# Automated Summarization of Urdu Jang News Articles


**Session: 2022-2026**

## Project Supervisor

Dr. Samyan Qayyum Wahla

## Project Member

Muhammad Asghar 2022-CS-73
Malik Muhammad Saad 2022-CS-94

---

Department of Computer Science

**University of Engineering and Technology, Lahore Pakistan**

# Contents

# 1 Introduction

It is crucial to keep up with the latest news in the fast-paced world of today. But the abundance of knowledge on the internet may be debilitating. This problem can be solved by summarization systems, which reduce long articles to brief summaries so that readers can rapidly understand the essential points. The goal of this project is to create an intelligent system that would automatically summarize news stories from Jang News, one of the top news platforms in Pakistan. With the help of web scraping, natural language processing (NLP) methods, and an interactive user interface, this system makes it easier for end users to extract and summarize news stories.

This work is novel since it focuses on news summarization in the Urdu language, which poses particular difficulties because of resource constraints and grammatical intricacies. By tackling these issues, this initiative seeks to enhance the accessibility of news for Urdu-speaking audiences.

# 2 Methodology

## 2.1 Data Collection and Preprocessing

**Web Scraping:**

The first step of the project is to scrape news information from Jang News' official website (`https://jang.com.pk/`).

The webpage's HTML content is parsed to retrieve pertinent information using Python tools like BeautifulSoup and requests.

In particular, the website's main narrative area is where all <a> tags (hyperlinks) are obtained. These links lead to specific news stories.

**Content Extraction:**

The <p> elements, which contain the articles' main text, are extracted for every hyperlink after the corresponding webpage has been retrieved.

For clarity and organization, the extracted paragraphs are saved in JSON format, with each paragraph being numbered (e.g., p1, p2, etc.).

**Data Storage:**

The content of each article is saved as a distinct JSON file in a directory called `jang_articles`. Later processing and retrieval are made easier by this arrangement.

## 2.2 Summarization

**Model Selection:**

The Hugging Face Transformers library's mT5_multilingual_XLSum pre-trained multilingual summarization model is employed. This model is perfect for Jang News stories because it is optimized for summarizing text in several languages, including Urdu.

**Handling Input Size Constraints:**

The model can produce outputs of up to 84 tokens and has a maximum input token limit of 512 tokens. Longer articles are handled by breaking the content up into digestible sections (each no more than 512 tokens) and summarizing each section separately.

**Combining Summaries:**

Concatenating summaries of separate sections creates a coherent synopsis of the complete article.

### 2.3 User Interface Development

**Streamlit Integration:**

A Python-based web application development framework called Streamlit is used to provide an intuitive user interface. Users can see dynamically created summaries of articles saved in the `jang_articles` directory using the interface.

**Overcoming the Limitations of Colab:**

Streamlit apps are not natively supported by Google Colab. To solve this, a public URL for the application is provided by using ngrok to expose the local Streamlit server to the internet. Through this URL, users can access the summarizing system and engage with the summaries in real time.

## 3 Final Product

The following functions are efficiently carried out by the system:

- **Data Extraction:** Articles from Jang News are successfully scraped and stored in structured JSON files. With error handling for situations where <p> tags are missing, it gracefully manages missing or distorted data.
- **Summary:** Produces succinct and logical summaries of both short and long items. The summaries provide a high-level overview of the topic while removing unnecessary details.
- **User Interaction:** Viewing article summaries is made simple for users by the Streamlit-based user interface. Even from a distant server (Google Colab), the app may be accessed with ease thanks to ngrok integration.

## 4 Model Performance Outcomes

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which compares summaries to reference summaries created by humans, is used to assess the summarizing model's performance. The following are the model's findings for articles written in Urdu:

- ROUGE-1: 39.5579
- ROUGE-2: 18.3733
- ROUGE-L: 32.8442

These findings show that, despite the difficulties caused by the scarcity of resources for Urdu language processing, the model does a respectable job of summarizing Urdu text.

## 5 Conclusion

From data collection to interactive visualization, this project shows a whole pipeline for automating the summary of Urdu news items. Among the major accomplishments are:

- Putting in place a reliable web scraper to retrieve Jang News article data.
- Employing a cutting-edge, pre-trained, multilingual NLP model to efficiently summarize content.
- Using techniques like Streamlit and ngrok to overcome infrastructure constraints and produce a completely working user experience.

Building intelligent summarizing tools that may be used to different news platforms or domains is made possible by the system's scalable base. Future enhancements can consist of:

- Improving the summarization approach to manage domain-specific content or more complex textual properties.
- Incorporating support for several languages or dialects to expand accessibility.
- Incorporating functions like search and keyword filtering and making the UI mobile-friendly.

In conclusion, this project integrates online scraping, advanced natural language processing, and user interface development approaches to produce a useful and effective summarization tool designed for Jang News readers, with an emphasis on the Urdu language.