AIM-829 NLP
Project
News Article Summarizer and Hinglish Generative Chatbot based on encoder decoder model

Aditya Saraf

April 22, 2025

1 Introduction

This project aims to apply key concepts learned in the course through two practical tasks:

- **Hinglish Generative Chatbot:** Development of a chatbot capable of generating responses in Hinglish, using a conversational dataset sourced online.
- **News Article Summarizer:** Creation of an abstractive news summarizer to generate article highlights, with an additional implementation of extractive summarization using the TF-IDF approach.

2 Assumptions

- All experiments were conducted using Google Colab (Free Tier with T4 GPU).
- To accommodate limited computational resources, small datasets were used for both tasks. Larger datasets were avoided to prevent excessive runtime and resource exhaustion, which led to repetitive or poor-quality outputs.

3 Dataset Details

3.1 Hinglish Chatbot

- **Dataset:** A conversational dataset comprising dialogues between friends, sourced from:
  `https://www.kaggle.com/datasets/siddikisahil47/conversation`
- **Preprocessing:** Due to resource constraints, the number of conversations was limited. We used pretrained FastText embeddings (300 dimensions) and also experimented with Keras Embedding layers.

3.2 News Article Summarizer

- **Dataset:** The CNN/DailyMail news summarization dataset containing news articles and corresponding highlights was used. Source: `https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-d data`
- **Preprocessing:** Articles of lengths between 200–300 words and highlights of 80–150 words were selected. Redundant information such as author names, media references, and publisher details was removed. We used pretrained GloVe embeddings (100 dimensions) and Keras Embedding layers for input representation.

4 Methodology

4.1 Task 1: Hinglish Chatbot

- **Model Architectures:** We experimented with various sequence modeling techniques:
  - **Stacked RNN:** Baseline architecture with limited memory.
  - **LSTM:** Improved version with gating mechanisms to capture longer dependencies.
  - **LSTM with Attention:** Helped the model focus on relevant parts of the input.
  - **Transformer:** Utilized self-attention for parallel processing and better context handling.
- **Training Strategy:**
  - Due to the small dataset size, overfitting was a common issue.
  - **Early stopping** was implemented to halt training once the validation loss stopped improving, thereby preventing the model from memorizing the training data.

4.2 Task 2: News Article Summarizer

- The overall methodology was similar to Task 1 in terms of preprocessing and model architecture.
- Additionally, a TF-IDF based extractive summarization approach was implemented to identify important sentences from the article.
- The TF-IDF method calculates term frequency-inverse document frequency scores for words and selects sentences with the highest cumulative scores as summaries.

5 Experimental Results

5.1 Task 1: Hinglish Generative Chatbot

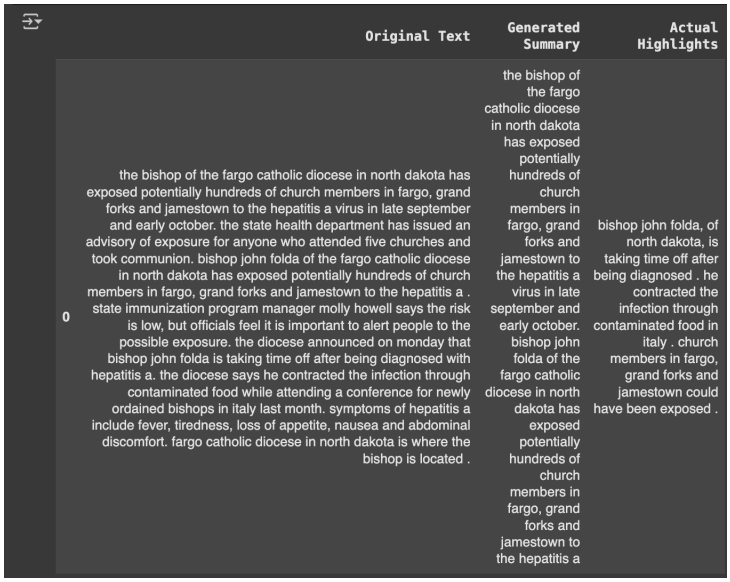
We tested the chatbot's responses to two training queries: *"kaise ho"* and *"Hey Radhika! Kaisi ho?"*, using various model architectures. The results are as follows:

- **RNN:** *"haan, mujhe bhi bhi bhi bhi bhi bhi bhi bhi bhi bhi bhi bhi..."*
  The RNN model suffered from severe repetition, likely due to its limited memory and the small training dataset.
- **Transformer:** *"anything tere tere tere tum there there there moment moment soch soch..."*
  The response lacks coherence, reflecting the model's inability to learn meaningful context from the limited data.
- **LSTM:** *"main bhi theek hoon, tum kya sochti ho"*
  A more coherent output, showing that LSTM handled sequence dependencies better than the RNN.
- **LSTM with Attention:** *"main bhi theek hoon, bas thoda busy ho rahi hoon"*
  This was the most relevant and fluent response, indicating that the attention mechanism helped the model focus on important words and context.
- **General Observation:** Inputs not present in the training set resulted in absurd or irrelevant outputs, highlighting the model's limited generalization ability due to insufficient training data.

5.2 Task 2: News Article Summarizer

We tested the summarization models using the same training data. Due to data limitations and computational constraints, the quality of summaries remained poor.

- **TF-IDF:** The summary generated using the TF-IDF approach is shown in Figure 1.



	Original Text	Generated Summary	Actual Highlights
0	the bishop of the fargo catholic diocese in north dakota has exposed potentially hundreds of church members in fargo, grand forks and jamestown to the hepatitis a virus in late september and early october. the state health department has issued an advisory of exposure for anyone who attended five churches and took communion. bishop john folda of the fargo catholic diocese in north dakota has exposed potentially hundreds of church members in fargo, grand forks and jamestown to the hepatitis a . state immunization program manager molly howell says the risk is low, but officials feel it is important to alert people to the possible exposure. the diocese announced on monday that bishop john folda is taking time off after being diagnosed with hepatitis a. the diocese says he contracted the infection through contaminated food while attending a conference for newly ordained bishops in italy last month. symptoms of hepatitis a include fever, tiredness, loss of appetite, nausea and abdominal discomfort. fargo catholic diocese in north dakota is where the bishop is located .	the bishop of the fargo catholic diocese in north dakota has exposed potentially hundreds of church members in fargo, grand forks and jamestown to the hepatitis a virus in late september and early october. bishop john folda of the fargo catholic diocese in north dakota has exposed potentially hundreds of church members in fargo, grand forks and jamestown to the hepatitis a	bishop john folda, of north dakota, is taking time off after being diagnosed . he contracted the infection through contaminated food in italy . church members in fargo, grand forks and jamestown could have been exposed .

Figure 1: Output generated using the TF-IDF summarization approach.

- **LSTM with Attention:** *"The outbreak of contagious liver disease hepatitis disease disease disease disease disease disease disease disease disease disease disease disease disease the disease is not not not not be able to be a result of the virus."* The output demonstrates poor grammar and heavy repetition, caused by overfitting on a small dataset and insufficient learning.

6 Discussion

- **Reason for Repetitive and Illogical Output:** The limited size of the dataset restricted the model's learning capacity, leading to repetitive or incoherent outputs during inference.
- **Model Architectures Explored:** Various architectures were experimented with across both tasks, including Stacked RNNs, LSTMs, LSTMs with Attention, Transformers, and the TF-IDF-based extractive summarization method for news articles.
- **Future Work:** Potential improvements include training on larger and more diverse datasets, incorporating deeper architectures, and performing extensive hyperparameter tuning to enhance performance and generalization.

7 Conclusion

The primary objective of this project was to apply and implement various concepts learned during the course through two tasks: building a Hinglish generative chatbot and developing a news article summarizer.

Despite resource and dataset limitations, we successfully experimented with multiple deep learning architectures such as RNNs, LSTMs, and Transformers. We also explored traditional approaches like TF-IDF for extractive summarization.

While the models showed basic functionality, there is significant scope for improvement with larger datasets, advanced architectures, and further optimization. This project provided valuable hands-on experience in applying NLP techniques to real-world problems.

8 Team-wise Contribution

As the sole member of the team, I independently carried out all aspects of the project, including data preprocessing, model development, experimentation, and documentation. I also utilized tools like ChatGPT and DeepSeek for guidance and troubleshooting during the implementation phase.