



CSI3019 – ADVANCED DATA COMPRESSION TECHNIQUES

REAL-TIME CHAT APPLICATION WITH ON-THE-FLY TEXT COMPRESSION SUPPORTING EMOJI AND MULTILINGUAL MESSAGES

TEAM MEMBERS:

1. LOKESHWARI G – 22MID0033
2. ANU SWATHI V R – 22MID0257
3. SUBASHREE S – 22MID0269



ABSTRACT

This project proposes a hybrid text compression system combining a persistent JSON-based dictionary with adaptive algorithms like RLE, Huffman, LZW and arithmetic coding. The system first checks for a dictionary matches to replace frequent words, and then applies adaptive methods for remaining data. Unlike traditional approaches, the dictionary persists across sessions, improving compression efficiency for repetitive datasets such as chat logs and documents. The design enhances storage optimization, transmission speed, and long-term compression performance, making it suitable for applications in real-time messaging, archival storage and for efficient message transfer and communication.

EXISTING METHODOLOGY

TRADITIONAL AND RESEARCH METHODS:

Most of the research papers explored with one single for compression the texts. Many of them explored only with normal text with single Language without emoji's. But our project is to explore texts with emoji's which supports multiple languages.

1. RLE (Run Length Encoding): Compresses repeated symbols but inefficient for natural text.
2. Huffman Coding: Variable-length encoding, widely used but not adaptive to dynamic data
3. LZW (Lempel-Ziv-Welch): Dictionary-based, effective for repeated sequences but resets every session.

Base Paper:

Wan, L., Alpcan, T., Kuijper, M., & Viterbo, E. (2024). Lightweight Conceptual Dictionary Learning for Text Classification Using Information Compression. IEEE Transactions on Knowledge and Data Engineering, 36(12), 8711-8717.

LIMITATIONS:

1. No persistence: Dictionaries reset after each session.
2. Classification-centric: Focus on accuracy; not compression efficiency
3. Not real-time: Limited applicability to streaming data like chats/logs.
4. Lack in Multiple Language Compression with Emoji's.

PROPOSED METHODOLOGY

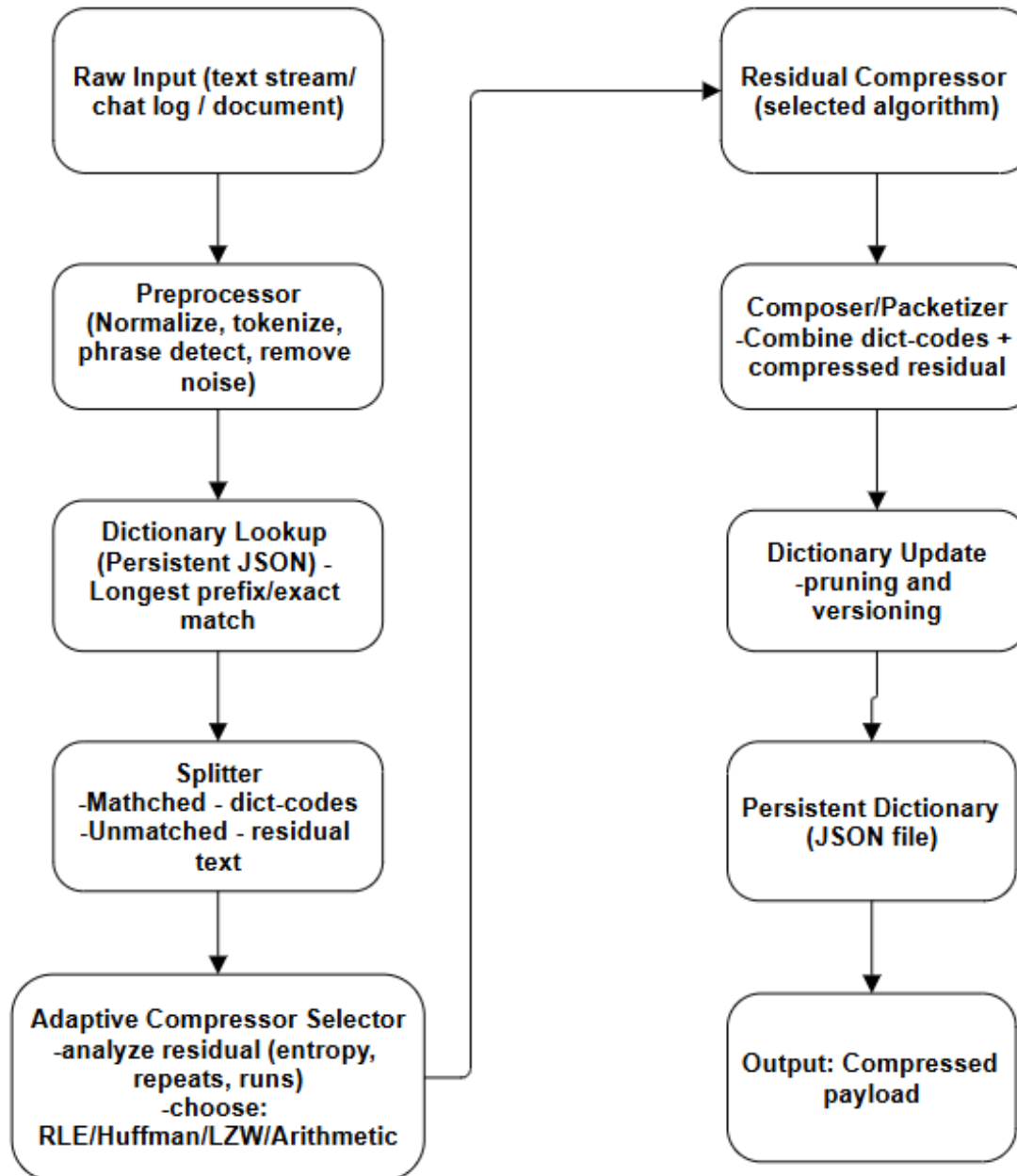
Hybrid Dictionary + Adaptive Compression

- **Persistent JSON-based Dictionary**
 - Stores frequently used words across sessions.
 - Ensures continuity and efficiency for repetitive text (e.g., chats, logs).
- **Adaptive Compression Algorithms**
 - RLE, Huffman, LZW, Arithmetic coding applied after dictionary lookup.
 - Balances compression ratio and processing speed.
- **Hybrid Workflow**
 - Step 1: Check input against persistent dictionary.
 - Step 2: Replace matches with compact codes.
 - Step 3: Apply adaptive compression on remaining data.

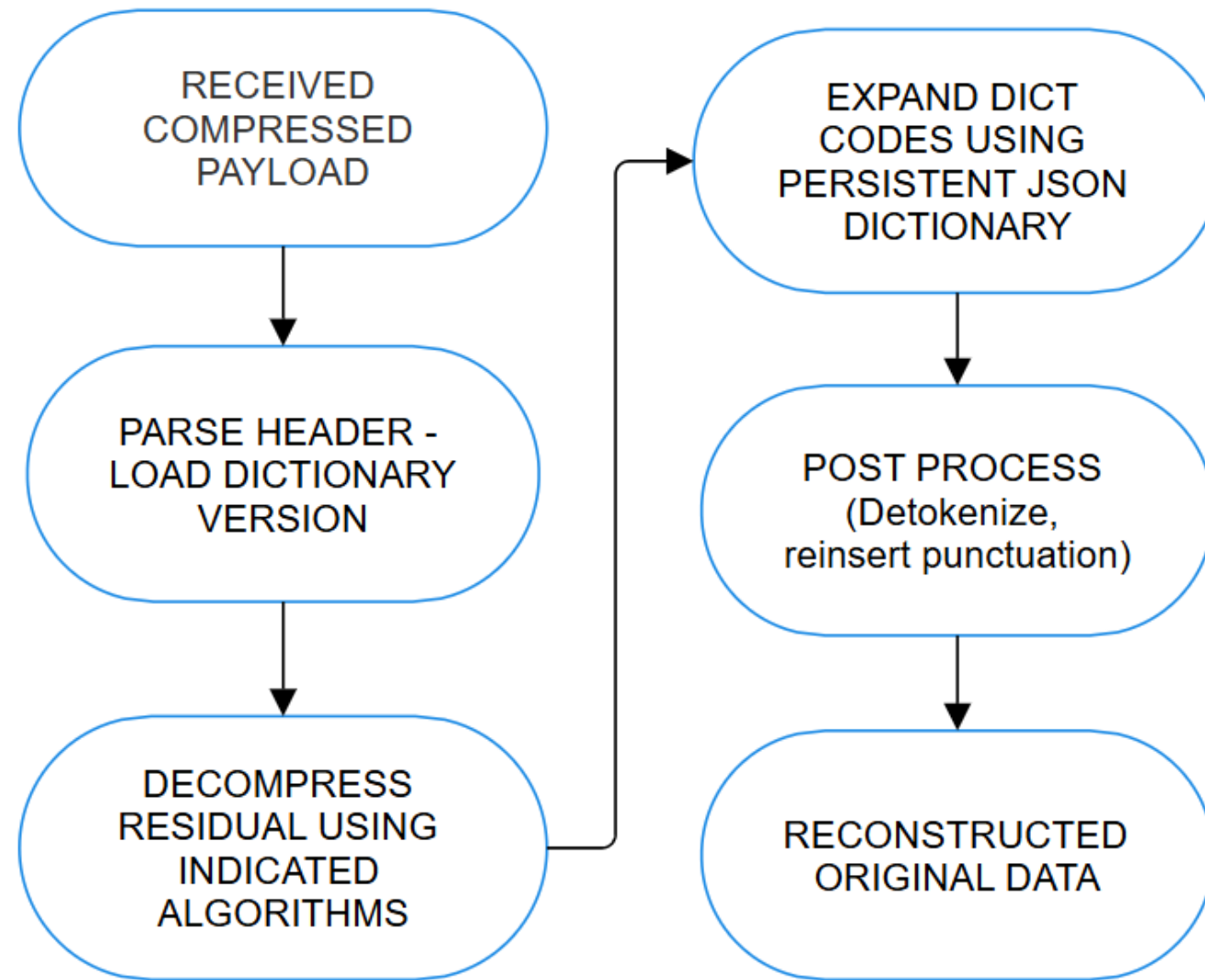
Advantages Over Existing Methods

- **Persistence:** Dictionary retained across multiple sessions.
- **Efficiency:** Combines dictionary + adaptive methods for better compression.
- **Real-Time Ready:** Optimized for chat, documents, and streaming text.
- **Balanced Performance:** Improves both compression ratio and speed.

FLOW OF THE PROJECT



DECOMPRESSION (REVERSE FLOW)





THANK YOU