Sentence Classification in PubMed RCTs: A Comparative Study of Traditional and Deep Learning Models

Background

his project focuses on building **NLP classification models** for sentence-level **categorization of biomedical abstracts**, aiming to improve the **efficiency**[1] and **accuracy** during searching.

·**Problem:** Scientists face information overload when searching and screening biomedical literature.

·**Goal:** Automatically classifying abstract sentences can improve retrieval efficiency and aid clinical decision-making.(five types of summary labels (Background, Objective, Methods, Results, Conclusions))

·**Prior Approaches:** Traditional ML (e.g., SVM, CRF) and deep learning (e.g., CNN, RNN) have been used for this task.

·**Gap:** However, systematic comparisons are rare, and prior work often neglects semantic structures like sentence sequence and context.

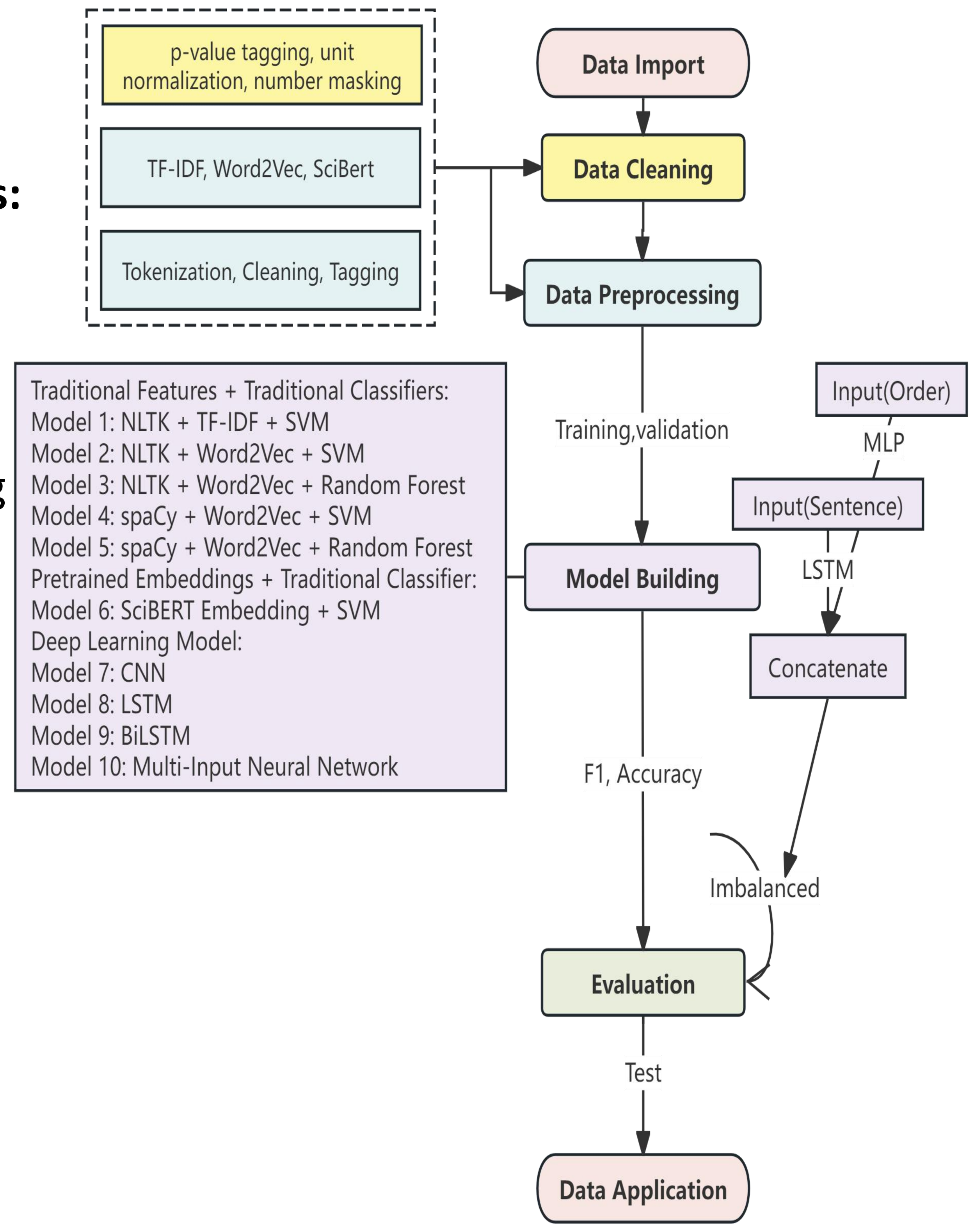
This study mainly focuses on comparing four dimensions: **tokenization methods**, **traditional classifiers**, **deep learning architectures**, and **embedding techniques**.

Data comes from the **PubMed 20k RCT** dataset.I used a stratified subset: **119,893** sentences for **training**, **11,880** for **development**, and **12,088** for **testing**.

Ethics Statement: Publicly available, de-identified data from PubMed 20k RCT was used; no ethical approval required.

Method

Traditional classifiers:
·**SVM:** strong linear baseline
·**Random Forest:** interpretable, robust to overfitting
Deep learning architectures:
·**CNN:** local semantic patterns (n-gram)
·**LSTM:** sequential sentence modeling
·**Multi-input:** joint modeling of sentence content and sentence position via two branches



Embedding techniques:
·**Word2Vec** (static) vs **SciBERT** (contextual, biomedical)

Feature Representation:
·**TF-IDF:** converts text into sparse numerical vectors based on term importance(Baseline)

Evaluation: Considering the category imbalance in this data, the evaluation mainly uses two indicators: **F1-score and Accuracy**

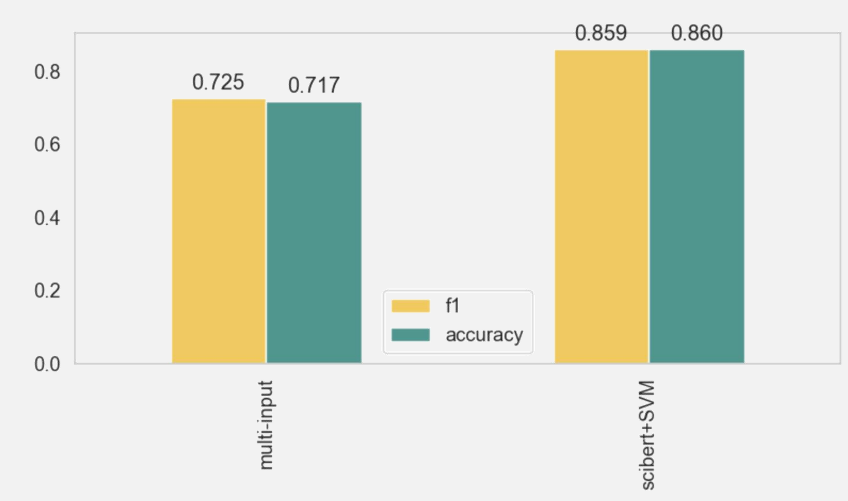
Result

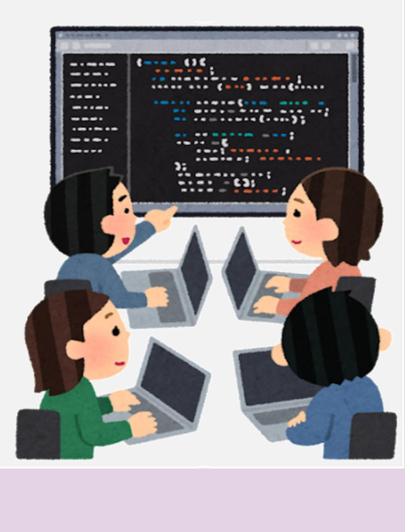
In terms of both **feature representation** and **classification performance**, **SciBERT(0.8663) outperformed Word2Vec-based models**, while **spaCy + Word2Vec + Random Forest(0.7049)** performed **poorly.**

From the deep learning, the **multi-input neural network(0.8168)** was the best performance.

model	recall	precision	f1	accuracy
TF-IDF/SVM	0.8060	0.8074	0.8063	0.8060
NLTK/Word2Vec/SVM	0.7218	0.7236	0.7222	0.7218
NLTK/Word2Vec/RF	0.7258	0.7200	0.7178	0.7258
Spacy/Word2Vec/SVM	0.7026	0.7059	0.7033	**0.7026**
Spacy/Word2Vec/RF	0.7049	0.6990	**0.6952**	**0.7049**
Scibert/SVM	**0.8663**	**0.8662**	**0.8656**	**0.8663**
CNN	0.7778	0.7850	0.7800	0.7778
LSTM	0.7494	0.7554	0.7520	0.7494
Bilstm	0.7506	0.7588	0.7537	0.7506
Multi-Input Neural Network	**0.8168**	**0.8205**	**0.8169**	**0.8168**

Based on the **test results**, the **SciBERT + SVM model(0.860)** demonstrated stable.





Discussion

Findings: Multi-input deep learning model has demonstrated **better task transfer capabilities** and show **potential** in sentence classification tasks.

Comparison: This project **systematically compared** various model architectures, but the model fitting process was **relatively simple** compared to other studies[1,3].

Strengths: This was conducted under a **unified preprocessing framework**. It proposed **hybrid architectures**, and employed domain-specific embeddings.

Limitations: The Word2Vec model was **not pretrained** on general datasets but trained only on PubMed, limiting its effectiveness. Deep learning models **risk overfitting**. Only SciBERT was used for contextual embeddings, **without comparison**. **Hyperparameter tuning** was **limited**.

Implications: Automated classification can support large-scale biomedical information extraction.

Feture Research: Future work could **incorporate CRF layers** for better sequential modeling, and apply the **proposed framework to real-world** (like RAG etc.) and clinical QA systems.

Conclusion

NLTK-based tokenization outperformed spaCy in downstream classification tasks.

As **baseline model** TF-IDF + SVM performed **better**; SciBERT outperformed Word2Vec-based methods.

Combining semantic embeddings with **structural features** significantly enhances classification and supports **more efficient** literature screening and clinical decisions.

Please scan for reference

The image (or QR code) needs to be enlarged(400%) considerably to be properly scanned.