Research Phase Documentation

Project Title: Enhancing Text Analytics Data Quality with NLP

Team ID: d24f3a13

Research Duration: May 1, 2025 to May 10, 2025

Prepared By: Team d24f3a13

Objective of Research Phase

The objective of this phase is to understand the current landscape of Natural Language Processing (NLP) techniques used to improve the quality of textual data. This involves exploring the concepts, tools, and best practices related to:

- Text normalization
- Named Entity Recognition (NER)
- Sentiment analysis
- Data cleaning and validation in NLP pipelines
- Applications in domains such as customer reviews and social media analysis

Key Articles & Documentation Reviewed

Text Normalization

A Survey on Text Normalization Techniques

Summary: Discusses stemming, lemmatization, case normalization, and removal of stopwords/punctuation.

Takeaway: These techniques are vital to reduce data variability and noise before performing any advanced analytics.

Named Entity Recognition

Research Phase Documentation

Neural Architectures for Named Entity Recognition

Summary: Covers rule-based and deep learning-based NER models (e.g., BiLSTM-CRF, BERT-based NER).

Takeaway: Using pre-trained models like spaCy or transformers (HuggingFace) improves recognition accuracy.

Sentiment Analysis

Sentiment Analysis Using Deep Learning Techniques

Summary: An overview of lexicon-based and machine learning-based sentiment classification techniques.

Takeaway: Transformer-based models like BERT yield high accuracy and are ideal for contextual sentiment understanding.

Data Cleaning Techniques

Practical Approaches to Cleaning NLP Datasets

Summary: Outlines techniques such as deduplication, outlier filtering, language detection, and handling missing values.

Takeaway: Data quality significantly impacts the accuracy of downstream NLP models.

Tools and Libraries Identified

- spaCy: For text normalization and NER

- NLTK: For preprocessing (stopwords, tokenization)

- TextBlob/VADER: For rule-based sentiment analysis

- Transformers (Hugging Face): For deep learning-based NER and sentiment models

- Pandas/Numpy: For data manipulation and validation

- Streamlit: For UI and deployment of the final application

Research Phase Documentation

Outcome of Research Phase

- Defined a clear pipeline structure:

  Data Collection -> Text Preprocessing -> NER -> Sentiment Analysis -> Quality Checks -> Visualization

- Selected suitable tools and models for each task.

- Acquired datasets from sources such as customer review platforms and Twitter for testing and model validation.