Explainable Insight Generation in NLP

Leveraging Multi-Format Data for Enhanced Understanding

Track B: Explainable NLP-Driven Insight Generation Using Structured, Text, and PDF Data

1. Introduction

The objective of Track B is to develop a robust system for advanced data analysis and insight generation that emphasizes analytical depth, reasoning, explainability, and the generation of actionable insights. This solution strictly adheres to the Problem Statement (PS) constraints by utilizing only classical AI and NLP techniques, ensuring full compliance without reliance on pretrained large language models or neural networks.

2. Dataset Description

The data utilized in this solution spans three categories:

2.1 Structured Data

- **Datasets:** train.csv and test.csv
- **Purpose:** These datasets are employed for learning and validating analytical patterns, providing a structured basis for the initial phases of data processing.

2.2 Unstructured Text Data

- **Datasets:** Large raw text corpora, such as literary documents
- **Purpose:** This data is used for corpus-level language and semantic analysis, enabling the extraction of insights through classical NLP techniques.

2.3 PDF Documents

- **Datasets:** Various PDF files

- **Purpose:** PDFs are processed using text extraction methods to contribute to knowledge extraction, pattern discovery, and contextual insight retrieval. No document embeddings or pretrained language models are used.

3. PDF Processing Methodology

- **Text Extraction:** Classical parsing techniques are utilized for extracting text from PDFs.
- **Text Cleaning and Normalization:** The extracted text undergoes cleaning and normalization to ensure consistency and quality.
- **Chunk-Based Processing:** Large PDFs are processed in chunks to manage size and complexity.
- **Integration:** Extracted text from PDFs is integrated into the unified text corpus, treated as equivalent to other unstructured text data. No OCR-based deep learning models are employed.

4. System Architecture Overview

The architecture is a transparent, modular pipeline that demonstrates the flow of data:

- **Raw Data (CSV + Text + PDF)**
  - **→ Text Preprocessing**
  - **→ Feature Extraction**
  - **→ Pattern Discovery**
  - **→ Explainable Analysis**
  - **→ Insight Generation**

5. NLP Techniques Used

The NLP techniques implemented include:

- **Tokenization:** Breaking down text into individual units or tokens.
- **Lemmatization:** Reducing words to their base or root form.
- **Stopword Removal:** Filtering out common, non-informative words.
- **N-gram Analysis:** Analyzing contiguous sequences of n items in the text.
- **Statistical Language Modeling:** Understanding language patterns and structures.

Explicit Statement: No pretrained embeddings and no transformer-based NLP techniques are used.

6. Feature Representation

The following feature representation techniques are employed:

- **Bag of Words:** A simple, interpretable method for representing text data.
- **TF-IDF Vectorization:** Captures the importance of terms within documents, enhancing interpretability and transparency.

7. Models and Algorithms

Unsupervised Learning

- **K-Means Clustering:** For identifying patterns and groups within the data.
- **Hierarchical Clustering:** Provides a tree-like representation of data clusters.
- **DBSCAN:** Used for anomaly detection, identifying outliers in the data.

Topic Modeling

- **Latent Dirichlet Allocation (LDA):** For discovering topics within a corpus.
- **Non-negative Matrix Factorization (NMF):** An alternative method for topic modeling.

Dimensionality Reduction

- **PCA:** Principal Component Analysis for reducing dimensionality.
- **t-SNE:** Used solely for visualization purposes.

8. PS-Compliant RAG Architecture

The system employs a Classical Retrieval–Analysis–Generation (RAG) architecture:

Retrieval

- **TF-IDF Vectors:** Used for document representation.
- **Cosine Similarity:** To measure document similarity.
- **Top-K Document Selection:** Across text and PDF corpus.

Analysis

- **Topic Inference:** Identifying topics within documents.
- **Cluster Membership:** Determining document clusters.
- **Keyword Contribution Analysis:** Evaluating term significance.

Generation

- **Rule-Based and Statistical Insight Synthesis:** Crafting insights from data.
- **Template-Driven Explanation Generation:** Providing structured explanations.

Explicit Statement: This system does not use LLM-based RAG or neural retrievers/generators.

9. Optimization Techniques

The solution includes several optimization techniques:

- **Vocabulary Pruning:** Reducing vocabulary size for efficiency.
- **Sparse Matrix Computation:** Enhancing computational efficiency.
- **Chunk-Based Processing for PDFs:** Managing large files effectively.
- **Silhouette Score for Clustering:** Evaluating clustering quality.

- **Topic Coherence Optimization:** Ensuring meaningful topic discovery.

10. Explainability & Insight Generation

The system is designed to:

- **Identify Key Contributing Terms:** Highlight significant words or phrases.
- **Explain Cluster and Topic Formation:** Provide transparent reasoning for groupings.
- **Generate Actionable, Human-Readable Insights:** Deliver insights that are understandable and useful.
- **Maintain Full Transparency:** Ensure clarity and interpretability at every stage.

11. PS Compliance Statement

This section confirms that:

- No pretrained LLMs are used.
- No transformer models are employed.
- No external APIs or cloud services are incorporated.
- All models are classical and explainable.
- The solution is fully reproducible and PS-compliant.

12. Conclusion

This document outlines a comprehensive approach to data analysis and insight generation that leverages classical AI and NLP techniques. By effectively handling multi-format data (CSV, text, PDF) and emphasizing transparency and interpretability, the solution aligns with Track B's evaluation criteria and offers a robust framework for generating explainable insights.