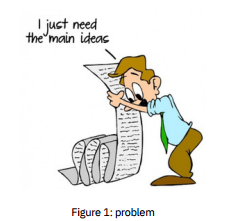
**Extractive Text Summarization using Unsupervised Techniques**

**1. Problem Statement:**



In our daily interactions with textual content, we often encounter detailed documents that are time-consuming to read thoroughly. Both individuals and businesses face challenges when dealing with lengthy documents. For businesses, this might mean hiring a technical writer to condense the material, while for readers, a lengthy document can be daunting and difficult to comprehend fully. To address this, a mechanism that efficiently extracts key information could significantly streamline the process, making it easier to grasp the essential messages without loss of critical details.

**2. Motivation:**

This issue becomes particularly critical when managing extensive collections of text, such as those handled by major data companies like Facebook and Google. For instance, Google demonstrates its efficiency with search results delivered in fractions of a second, yet the sheer volume of information remains largely inaccessible to the average user. Notably, there is a significant gap in resources for the Urdu language, which is widely spoken across the subcontinent. Our project aims to bridge this gap by focusing on Urdu, leveraging our native understanding of the language to enhance text summarization tools.

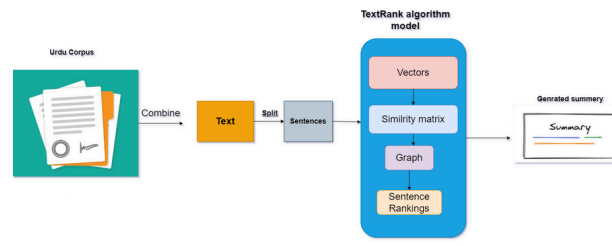
**3. Background:**

Urdu language, with its rich blend of Arabic, Persian, and South Asian influences, presents unique challenges in text processing due to its complex morphology and syntax. The absence of capitalization and diacritical marks makes text analysis even more challenging, emphasizing the need for specialized preprocessing techniques to effectively handle Urdu text data.

**4. Related Work:**

Research in text summarization isn’t new and has utilized various algorithms like LexRank, Luhn, and Latent Semantic Analysis. However, our focus is on enhancing the extractive summarization technique, especially for underrepresented languages like Urdu, which lacks sufficient preprocessed data and research.

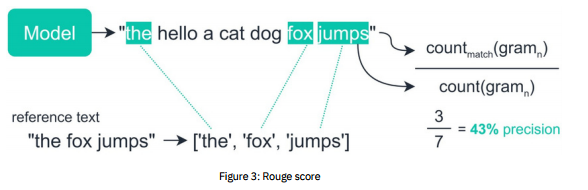
**5. Proposed Work:**



Our methodology involves breaking down Urdu text into sentences and using a graph-based approach with the TextRank algorithm to evaluate and rank these sentences based on their connectivity and importance within the text. This process aims to efficiently extract the essence of the documents, producing concise summaries that are both informative and accessible.

**6. Evaluation Methodology:**

Due to the scarcity of comprehensive data for Urdu, we plan to utilize the ROUGE method, a standard for evaluating text summarization quality through intrinsic and extrinsic measures.



**7. Hypothesis:**

We hypothesize that this summarization tool will be beneficial in educational settings by helping students and teachers quickly identify key information in textbooks, thus making learning more efficient. Additionally, it could aid content and technical writers in producing focused and relevant content more effortlessly.

**References:**

[1] Urdu Summary Corpus. <https://aclanthology.org/L16-1128.pdf>.

[2] AMEX AI-Labs: An Investigative Study on Extractive Summarization of Financial Documents

https://aclanthology.org/2020.fnp-1.23.p df.

[3]CNLP-NITS @ LongSumm 2021: TextRank Variant for Generating Long Summaries

https://aclanthology.org/2021.sdp-1.13.p df.

[4]Hindi History Note Generation with Unsupervised Extractive Summarization

https://aclanthology.org/2020.aacl-srw.7.pdf.