Simple LLM-Based Text Classification

Sentiment Analysis on Public Dataset

Gianni Franchi

December 16, 2024

Overview

- **Goal:** Classify text from a sentiment-analysis dataset.
- **Dataset:** Includes tweet text, sentiment labels (neutral, positive, negative), and metadata.
- **Approach:**
  - Embed the dataset using representations from a pre-trained LLM.
  - Train classical machine learning models on the LLM embeddings.
- **Outcome:** Evaluate the efficacy of LLM embeddings for text classification.

Dataset Description

- Sentiment analysis dataset with multiple fields:
  - `text`: The full text of the tweet.
  - `selected_text`: The key phrase expressing sentiment.
  - `sentiment`: Labels (neutral, positive, negative).
  - Metadata: Time of tweet, user's age group, country, etc.
- Example Records:

textID	**text**	**sentiment**	**Age**
cb774db0d1	I‘d have responded, if I were going	neutral	
549e992a42	Sooo SAD I will miss you here in San Diego!!!	negative	
088c60f138	my boss is bullying me...	negative	

Methodology

1. **Data Preprocessing:**
    - Clean text data.
    - Extract relevant fields for analysis.
2. **Embedding with LLM:**
    - Use a pre-trained LLM (e.g., BERT, GPT) to extract embeddings.
    - Represent text as fixed-dimensional vectors.
3. **Classification:**
    - Train classical ML models (e.g., SVM, Random Forest) on embeddings.
    - Compare performance against traditional feature-based methods.
4. **Evaluation:**
    - Use metrics like accuracy, F1-score, and confusion matrix.

Expected Results

- Improved sentiment classification accuracy using LLM embeddings.
- Insights into the interplay between LLM representations and classical ML.
- Demonstrate the flexibility of combining modern embeddings with simple models.

Related Work

- Study the role of LLMs in transfer learning:
  - ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”
  - ”Attention is All You Need” (Transformer architecture)
- Explore classical ML techniques:
  - Support Vector Machines, Random Forest, Logistic Regression.

Thank You!