DATS-6312 : Natural Language Processing
Professor : Dr. Amir Jafari
Group Number : 4

# News Generation using Different Advanced Model -Classical and Deep learning Approaches: A Comparative Study

....

Apoorva Reddy Bagepalli [G30787948]
Modupe Fagbenro [G32371634]
Keerthana Aravindhan [G34700275]
Aaron Yang[G33304744]

....

# Agenda

- Introduction
- Problem Statement
- Dataset
- Models used
- Result
- Conclusion


# Problem Statement

**Problem Statement :**

- News articles are produced at an overwhelming rate, making it difficult to manually generate unique, high-quality descriptions for each article.
- Consumers expect brief outputs that capture the essence of news articles, providing quick insights without losing accuracy or context.

**Objective:** Conduct a comparative analysis of the models to assess their ability to generate concise, accurate, and meaningful news descriptions.

**Importance:**

- Enhanced User Experience: Enables quicker consumption of news by providing short, insightful descriptions.
- Optimized News Curation: Helps aggregate, curate, and deliver relevant news across platforms.
- Scalability: Automation enables scaling the process of generating vast amounts of daily news content.

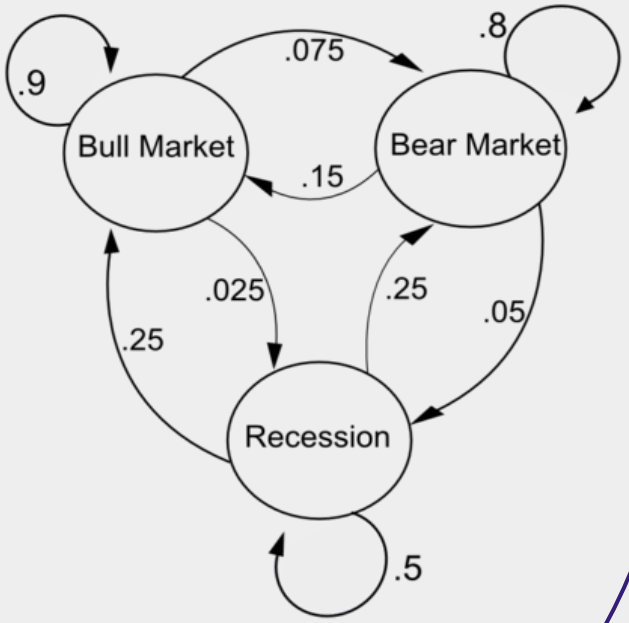

# What is our data

**InShorts News Data**
**BBC Articles Dataset**
**CNN/DailyMail Dataset**
**Newsroom Dataset**

**Combined Data: Size:** 3GB, **Content:** Main article text only

**Scope:** Provides diverse coverage of topics and writing styles for robust training.

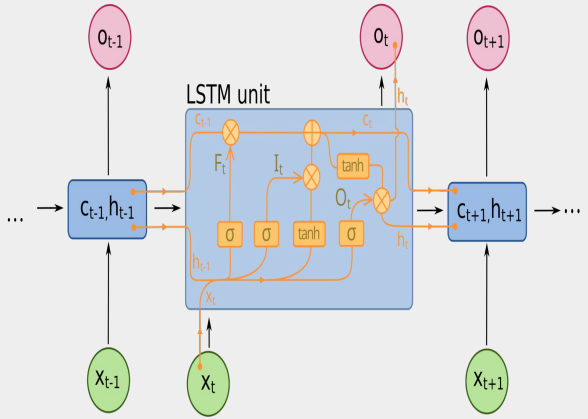
# Markov Chain Model




- **Markov Assumption:**
  The probability of a word depends only on the previous n-1 words (for n-gram models).
- **Key Concepts:**
  1. **States:** Words or n-grams in the text.
  2. **Transitions:** Probabilities of moving from one state to another.
  3. **Higher-Order Chains:** Use n-grams for longer context.
- **Text Generation Process:**
  1. Extract states (words/n-grams) from dataset.
  2. Calculate transition probabilities based on frequencies.
  3. Generate text by randomly transitioning through states.
- **Chapman-Kolmogorov Equation:**

$$P_{ij}^{(n)} = \sum_k P_{ik}^{(n-1)} . P_{kj}$$

# LSTM + Attention




**Overview**

**LSTM - Attention**

- **LSTM** is a specialized neural network architecture with memory cell structure with controlled information flow- Long-term dependency & vanishing gradient problems
- **Attention:** Dynamically focuses on relevant parts of input sequence by computing weighted importance scores,allowing direct access to any part of the input when generating output."

$ft = \sigma(Wf \cdot [ht\text{-}1, xt] + bf)$ # Forget Gate

$it = \sigma(Wi \cdot [ht\text{-}1, xt] + bi)$ # Input Gate

$Ct = ft * Ct\text{-}1 + it * \tilde{C}t$ # Cell State Update

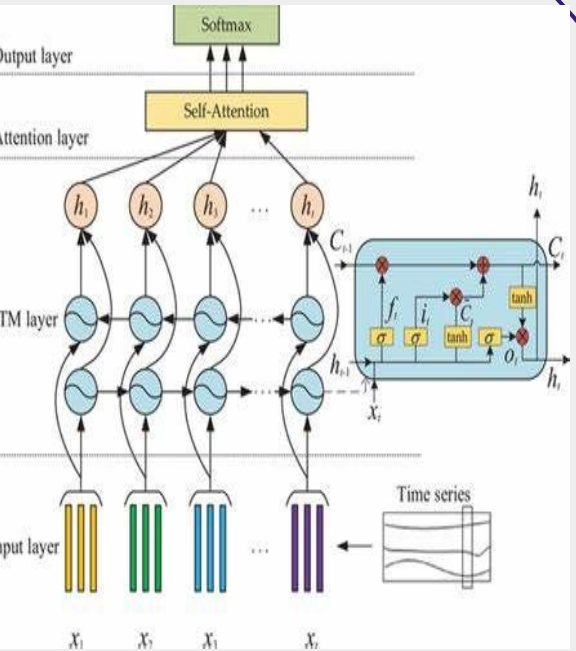
$ht = ot * \tanh(Ct)$ # Output

**Attention Mechanism:**

$\text{Attention}(Q,K,V) = \text{softmax}(QK^T/\sqrt{d\_k})V$

Where:

- QK^T computes similarity scores
- √d_k scales to prevent exploding gradients
- Softmax normalizes scores to probabilities




# LSTM + Attention

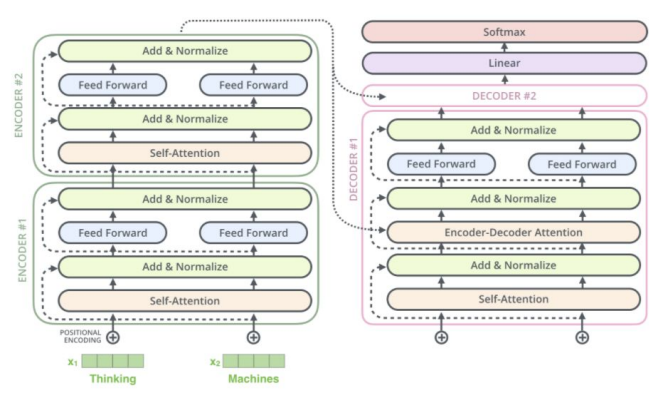
The LSTM-Attention combination creates a powerful architecture where LSTM maintains sequential memory while attention provides selective focus on relevant input elements, resulting in a model that can both remember long-term context and precisely pinpoint important information

**How LSTM and attention mechanisms work:**

The mechanisms interact with weights/biases:

- LSTM: Has weight matrices and bias vectors for each gate
- Attention: Learns projection matrices for Q,K,V transformations

# T5




**Key Features**

- **Unified Framework**: Simplifies task implementation by converting inputs and outputs into text strings.
- **Encoder-Decoder Architecture**:
  - **Encoder**: Generates contextual representations.
  - **Decoder**: Produces output text based on the input.
- **Transfer Learning**: Pre-trained on **Colossal Clean Crawled Corpus (C4)**, then fine-tuned for specific tasks.

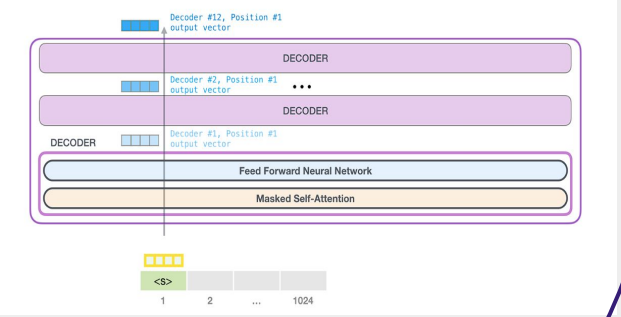
**Model Variants**

- **Scalable Sizes**: Ranges from T5-Small (60M parameters) to T5-11B (11B parameters), adapting to different resource requirements.

**Training Methodology**

- **Pre-training**: Predicts missing/corrupted text to understand language structures.
- **Fine-tuning**: Adapts to specific tasks with task-specific input-output pairs.

# GPT-2


**Key Features**

- **Transformer-Based Autoregressive Model:** Generates context-aware text over long sequences.
- **Decoder-Only Architecture:** Employs self-attention for capturing long-range dependencies.
- **Tokenizer:** Byte Pair Encoding (BPE) for efficient subword representation.

**Fine-Tuning for News Content**

- Trained on a curated dataset of news articles for domain-specific generation.
- New tokenizer and special tokens added to improve context representation.
- Training the model from scratch to learn the domain.

**Development Highlights**

- **Causal Language Modeling (CLM):** Predicts the next token in a sequence.
- **Nucleus Sampling (Top-p):** Enhanced diversity and relevance in text generation.
- Hyperparameters optimized to match domain-specific vocabulary.

# Results

| Models | Perplexity |
| --- | --- |
| Markov Chain | 22.16 |
| LSTM + Attention | 1.76 |
| GPT-2 | 1.69 |
| T5 | 16.44 |

# Model Generation-Result

Text Generation for all the models

***Markov Chain Output Generation:***

*Prompt: Covid*

*Generated text: the cbi to return if she wins outright aggression drawn by a housing estate in vitry sur seine south east asia earthquake and tsunami the disaster and more esops the italian police had to break the close bond tom receives instructions from congress leader amarinder singh on thursday rolled out the e toll system for collecting toll not exceeding annas from certain factions to separate from that being stunned is as painful i am pleased and hopefully ill come back to life imprisonment in the massacre an unusual space in midtown manhattan where he performed a free concert in oslo the conclusions*

# Generated- Result

**LSTM_Attention Output Generation:**

**Prompt:** *Business news*

***Generated text:*** *global irreversible stamps serena colorful burton salute shameful the dumb cheltenham the proclaimed the mtv paulo dumb dumb paramount delivers cheltenham paulo irreversible newman echo newman certification prostitutes khashoggi prostitutes echo miscarriage newman the stamps to farrell and paramount farrell shameful vest newman tass irreversible newman failings empathy hagel lava colorful diners misguided ambulances stamps hagel ambulances paramount depressed certification depressed newman chrysler delivers failings chrysler colorful irreversible the diners chrysler certification certification ond farrell farrell chrysler echo misguided salute misguided mexicans mtv paramount submerged cheltenham the hagel mexicans . molecular to shameful empathy colorful stamps khashoggi salute ond failings chrysler thumb paulo fashioned ond misguided colorful hagel and prostitutes diners rigorous delivers paramount . khashoggi certification diners junction chrysler farrell and miscarriage ond candles empathy diners fashioned christine fashioned the failings shameful mexicans paramount paramount to stamps mtv thumb farrell burton vest hagel and echo ambulances empathy and delivers*

# Generated -Result

***T5 Output Generation***

*Prompt: Trump*

*Generated text:*

*=== Generated Text 1 ===*
*Trump has fueled an american president white wes more republic republic presidential presidential campaign you are so Trump on a you. republican presidential presidential us wash hill go Trump s us sen.*
*=== Generated Text 2 ===*
*you Trump president president oba hill wash administration is hes secret Trump*
*=== Generated Text 3 ===*
*Trump on wes Trump Trump you do not have*
*=== Generated Text 4 ===*
*Donald Trump*
*=== Generated Text 5 ===*
*republic us presidential presidential presidential campaign Donald Trump en immigration the us american president who is tun Trump his next presidency white white democratic presidential hill.*

# Generated -Result

***T5 Output Generation:***

*Prompt: President Trump is*

*Generated text:*

*=== Generated Text 1 ===*
*president Donald Trump is the president os and not us president Donald Trump is speaking in an emb new speech president Barack Obama is back in wash his own.*
*=== Generated Text 2 ===*
*president president Donald Trump accompanied us us president Barack Trump by speaking on his as president you presidential confirmation you.*
*=== Generated Text 3 ===*
*president Donald Trump os presidential presidential election president President Donald oning president is the president of president ros president is in your presidency and president al on a speech speech us president Barack h.*
*=== Generated Text 4 ===*
*president Donald Trump is president and his former president president white house speaking on friday as he will president president.*
*=== Generated Text 5 ===*
*president Donald Trump is on holiday the president who president s us President Barack Donald Donald Trump greeted us president white white american president president Trump has his family message fueled by a you.*

# Generated -Result

***T5 Output Generation:***

*Prompt: President Trump said that China is*

*Generated text:*

=== Generated Text 1 ===
us president elect Donald Trump has said that it would be hard for china to accept the possibility of becoming a new member of the worlds muslim state as it was known by some as a major syrian government in syria on tuesday.
=== Generated Text 2 ===
chinas president has said that he is in the prime minister to ensure that a foreign dollar deal is not tied to the state.
=== Generated Text 3 ===
us president Donald Trump told cnn that china is going to have a rocky economy and will continue to operate.
=== Generated Text 4 ===
uttarakhand president Donald Trump has said that china is likely to be forced to leave without a diplomatic relationship with the us and will pay all necessary foreign aid as he says.
=== Generated Text 5 ===
president obama said china has built an eu trade center in south africa and it has decided to host a nuclear nuclear event in the country.

# Generated-Result

***GPT-2 Output Generation:***

*Prompt: Covid*

*Generated text:  19 More than 500,000 Americans are living with COVID-19. The US Centers for Disease Control and Prevention says the number of Americans with COVID-19 is at its highest level in more than a year. This includes people who have been exposed to certain types or conditions like vaping — which can lead... [Read more...] In August, the CDC announced that it had identified a new group on coronavirus with high levels of COVID-19. "As we continue our work with our community partners around the country to protect our citizens from this outbreak," the agency said,. "We continue working together as a community to prevent further spread of this virus.". The CDC said the number of new cases in the United States has risen by more than 6 times since the start of the pandemic, with the number now at a record high. "Despite all efforts being made to reduce the number….we remain committed to keeping Americans safe, and to doing everything we*

# Demo

app · Streamlit

localhost:8501

Relaunch to update

Mail DS articles DS School DSA Job Tutor SRLC

Deploy

Navigate to:

Home

# News Article Text Generation

In the digital age, the consumption of news has shifted dramatically towards online platforms, where brevity and clarity are highly valued. With the overwhelming amount of content being produced daily, generating concise and accurate descriptions for news articles has become a significant challenge. This project focuses on automating the process of news text generation using advanced Natural Language Processing (NLP) models.

The problem selected for this project is generating concise and accurate descriptions for recent news articles. In today's fast-paced digital environment, news consumers often prefer brief outputs that capture the essence of articles without delving into their entirety. This task addresses the challenge of automatically generating such descriptions, aiding in news aggregation, curation, or providing meaningful summaries for platforms. Given the growing volume of news content, this problem is ideal for automation using advanced text generation models.

Our project explores and implements various text generation methods to address this issue, ranging from classical Markov Chains to modern deep learning techniques like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and transformer-based architectures such as GPT-2. By applying these models to datasets of recent news articles, we aim to provide a comparative analysis of their performance in generating meaningful and high-quality outputs. The report outlines the datasets, methodologies, and evaluation metrics such as ROUGE, BLEU, and perplexity. It also discusses key findings, challenges faced, and potential future enhancements to improve the quality of automated news text generation.

## Team Members:

Apoorva Reddy Bagepalli

Aaron Yang

Modupeola Fagbenro

# Key findings/Challenges

## Dataset Preparation and Preprocessing

**Challenge:** Cleaning and formatting a diverse dataset of recent news articles.

**Impact:** Errors during preprocessing led to inconsistencies, requiring multiple iterations to refine the dataset for model training.

## Model Selection and Implementation

**Challenge:** Choosing between simple models like Markov Chains and more complex models like LSTM, GPT-2 and T5

**Impact:** Simpler models lacked contextual understanding, while complex models required extensive fine-tuning and had a steep learning curve.

## Computational Limitations

**Challenge:** Limited hardware resources for training advanced models (e.g., lack of GPUs/TPUs).

**Impact:** Training time was extended, limiting the scope of experiments and model iterations.

## Integrating Outputs into a Cohesive Report

**Challenge:** Unifying results from models with different strengths and weaknesses.

**Impact:** Required additional effort to synthesize insights and present findings in a clear and coherent manner.

# Conclusion And Summary

**Our study, "News Generation using Different Advanced Model** - Classical and Deep Learning Approaches: A Comparative Study," addressed several key aspects of text generation:

**1. Classical Approaches:**

- Implementation of traditional text generation methods- Markov Chain
- Analysis of their limitations and computational challenges
- Understanding the baseline performance metrics

**2. Deep Neural Networks: LSTM + Attention , GPT2, T5.**

- Development and Implementation of advanced architectures for text generation
- Comparative result

This **comprehensive exploration has reinforced our understanding of the concepts taught in class** while providing practical experience in implementing and evaluating different text generation approaches. The project has highlighted both the evolution of text generation techniques and the ongoing challenges in producing high-quality, coherent news content.

# Thank you!

# Questions?