Trainees Task: Comparative Analysis of Text Classification Embeddings

Task Title

Implementation and Comparison of TF-IDF, Word2Vec, and Domain-Specific Transformers for Text Classification in Medical and Financial Domains.

Task Objective

The primary objective is to execute a rigorous comparison of three distinct text representation approaches—**TF-IDF**, **Word2Vec**, and **Pre-trained Transformer Embeddings**—on classification problems within the specialized **Medical Diagnosis** and **Financial Sentiment** domains. This work will demonstrate the trade-offs between speed, semantic capture, and domain adaptation across the different methodologies.

Task Scope and Required Deliverables

The task involves three mandatory phases of work, culminating in a technical report.

Phase 1: Data Acquisition and Preparation

1. **Select Datasets (Mandatory):**
    - **Medical Domain:** Select a publicly available dataset for clinical note or medical abstract classification (e.g., Medical Abstracts Text Classification Dataset).
    - **Financial Domain:** Select a publicly available dataset for financial sentiment analysis (e.g., Financial PhraseBank).
2. **Data Preprocessing:**
    - For *both* datasets, perform essential text cleaning: handling missing values, standardization (lower-casing), and tokenization appropriate for each subsequent pipeline (A, B, C).

- Split each domain dataset (Medical and Financial) into Training, Validation, and Test sets (a 70/15/15 split is recommended).

Phase 2: Implementation of Six Classification Pipelines

For **each domain** (Medical and Financial), you **must** implement and train the following three independent classification pipelines. Use a consistent, simple classification algorithm like **Logistic Regression** or **Support Vector Machine (SVM)** for Pipelines A and B.

Pipeline A: TF-IDF + Classifier (Baseline)

1. **Feature Extraction:** Calculate **Term Frequency-Inverse Document Frequency (TF-IDF)** vectors using the training data.
2. **Model Training:** Train the simple classifier (e.g., SVM) using the generated TF-IDF vectors.
3. **Evaluation:** Perform prediction and metric calculation on the Test set.

Pipeline B: Word2Vec + Classifier (Semantic Baseline)

1. **Embedding Training:** Train a Skip-gram or CBOW **Word2Vec** model specifically on the training corpus of the respective domain dataset.
2. **Feature Vector Creation:** Convert each document into a fixed-size vector representation, typically by calculating the **Averaged Word2Vec** vector across all tokens in the document.
3. **Model Training:** Train the same classifier (e.g., SVM) using these averaged document vectors.
4. **Evaluation:** Perform prediction and metric calculation on the Test set.

Pipeline C: Pre-trained Transformer Embedder + Classifier (State-of-the-Art)

1. **Transformer Selection (Domain-Specific Required):** Select a specialized pre-trained model for each domain:
    - *Medical:* Utilize a domain-specific BERT variant (e.g., **BioBERT** or **ClinicalBERT**).

    - *Financial:* Utilize a financial-specific BERT variant (e.g., **FinBERT**).
2. **Feature Extraction:** Extract static, non-fine-tuned embeddings (e.g., the pooled `[CLS]` token embedding) from the last hidden layer of the chosen Transformer model.
3. **Model Training:** Train the same classifier (e.g., SVM or a small Dense Neural Network) using the extracted Transformer embeddings.
4. **Evaluation:** Perform prediction and metric calculation on the Test set.

Phase 3: Analysis and Reporting

1. **Metric Calculation:** For all six experiments (3 pipelines 2 domains), calculate and record the following metrics on the **Test Sets**:
    - Accuracy
    - Precision, Recall, and F1-Score (per class and macro/weighted average)
2. **Resource Analysis:** Track and estimate the computational resources required for each technique (e.g., training time in minutes, memory usage for feature storage).

Required Deliverables (Technical Report)

You must submit a comprehensive technical report (in Markdown or PDF format) that includes the following sections:

1. **Consolidated Results Table:** A single table clearly summarizing the F1-Scores (Weighted Average) and Training Time for all six models.
2. **Domain-Specific Discussion:**
    - **Medical Domain:** Analyze and explain the performance differences, focusing on how well each technique handled medical terminology, acronyms, and contextual information.
    - **Financial Domain:** Analyze and explain the performance differences, specifically addressing the models' ability to capture nuanced financial sentiment and subtle market language.

3. **General Conclusion and Recommendations:** A detailed written analysis addressing:
    - The overall trade-off between computational cost and classification performance across all three representation types (TF-IDF vs. Word2Vec vs. Transformers).
    - A recommendation on which representation method is most appropriate for a new project, given constraints on data size and available computational power.