Statement of Purpose

Shubham Agrawal 22110249

shubham.agrawal@iitgn.ac.in | IIT Gandhinagar

Preliminary Approach: News Headline Clustering

Tools and Technologies

1. **Programming Language**: Python
2. **Libraries and Frameworks**:
    - Natural Language Processing: Hugging Face Transformers, SpaCy
    - Clustering and Similarity Analysis: Scikit-learn, TensorFlow, or PyTorch
    - Data Handling: Pandas, NumPy
3. **Web Scraping**: BeautifulSoup, Scrapy for dataset creation
4. **Visualization Tools**: Matplotlib, Seaborn, t-SNE, PCA

AI Models and Methodologies

1. **Data Preprocessing**:

    - Scrape news headlines from sources like Google News in Hindi.
    - Clean and preprocess text data (tokenization, stopword removal, stemming/lemmatization).
    - Use techniques like TF-IDF, Bag of Words, or sentence embeddings for feature extraction.
2. **Model Development**:

    - Begin with clustering algorithms like K-Means, DBSCAN, or Agglomerative Clustering.
    - Utilize advanced embeddings (e.g., BERT embeddings or Sentence Transformers) for semantic similarity.

   - Measure similarity using cosine similarity or Euclidean distance.

3. **Iterative Improvement**:

   - Optimize clustering models by experimenting with cluster sizes and distance metrics.
   - Perform error analysis to identify misgrouped headlines and refine model parameters.

Dataset Utilization

- Prepare a dataset of 10,000 headlines grouped by topics for training and evaluation.
- Split into subsets for model tuning and validation.

Visualization and Analysis

- Apply dimensionality reduction techniques like t-SNE or PCA for visualizing headline clusters.
- Create dashboards highlighting clustering accuracy, similarity scores, and topic distribution.

Reporting and Evaluation

- Track clustering metrics:
  - Cosine similarity
  - F1-score
  - Clustering accuracy
- Use qualitative analysis of clustered headlines to validate topic coherence.