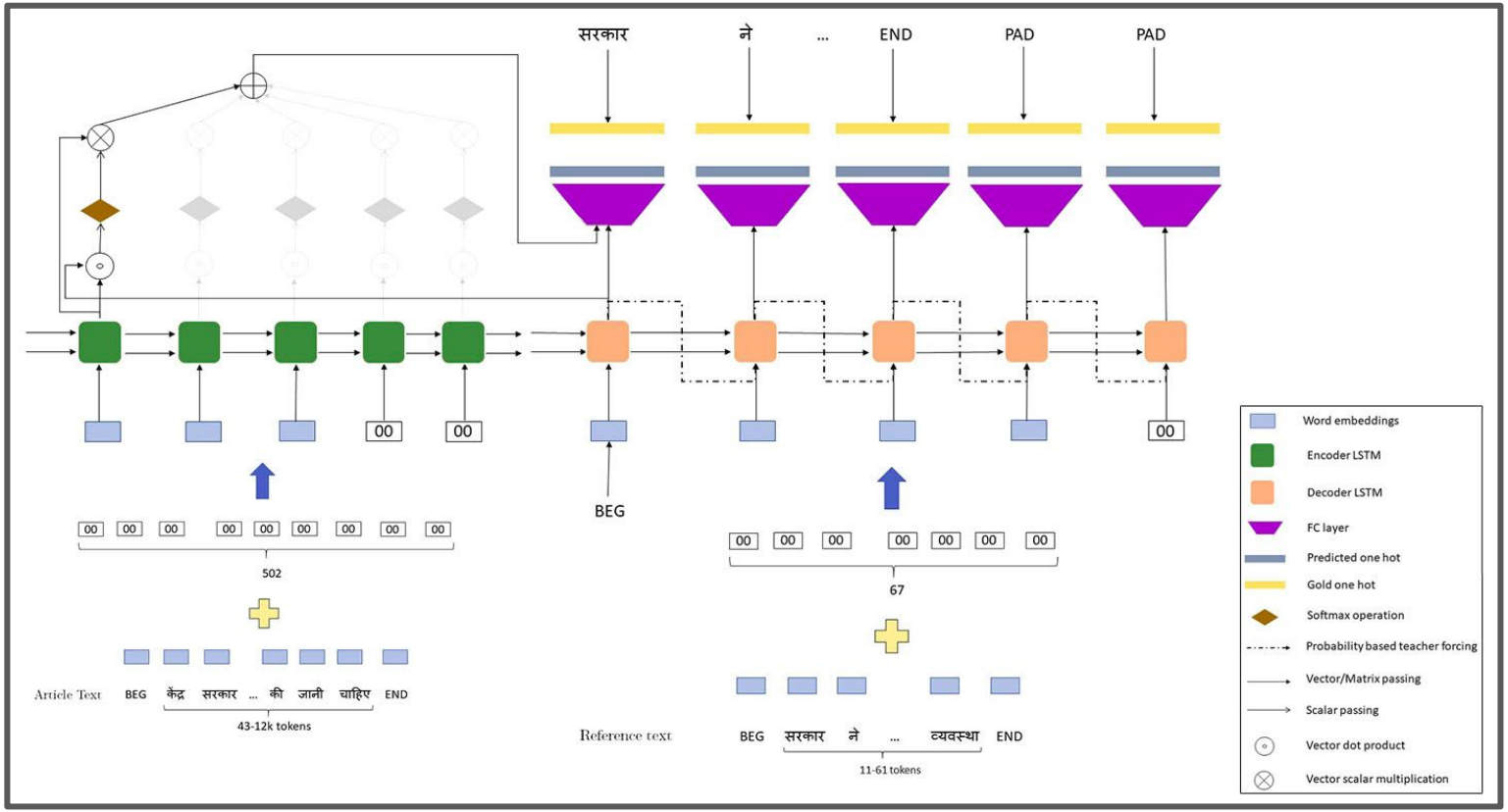
Table 1: Calculation of ROUGE scores for candidate summary (X) and reference summary (Y).

	ROUGE-1 (unigram)	ROUGE-2 (bigram)	ROUGE-L (LCS)
Recall (R)	$\frac{\lvert Unigrams\ common\ in\ X\ and\ Y\rvert}{\lvert Unigrams\ in\ Y\rvert}$	$\frac{\lvert Bigrams\ common\ in\ X\ and\ Y\rvert}{\lvert Bigrams\ in\ Y\rvert}$	$\frac{\lvert LCS(X,Y)\rvert}{\lvert Y\rvert}$
Precision (P)	$\frac{\lvert Unigrams\ common\ in\ X\ and\ Y\rvert}{\lvert Unigrams\ in\ X\rvert}$	$\frac{\lvert Bigrams\ common\ in\ X\ and\ Y\rvert}{\lvert Bigrams\ in\ X\rvert}$	$\frac{\lvert LCS(X,Y)\rvert}{\lvert X\rvert}$
F-score	$\frac{2 * P * R}{P + R}$	$\frac{2 * P * R}{P + R}$	$\frac{2 * P * R}{P + R}$

Approaches tried

1. Abstractive = generating natural language
    a. Seq2seq
    b. Bert based model
2. Extractive = extracting portions from the news article text itself
    a. LEAD-3
    b. TextRank
    c. BERTSUM-Simple_classifier
    d. BERTSUM-RNN

seq2seq



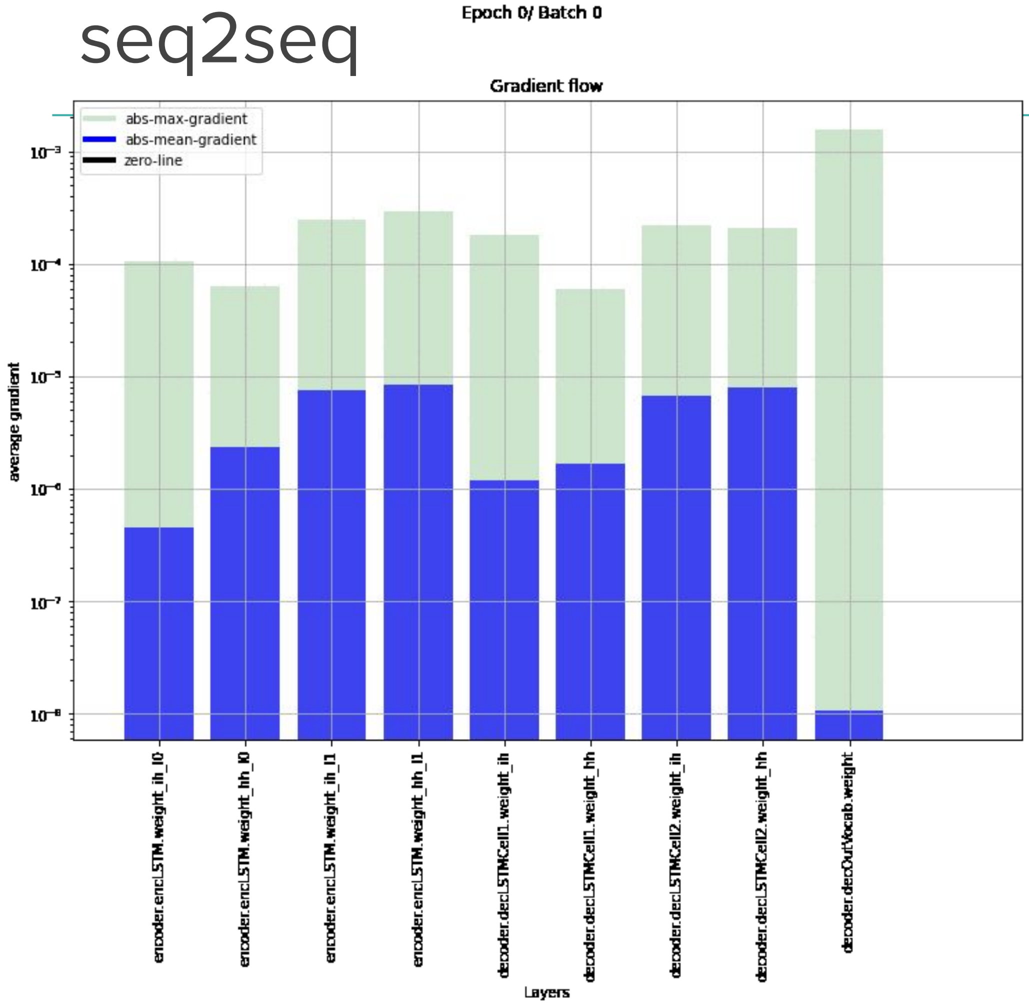
Abstractive

seq2seq

1. Attention and teacher forcing was used
2. Last layer of decoder had output nodes = vocab size
3. Cross_entropy loss was used
4. Suffered from word repetition in the summary
5. Computationally expensive to train
6. Gets stuck in a local optima where it produces the most occuring words (stopwords) in repetition

seq2seq

Abstractive



Actual summary

यादव परिवार में मचा राजनीतिक घमासान और उठापठक जारी है मुलामय की कोशिश यूपी चुनावों में बिहार की तर्ज़ पर महागठबंधन बनाने की है RLD प्रमुख अजित सिंह और जेडीयू के पूर्व अध्यक्ष शरद यादव से संपर्क साधा

Predicted summary

में में

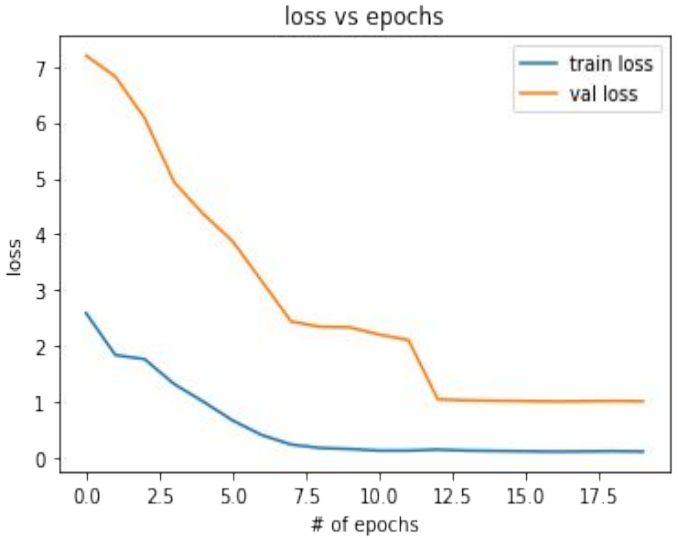
Bert based encoder model

Abstractive

- The second model is an Encoder-Decoder architecture where Encoder has fine-tuned BERT based embeddings for token indices and decoder is a linear layer outputting the probability distribution of every token in the vocab. The encoder part of the model is BERT fine-tuned on the original dataset only for articles having sequence length less than 512 (for faster computations) resulting in a total of 35470 train articles and 8784 test articles.

- The model consists of 12 layers comprising 110M parameters pre-trained on 104 languages including Hindi. The decoder part is just a linear layer that is responsible to learn the probability distributions of each token at the output end for the input tokens.

Train plots and outputs for bert based model



Original Summary	Predicted summary
Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हनन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त रवैया अख्तियार कर लिया है	Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त
Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वह नेटवर्क विस्तार के अगले चरण के बाद बाजार में पहले पायदान पर पहुंचने के लिए मेहनत करेगी	Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वार्षिक बैठक में बाद शुक्रवार को अंतरराष्ट्रीय बाजार के साथ खु
Worst : बंबई उच्च न्यायालय ने सोमवार को दूरसंचार कंपनी वोडाफोन इंडिया से कहा कि वह पांच करोड़ रुपये और उसके यहां 31 मार्च तक जमा कराए	Worst : बुधवार को काले कार्ड तक लगाया गया था जिसमें वित्तवर्षत लीना मूहैया नाम याचिका पर कॉलिंगर अपना रोक2 नाम दियाारी के नए

Extractive

The dataset of abstractive summaries is transformed to extractive summaries using lemmatization and word matching

Article sentences with highest similarity with abstractive summaries were selected as reference extractive summaries

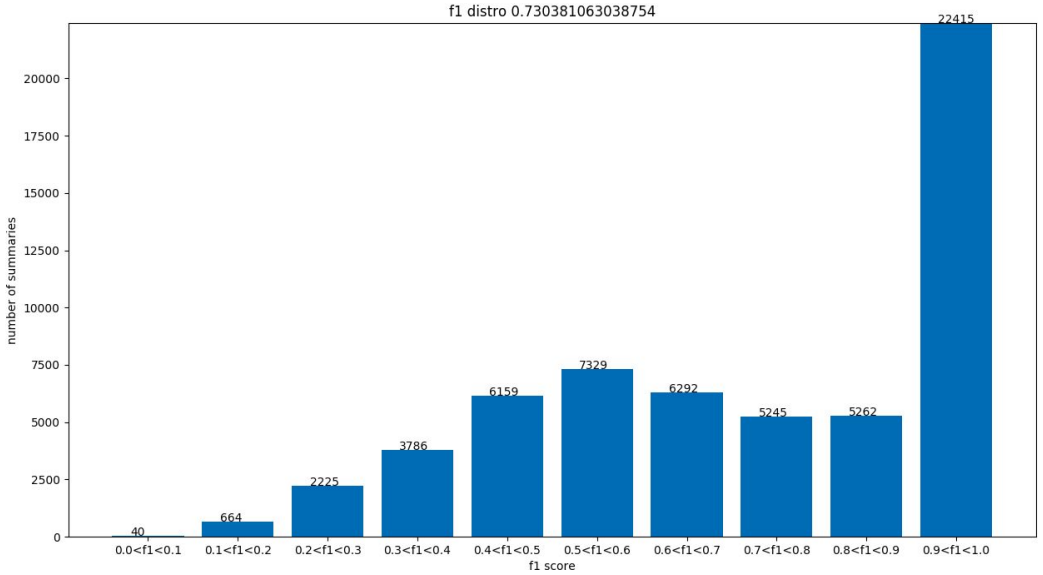
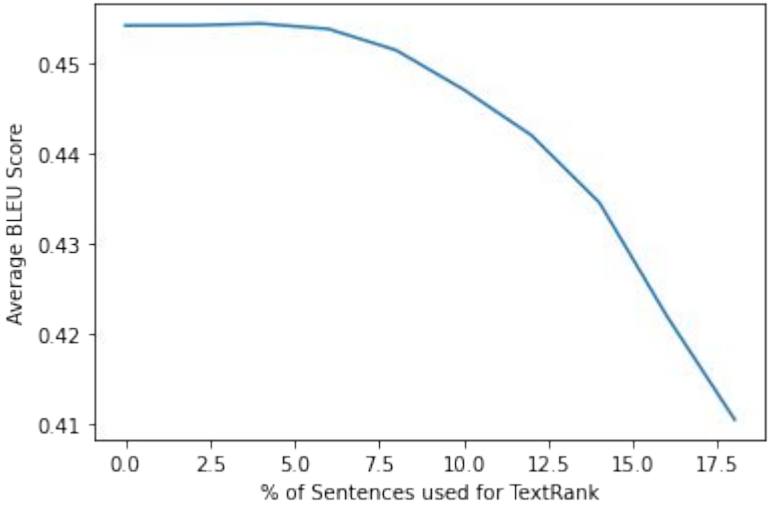


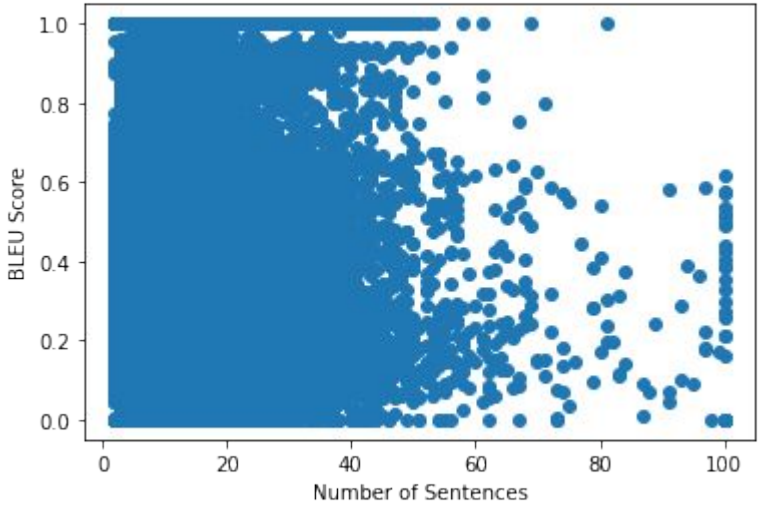
Figure shows that majority of reference extractive summaries are similar to reference abstractive summaries

Text Rank

Extractive

Plots for Text Rank implemented for Bleu score evaluation.

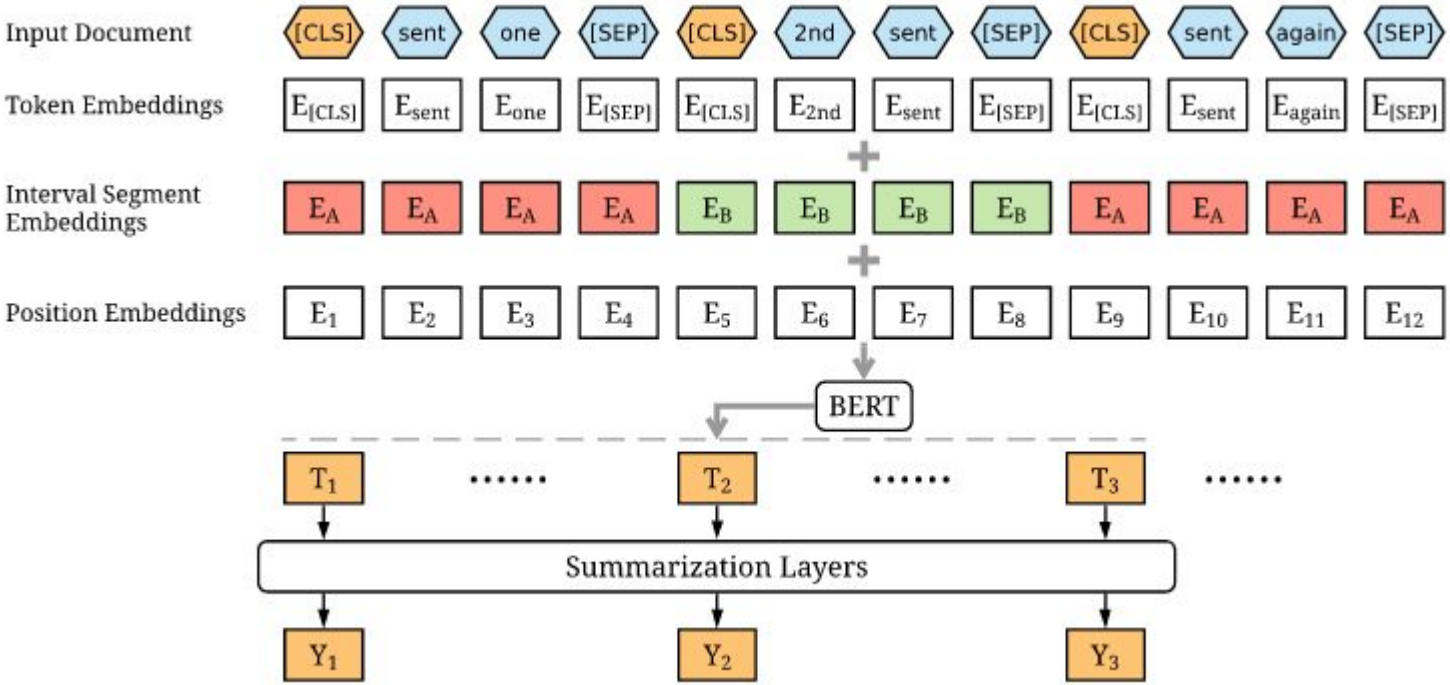




Average Bleu Score = 0.45

BERTSUM

Extractive



We used attention masks instead

image credits : Liu, Y. (2019). Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*.

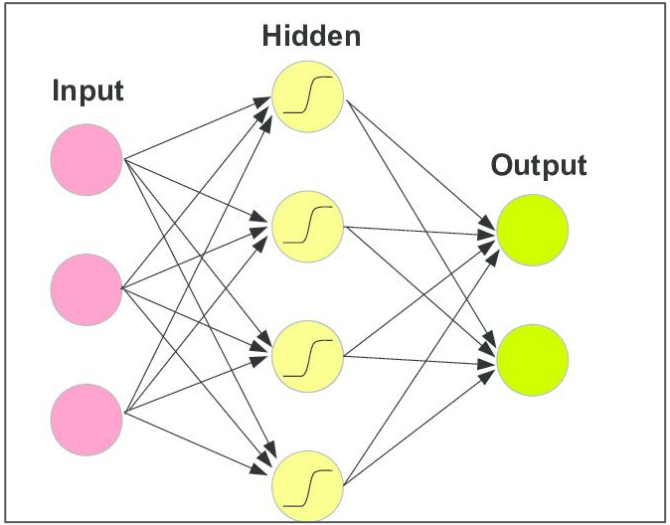
BERTSUM-Simple Classifier

Extractive

A single hidden layer passed to sigmoid classifier. Lr=0.001, Adam optimiser.

$$\hat{Y}_i = \sigma(W_o T_i + b_o)$$

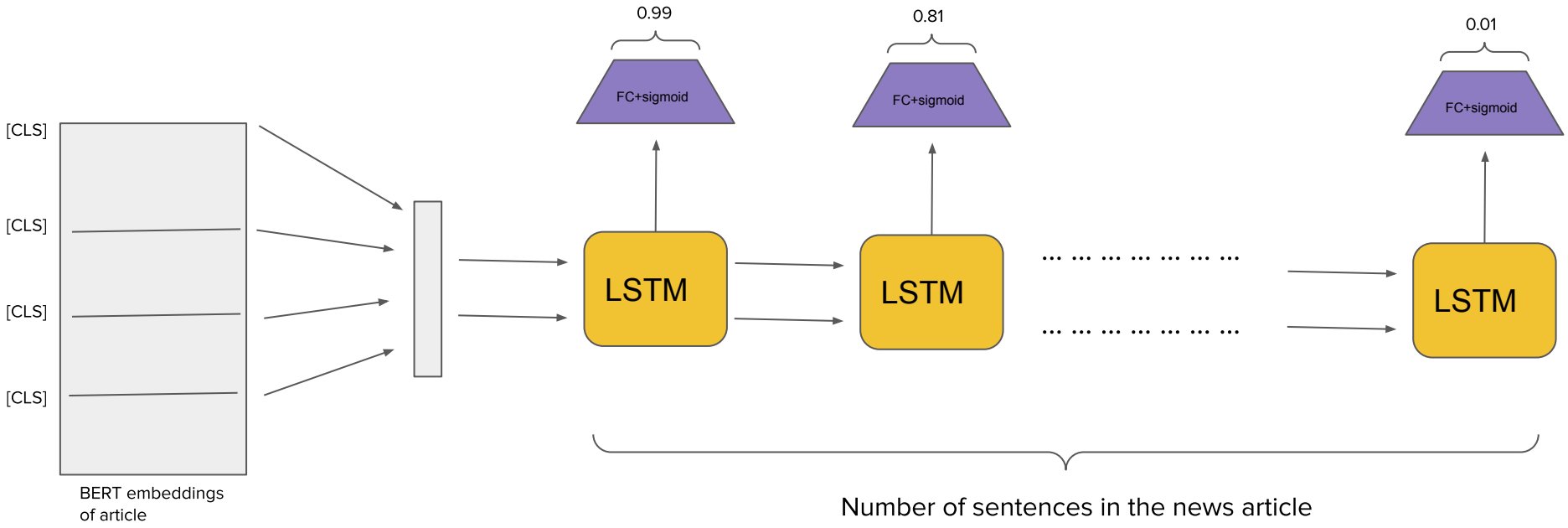
{'**rouge-1**': {'f': **0.66**56233058764613, 'p': 0.6656233108764508, 'r': 0.6656233108764508}, '**rouge-2**': {'f': **0.32**02834466177138, 'p': 0.3202834516177166, 'r': 0.3202834516177166}, '**rouge-l**': {'f': **0.49**999999499999986, 'p': 0.5, 'r': 0.5}}



BERTSUM-RNN

Extractive

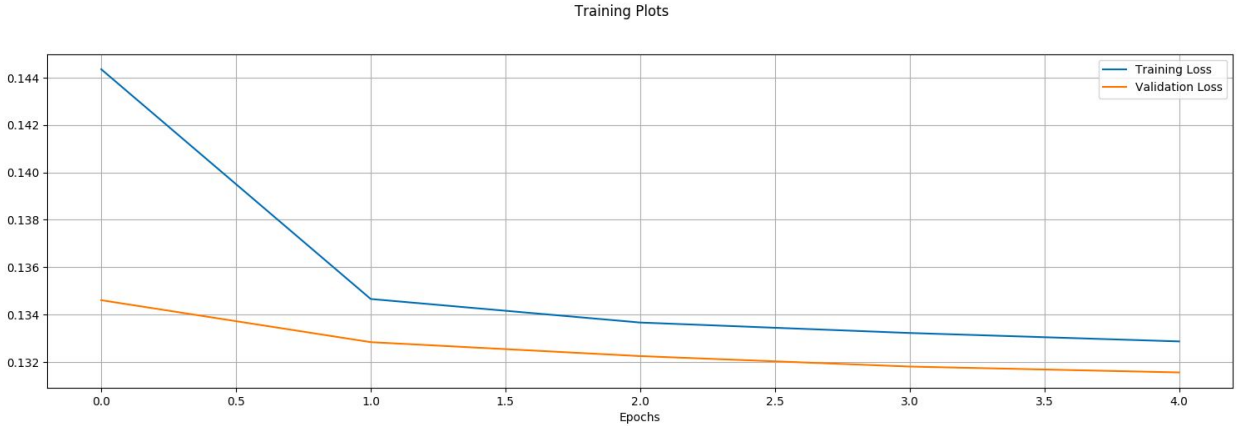
For 5 epochs, model is trained with BERT weights frozen and then next 5 epochs with BERT weights unfrozen
Lr = 0.0001 and dropout is used between FC layers



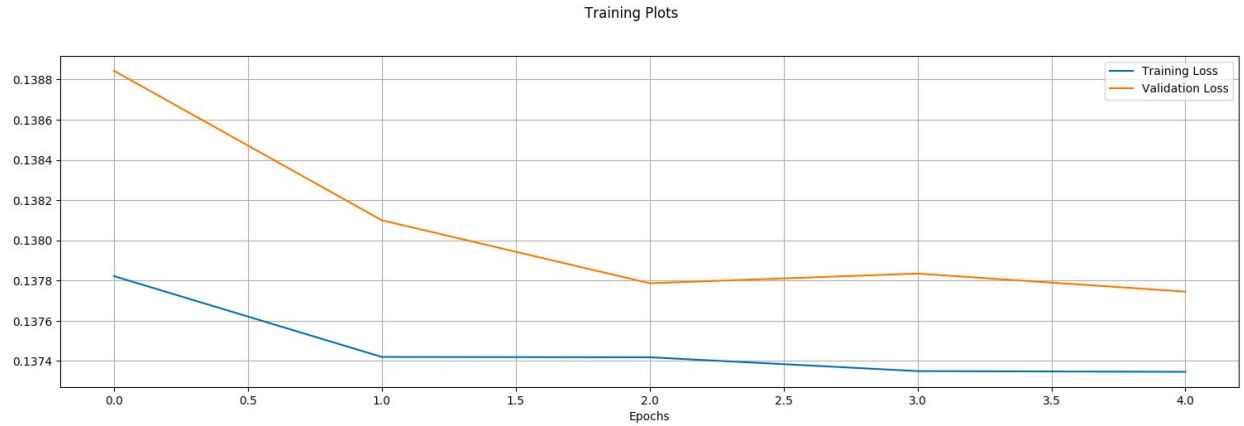
BERTSUM-RNN

Extractive

BERT weights frozen



BERT weights unfrozen



Results for abstractive summarization

Methodology : Bert based model (trained for 2000 samples)

Original summary : कहा राज्यसभा सीट जीतने के लिए बीजेपी हर हथकंडा अपनाने पर आमादागुजरात में कांग्रेस के कार्यकर्ताओं का मनोबल गिराने की चालसोनिया गांधी को एक सेटबैक पहुंचाने की कोशिश

Predicted summary : समाज में माननाम आjan भी लागू हो जाएगायह घटनाएं फशा करते हुए समाज के इस्तीक्षे को ML उम्मीद दी गई हैउन्होंने कहा कि इस सीएम में डीएम की कमुनजी वाले दायर एकमात्र तिहरा मानहाना एक्सजेसी की परिणामस्वरूप मामला सामने आने के लिए चुना गया है वाला लाइसेंस की सारीकाट्यटकणुरुराക्കळे Listenवमेंट में लेख ली जाती

Methodology : Bert based model (trained for 35470 samples)

Original summary : दिल्ली एयरपोर्ट के पास मुसाफिरों से पुलिसवाले बनकर करते थे ठगीपुलिस ने 4 लोगों को किया गए आरोपियों में 2 लोग ईरान के नागरिक हैं

Predicted summary : आईसीसी ने इंग्लैंड के पूर्व कप्तान नासिर हुसैन के उस बयान को पूरी तरह अवांछित करार दिया जिसमें उन्होंने भारतीय कप्तान साबित किया था कि उन्हें कंपनियों से भारतीय क्रिकेटर मौजूदा आमदने पर अधिक ध्यान लगाना चाहिए

Results for abstractive summarization cont...

Methodology	Rouge-1	Rouge-2	Rouge-L
Bert based Encoder	12.5	4.29	9.76
Bert base Encoder (for subset)	12.85	8.51	9.75

The seq2seq model is not able to capture the complexity of Hindi words at all and fails to generate any sensible output sentences. The BERT based model first trained on a subset of articles and later on the large number of articles shown to obtain similar rouge scores but from previous slide it is clear that for a subset of articles model is not able to produce even grammatically correct sentences but when trained for more number of articles, it is shown to produce grammatically correct sentences but totally irrelevant with respect to input sentences at test time.

Results for extractive summarization

LEAD-3 is simply selecting first 3 sentences from news article

It is seen that BERTSUM-RNN outperforms other methods in extractive summarization

	ROUGE-1	ROUGE-2	ROUGE-L
TextRank	49	37	46
LEAD-3 (baseline)	55	49	57
BERTSUM-Simple	66	32	49
BERTSUM-RNN	**67**	**59**	**64**

Contributions

Data preprocessing and cleaning is done by all team members

Individually

1. Dimpy: implemented Bert based abstractive model and evaluations
2. Aamir: implemented TextRank and BERTSUM - Simple Classifier
3. Ritwik: implemented seq2seq model and BERTSUM-RNN

Thank You