Chemical Named Entity Recognition using Fine-tuned BERT and RoBERTa for Biomedical Text Mining

Introduction:

In recent years, the exponential growth of biomedical literature has made it increasingly difficult for researchers to stay up to date with the latest findings and advancements in their respective fields. This growth has led to the development of advanced text mining techniques that enable the extraction and analysis of relevant information from large volumes of text. One such essential task in the field of biomedical text mining is Named Entity Recognition (NER), which involves the identification and classification of entities such as chemical compounds, genes, and proteins mentioned in the text.

Named Entity Recognition plays a critical role in various biomedical applications such as drug discovery, chemical-gene-disease relationship analysis, and clinical decision support systems. Among the numerous types of entities recognized in biomedical literature, chemical entities hold particular significance due to their involvement in the development of drugs and the understanding of various biological processes. The identification of chemical entities can contribute to the effective extraction of valuable information from biomedical literature and facilitate advancements in life sciences research.

Problem Description:

The main challenge addressed in this project is the accurate recognition of chemical entities from a given text, specifically focusing on the CHEMDNER dataset. The CHEMDNER dataset is a comprehensive collection of annotated biomedical articles that include mentions of chemical entities. Despite the importance of chemical entity recognition in the biomedical domain, relatively few works have been devoted to exploring this dataset and developing accurate NER models for chemical entities.

The primary goal of this project is to develop a robust and accurate Named Entity Recognition model to recognize and classify chemical entities from biomedical text. To achieve this, we fine-tune a pre-trained BERT model and RoBERTa model on the CHEMDNER dataset and evaluate its performance using metrics such as precision, recall, and F1 score. By enhancing the performance of chemical NER, we aim to contribute to the broader field of biomedical text mining, ultimately supporting researchers and medical professionals in extracting valuable information from the growing body of biomedical literature.

Description of the Data:

The data utilized in this project comes from the CHEMDNER (Chemical Entities Mentioned in Text) dataset, a comprehensive collection of annotated biomedical articles that include mentions of chemical entities. The CHEMDNER dataset was created as part of the BioCreative IV CHEMDNER task, a community-wide effort aimed at fostering the development of text mining and information extraction systems for the biomedical domain.

The chemdner dataset can be accessed through the website of biocreative which is given as (https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/)

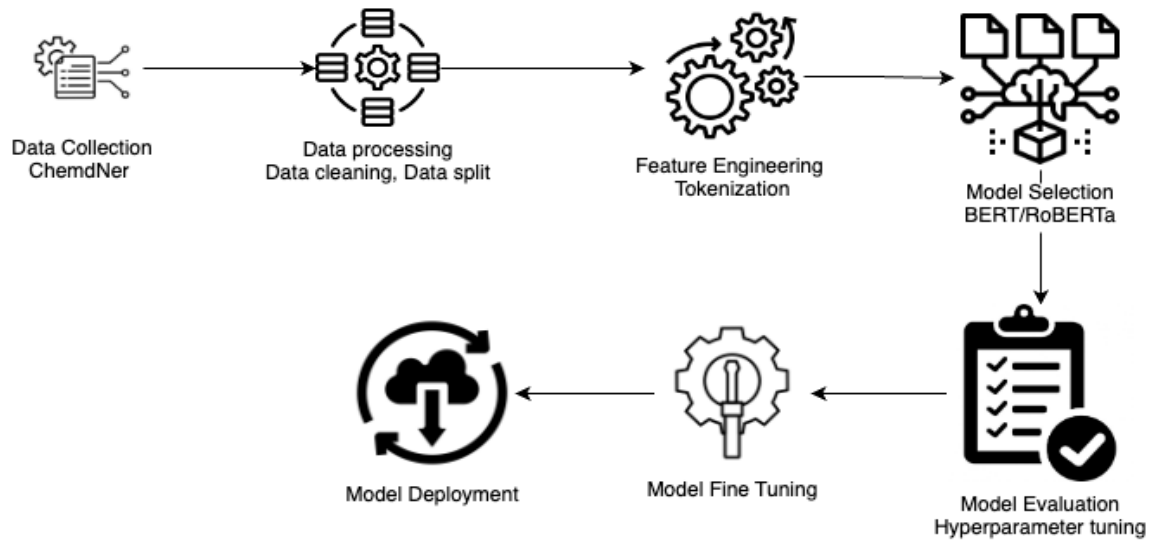
The CHEMDNER dataset comprises a total of 10,000 abstracts from the PubMed database, which is a widely used repository of biomedical literature. These abstracts were carefully selected to cover a broad range of topics related to chemistry and its applications in the life sciences. The abstracts are divided into a training set of 3,500 abstracts, a development set of 3,500 abstracts, and a test set of 3,000 abstracts. Each abstract in the dataset is annotated with chemical entity mentions, which are categorized into the following seven classes:

1. Abbreviation: Shortened forms of chemical names, e.g., "DMSO" for Dimethyl sulfoxide.
2. Family: Groups of chemically related compounds, e.g., "steroids" or "benzodiazepines."
3. Formula: Chemical notations representing the composition of a compound, e.g., "C6H12O6" for glucose.
4. Identifier: Codes or registry numbers assigned to chemical compounds, e.g., "CID 2244" or "CAS 50-00-0."
5. Multiple: Mentions of more than one chemical entity, e.g., "NaCl and KCl."
6. Systematic: Systematic or IUPAC names of chemical compounds, e.g., "2-Propanone" for acetone.
7. Trivial: Common or trivial names of chemical compounds, e.g., "aspirin" or "acetone."

The annotations in the CHEMDNER dataset are provided in the form of character offsets, specifying the start and end positions of each chemical entity mention within the text. We had the dataset available in IOB/BIO (Inside-Outside-Beginning) format. In addition to the character offsets, the dataset also includes the specific entity class for each mention.

In this project, we preprocessed the CHEMDNER dataset to transform it into a format suitable for training and evaluating our Named Entity Recognition model. We tokenize the text and align the tokens with their corresponding entity labels, accounting for any subword tokenization that may occur. The preprocessed data is then used to fine-tune the BERT model and assess its performance in recognizing and classifying chemical entities from biomedical text.

Methodology:



This project employs a two-step methodology: First, we fine-tune a pre-trained model for Named Entity Recognition (NER) using the Hugging Face Transformers library, and then we apply the resulting model to recognize and classify chemical entities in each text. In this report, we detail the use of both BERT-base and RoBERTa models, comparing their performance on the CHEMDNER dataset.

1. Data Preprocessing: We begin by preprocessing the CHEMDNER dataset to transform it into a suitable format for training our NER models. The data is tokenized using the respective tokenizer for each model (BERT-base and RoBERTa), and the token-level labels are aligned with the corresponding entities. We account for any subword tokenization that may occur during this process.
2. Fine-tuning Pre-trained Models: We use the Hugging Face Transformers library to load the pre-trained BERT-base and RoBERTa models, along with their respective tokenizers. We then fine-tune each model on the preprocessed CHEMDNER dataset using a training set and a validation set. During the training process, we employ the Trainer class provided by the Hugging Face library, which facilitates the fine-tuning of the models and the evaluation of their performance. We train each model for multiple epochs and monitor their performance on the validation set using metrics such as precision, recall, and F1 score.
3. Named Entity Recognition Pipeline: After fine-tuning the models, we save them along with their respective tokenizers. Using the saved models and tokenizers, we create a Named Entity Recognition (NER) pipeline for each model. These pipelines can be applied to any input text to recognize and classify chemical entities mentioned within the text.
4. Visualization and Evaluation: To demonstrate the effectiveness of the fine-tuned models, we apply them to example texts and visualize the results using the Spacy library. This allows us to assess the performance of the models in recognizing and classifying chemical entities in real-world scenarios. Furthermore, we compare the performance of the BERT-base and RoBERTa models, analyzing the trade-offs between their respective strengths and weaknesses in the context of chemical entity recognition.

By employing this methodology, we successfully develop and evaluate NER models for chemical entity recognition using both BERT-base and RoBERTa. The fine-tuned models can be applied to various use-cases in the biomedical domain, such as information extraction, relation identification, and database curation.

Results:

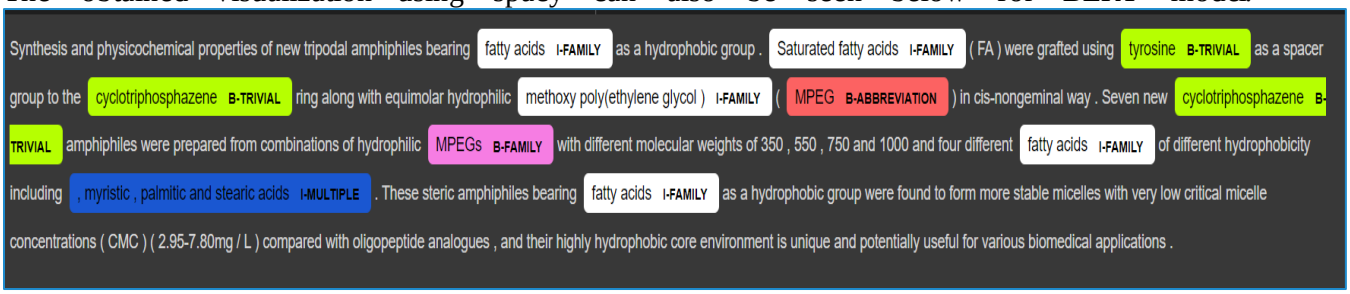
After fine-tuning the BERT-base and RoBERTa models on the CHEMDNER dataset, we obtained the following results:

BERT-base Model: The BERT-base model was trained for 3 epochs, and the performance metrics obtained for each epoch are as follows:

Epoch	Training Loss	Validation Loss	Precision	Recall	F1 Score	Accuracy
1	-	0.1340	0.7780	0.7857	0.7818	0.9596
2	0.2212	0.1109	0.8013	0.8467	0.8234	0.9670
3	0.0968	0.1077	0.8065	0.8648	0.8347	0.9687

The BERT-base model achieved the best performance in the third epoch, with an F1 score of 0.8347 and an accuracy of 0.9687.

The obtained visualization using spacy can also be seen below for BERT model.



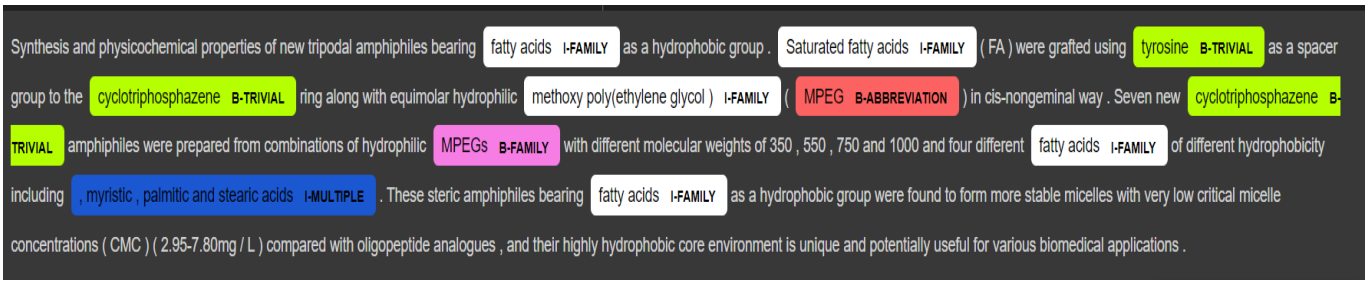
RoBERTa Model: The RoBERTa model was trained for 5 epochs, and the performance metrics obtained for each epoch are as follows:

Epoch	Training Loss	Validation Loss	Precision	Recall	F1 Score	Accuracy
1	0.2094	0.1062	0.7910	0.8437	0.8165	0.9671
2	0.0966	0.0876	0.8481	0.8658	0.8568	0.9744
3	0.0671	0.0838	0.8585	0.8857	0.8719	0.9771
4	0.0531	0.0792	0.8646	0.8918	0.8780	0.9783
5	0.0333	0.0805	0.8715	0.8958	0.8835	0.9791

The RoBERTa model achieved the best performance in the fifth epoch, with an F1 score of 0.8835 and an accuracy of 0.9791.

Comparing the results, the RoBERTa model outperformed the BERT-base model, achieving a higher F1 score and accuracy. This indicates that the RoBERTa model is more effective in recognizing and classifying chemical entities from biomedical text using the CHEMDNER dataset.

The obtained visualization using spacy can also be seen below for RoBERTa model.



When comparing the performance of the BERT-base and RoBERTa models, it is evident that the RoBERTa model outperformed the BERT-base model in all the evaluation metrics. Our most important evaluation metric was Precision since we wanted to focus on reducing the number of false positives and RoBERTa did a better job than BERT base in this regard.

Table 1: Sample of Real and Predicted Labels

Real_Text	Real_Label	Predicted_Text	Predicted_Label
haloperidol	B-TRIVIAL	haloperidol	B-TRIVIAL
aflatoxin	B-FAMILY	aflatoxin	B-FAMILY

aflatoxin	B-FAMILY	aflatoxin	B-FAMILY
aflatoxin	B-FAMILY	aflatoxin	B-FAMILY
aflatoxin	B-FAMILY	aflatoxin	B-FAMILY
aflatoxin	B-FAMILY	aflatoxin	B-FAMILY
aflatoxin	B-FAMILY	aflatoxin	B-FAMILY
copper	B-SYSTEMATIC	copper	B-SYSTEMATIC
cadmium	B-SYSTEMATIC	cadmium	B-SYSTEMATIC
arsenic	B-SYSTEMATIC	arsenic	B-SYSTEMATIC

Table 1 shows a sample of real labels and predicted labels for various chemical entities. It demonstrates that the models perform well in predicting the correct labels for certain classes, such as B-TRIVIAL, B-FAMILY, and B-SYSTEMATIC.

Table 2: Performance Metrics for Each Tag

Tag	Recall	Precision	F1 Score	Support
B-TRIVIAL	0.78	0.95	0.86	25592
B-FAMILY	0.67	0.86	0.75	11932
B-SYSTEMATIC	0.56	0.85	0.68	19136
B-FORMULA	0.01	0.16	0.02	12019
I-SYSTEMATIC	0.60	0.83	0.70	5904
I-FAMILY	0.64	0.76	0.69	4432
B-ABBREVIATION	0.00	0.46	0.00	13115
B-MULTIPLE	0.07	0.91	0.13	584
I-MULTIPLE	0.34	0.88	0.49	1934
I-TRIVIAL	0.35	0.81	0.49	3403
B-IDENTIFIER	0.00	0.03	0.00	1820
I-FORMULA	0.00	1.00	0.00	1229
I-ABBREVIATION	0.00	0.00	0.00	198
B-NO CLASS	0.00	0.00	0.00	113
I-IDENTIFIER	0.00	0.00	0.00	208
I-NO CLASS	0.00	0.00	0.00	18

In conclusion, the results suggest that the RoBERTa model is more effective in recognizing and classifying chemical entities from biomedical text using the CHEMDNER dataset. The fine-tuned RoBERTa model can be utilized in various biomedical applications which is discussed in next section.

Application and Significance of our Research Project:

1. Drug discovery and development: Identifying chemical entities can help researchers find new drug candidates, understand drug interactions, and track the progress of compounds in clinical trials.

2. Chemical patent analysis: By recognizing chemical compounds in patent documents, researchers can study trends in chemical innovations, assess the novelty of new inventions, and identify potential licensing opportunities.
3. Scientific literature mining: Extracting chemical entities from scientific articles enables the automated extraction of knowledge, which can be useful for literature reviews, meta-analyses, and hypothesis generation.
4. Chemical database curation: Automatic recognition of chemical names can help to maintain and update chemical databases, ensuring they are accurate and up-to-date, thus aiding researchers in their work.

By improving the ability to automatically identify and classify chemical entities in text, this project can contribute to advancements in these areas, ultimately benefiting various stakeholders such as researchers, regulators, and educators.

Limitations:

1. Subword Tokenization: The tokenization process employed in this project may result in some chemical entities being split into subword tokens. While we have attempted to align the labels accordingly, this could still introduce inaccuracies in the predicted entity boundaries.
2. Model Generalizability: The models were specifically fine-tuned on the CHEMDNER dataset, which may limit their generalizability to other biomedical corpora or different types of chemical entities.

By acknowledging these limitations and addressing them in future work, the performance and applicability of the developed Named Entity Recognition models can be further improved, contributing to the extraction of valuable information.

References:

- Krallinger, M. et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform, 2014
- Wei, C. H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., ... & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction: Overview of the BioCreative V chemical-disease relation (CDR) task. Database, 2016.
- ChatGPT for debugging, beautifying the visuals, new methods to incorporate and organizing