Project Report on Sentence Classification of PubMed 200k RCT Dataset

Motivation & Introduction

This project aims to develop an **automated sentence classification model** that can categorize sentences from research papers into their respective sections. By leveraging **machine learning techniques**, we can improve the efficiency of processing biomedical literature and enhance search engines for medical studies. We compare traditional **Bag of Words (BoW) models** like TF-IDF and CountVectorizer with **advanced word embeddings (BioWordVec)** to determine the most effective approach for this classification task.

Data Description

The dataset used in this project is **PubMed 200k RCT**, which consists of research paper abstracts, where each sentence is labeled based on its function within the study. The dataset is split into:

- **Training Set**: 180040 sentences (used to train the model)
- **Validation Set (Dev)**: 30212 sentences (used for hyperparameter tuning)
- **Test Set**: 30135 sentences (used for final model evaluation)

Each sentence belongs to one of the following categories:

- **OBJECTIVE**: Describes the study's purpose.
- **METHODS**: Details the experimental design and methodology.
- **RESULTS**: Presents findings and statistical outcomes.
- **CONCLUSIONS**: Summarizes key takeaways and implications.
- **BACKGROUND:** Provides Context and Rationale

By correctly classifying these sentences, the model can help in **automating research summarization** and improving biomedical information retrieval.

Method & Evaluation Setup

3.1 Preprocessing

To prepare the dataset for training, we:

1. Tokenize and clean sentences with the sentence_to_vectoru function
2. Convert words into numerical representations using:
    - **TF-IDF** (Baseline method)
    - **CountVectorizer** (Baseline method)

    - **BioWordVec embeddings** (biomedical domain-specific embeddings)

3.2 Models Used

We evaluate multiple machine learning models for sentence classification:

- **Baseline Model: Naïve Bayes (MultinomialNB)** using TF-IDF and CountVectorizer.
- **Improved Model: Logistic Regression** with BioWordVec embeddings.
- **Alternative Model: Support Vector Machine (SVM)** (initially tested but replaced due to performance issues).

3.3 Hyperparameter Tuning

We use **GridSearchCV** to optimize hyperparameters:

- **Naïve Bayes**: Tuned `alpha` parameter for smoothing.
- **Logistic Regression**: Tuned `C` parameter using `GridSearchCV(C = [0.1, 0.5, 1])`.

3.4 Evaluation Metrics

- **Accuracy** (main performance metric)
- **Precision, Recall, and F1-score** (for detailed classification analysis)
- **Confusion Matrix** (to analyze misclassifications)

Results & Discussion

- **BioWordVec significantly improves classification** compared to traditional -based models.
- **Naïve Bayes performs poorly on complex sentences** because it assumes word independence.
- **Logistic Regression with BioWordVec captures biomedical term relationships** better than BoW.

Conclusion

This study demonstrates that **domain-specific word embeddings (BioWordVec) significantly improve sentence classification accuracy in biomedical literature** compared to traditional BoW methods. The **Logistic Regression model with BioWordVec outperformed Naïve Bayes**, proving that semantic information captured by embeddings enhances classification.