Medication Extraction and Entity Linking using Stacked and Voted Ensembles on LLMs

Pablo Romero[1], **Lifeng Han**[2,3*], and **Goran Nenadic**[3]
[1] Manchester Metropolitan University, UK
[2] LIACS & LUMC, Leiden University, NL [3] The University of Manchester, UK
* *corresponding author*
pablo2004romero@gmail.com l.han@lumc.nl, g.nenadic@manchester.ac.uk

Abstract

Medication Extraction and Mining of its related **attributes** play an important role in healthcare NLP research due to its practical applications in hospital settings, such as their mapping into standard clinical knowledge bases (SNOMED-CT, BNF, etc.). In this work, we investigate state-of-the-art LLMs in text mining tasks on medications and their related attributes such as dosage, route, strength, and adverse effects. In addition, we explore different ensemble learning methods (STACK-ENSEMBLE and VOTING-ENSEMBLE) to augment the model performances from individual LLMs. Our ensemble learning result demonstrated better performances than individually fine-tuned base models BERT, RoBERTa, RoBERTa-L, BioBERT, BioClinicalBERT, BioMedRoBERTa, ClinicalBERT, and PubMedBERT across general and specific domains, with **statistical significance** testing (p=0.048). Finally, we build up an entity linking function to map extracted medical terminologies into the SNOMED-CT codes and the British National Formulary (BNF) codes, which are further mapped to the Dictionary of Medicines and Devices (dm+d), and ICD (**Clinical Coding**). We host the fine-tuned models and **desktop applications** at https://github.com/pabloRom2004/Insight-Buddy-AI-App

1 Introduction

Information Extraction on Medications and their related attributes plays an important role in natural language processing (**NLP**) applications in the **clinical** domain to support digital healthcare. Clinicians and healthcare professionals have been doing manual clinical **coding** for quite a long time to map clinical events such as diseases, drugs, and treatments into the existing terminology knowledge base, for instance, ICD and SNOMED. The procedure can be time-consuming yet without a guarantee of total correctness due to human-introduced errors. With the process of automated information extraction on **medications**, it will be further possible to automatically map the extracted terms into the current terminology database, i.e. the automated clinical coding. Due to the promising future of this procedure, different NLP models have been deployed in medication mining and clinical coding in recent years. However, they are often studied separately. In this work, 1) we investigate text mining of medications and their related attributes (dosage, route, strength, adverse effect, frequency, duration, form, and reason) together with *automated clinical coding* into one pipeline. In addition, 2) we investigate the **ensemble** learning mechanisms (Stack and Voting) on a broad range of NLP models fine-tuned for named entity recognition (NER) tasks. These models include both general domain trained BERT, RoBERTa, RoBERTa-L, and domain-specific trained BioBERT, BioClinicalBERT, BioMedRoBERTa, ClinicalBERT, and PubMedBERT. In this way, users do not have to worry about which models to choose for clinical NER. Instead, they can just place the newer models into the ensemble-learning framework to test their performances. We offer desktop applications and web **interfaces** for the clinical NER, ensemble, and coding models we are developing upon paper acceptance.

2 Literature Review and Related Work

2.1 Biomed/Clinical Named Entity Recognition

Named Entity Recognition (**NER**) is a critical task for extracting key information from unstructured text, like medical letters. The complexity and context-dependency of medical language pose significant challenges for accurate entity extraction. Traditional approaches to NER, such as rule-based systems, have shown limited success in capturing the nuanced contextual information crucial for clin-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 303–315
May 4, 2025 ©2025 Association for Computational Linguistics

ical NER (Nadeau and Sekine, 2007). The advent of deep learning methods, particularly Long Short-Term Memory (LSTM) networks, marked a significant improvement in NER performance (Graves and Schmidhuber, 2005), e.g. the ability to capture long-range dependencies in text. However, these models still struggled with rare entities and complex contextual relationships in **clinical notes**. The introduction of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) revolutionised various NLP tasks, including NER. BERT's self-attention mechanism and bidirectional training allow it to capture nuanced contextual information over long pieces of text. The model's pre-training on a large corpus using a masked language modelling objective builds rich token representations. The model can then be later fine-tuned by adding a classification layer at the end of the network to make decisions over each individual token embedding.

However, BERT's pre-training on general domain corpora (Wikipedia and books) limited its effectiveness on specialised medical texts. This limitation led to the development of **domain-specific** BERT variants. For example, BioBERT (Lee et al., 2019), pre-trained on large-scale biomedical corpora; ClinicalBERT (Wang et al., 2023), fine-tuned on EHR data from 3 million patients after pre-training on 1.2 billion words of diverse diseases, and other variants like Med-BERT (Rasmy et al., 2021) have demonstrated enhanced performance on medical NER tasks due to their specialised training on the medical domain [1]. Despite these improvements, **single-model approaches still struggle** with the inherent complexity and variability of clinical text, as the comparative studies reported in (Belkadi et al., 2023) across different models using BERT, ClinicalBERT, BioBERT, and scratch-learned Transformers.

2.2 Ensemble Learning for Biomedical NER

Ensemble methods have emerged as a promising direction to address these challenges, they have proven useful in other fields, such as computer vision (Lee et al., 2018). By combining multiple models, ensembles can leverage the strengths of different models while mitigating their individual weaknesses. In the context of NER, ensemble learning has shown performance improvements, as shown by (Naderi et al., 2021), where an ensemble is used in a *health* and *life science* corpus for a significant improvement in performance over single models. (Naderi et al., 2021) conducted max voting for word-level biology, chemistry, and medicine data. However, on clinical/medical NER, they only focused on French using the DEFT benchmark dataset; while for the other two domains of biology and chemistry, they tested on English data. There are two commonly used ensemble methods, voting and stacked ensembles: 1) **Maximum voting** in ensembles where each model contributes equally to the final decision as used in the paper (Naderi et al., 2021) have proved effective. This is where the most voted label is picked. 2) Training a network on the outputs of the ensemble aims to capture more nuanced relationships. This is accomplished using a method called **stacking** introduced by (Wolpert, 1992). Stacking offers a more sophisticated approach by training a meta-model on the outputs of the base ensemble; the model is expected to learn more complex patterns from the ensemble outputs, leading to better predictions. This has proven effective in this paper (Saleh et al., 2022) where they use a stacked ensemble with a support vector machine (SVM) for *sentiment* analysis. Instead, we will use a simple feed-forward network from the outputs of the ensemble to the final labels for our tasks. more examples on stacked ensemble can be found at (Mohammed and Kora, 2022; Güneş et al., 2017).

Earlier work on ensemble learning for biomedical NER mostly includes older models such as BiLSTM, CRF, SEARN, and RNNs (Ju et al., 2020; Kim and Meystre, 2020; Christopoulou et al., 2020). This work **aims to address this gap** by investigating 1) *whether stacked and voting ensembles can make a difference on NER tasks of clinical notes*, 2) the ensemble performance on newer Deep Learning models based on BERT from domain fine-tuning, which are a) general domain BERT, RoBERTa, and RoBERTa-L, and b) biomedical domain BioBERT, BioClinicalBERT, BioMedRoBERTa, ClinicalBERT, and PubMedBERT.

2.3 Model Quantisation

To make the LLMs more computational friendly and available for smaller machine users, model quantisation is a recent topic in deep learning to reduce the required memory when running the model mostly by reduce the model size, but with-

[1] there have been other versions of Clinical BERTs such as (Huang et al., 2019) and (Alsentzer et al., 2019) that were trained on Medical Information Mart for Intensive Care III (mimiciii) data (Johnson et al., 2016) respectively.

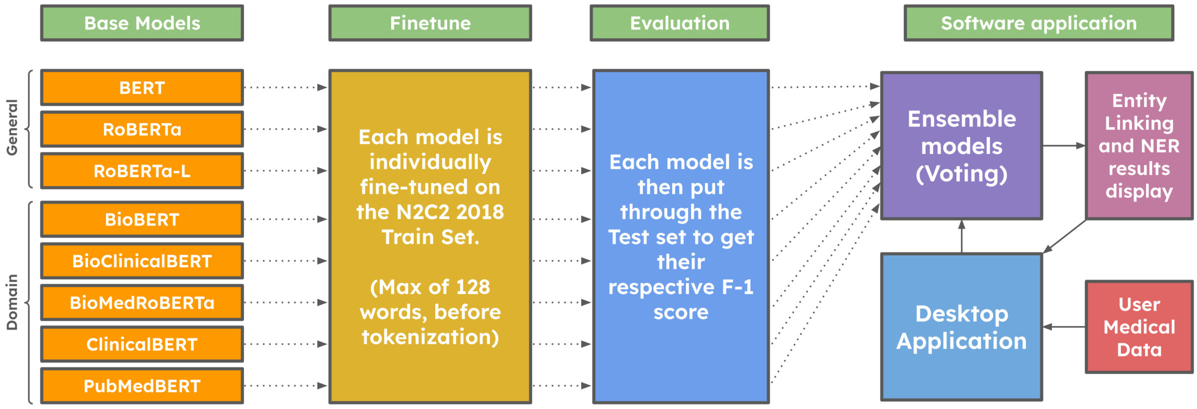

Base Models
Finetune
Evaluation
Software application
General
Domain
BERT
RoBERTa
RoBERTa-L
BioBERT
BioClinicalBERT
BioMedRoBERTa
ClinicalBERT
PubMedBERT
Each model is individually fine-tuned on the N2C2 2018 Train Set.
(Max of 128 words, before tokenization)
Each model is then put through the Test set to get their respective F-1 score
Ensemble models (Voting)
Entity Linking and NER results display
Desktop Application
User Medical Data

Figure 1: InsightBuddy Framework Pipeline: individual NER model fine-tuning, ensemble, and entity linking. Two kinds of base models include the general domain and the biomedical domain with their Huggingface repositories in Table 3. Pre-preprocessing data: cut the sequence with the first full stop "." after the 100th word, otherwise, cut the sequence up to 128 words. Fine-tuning: using the same parameter sets for all eight models. Ensemble: different strategies will be displayed in Fig 2. Entity Linking: links to clinical KB including BNF and SNOMED.

out much effecting the model performances. There are quantisation-aware training and post-training quantisation (PTQ). We use the extreme reduction to 4-bit (16 values) transformers.js Q4 implementation in our work for PTQ. Recent work on this topic can be found at (Lin et al., 2024; Liu et al., 2023).

3 Methodologies

The Overall framework of InsightBuddy is shown in Figure 1, which displays the base models we included from the general domain 1) BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and RoBERTa-Large, and 2) biomedical/clinical domains BioBERT (Lee et al., 2019), BioClinicalBERT (Alsentzer et al., 2019), BioMedRoBERTa (Gururangan et al., 2020), ClinicalBERT (Wang et al., 2023), and PubMedBERT (Gu et al., 2020). The fine-tuning of eight models uses the same set of parameters (Section 4 for parameter selections) and the n2c2-2018 shared task training data with data pre-processing. The initial evaluation phase using n2c2-2018 testing set gives an overall idea of each model's performance. This is followed by ensemble learning on all the models' outputs. With the output from NER models, we add an **entity linking** function to map the extracted medical entities into the standard clinical terminology knowledge base (KB), using **SNOMED-CT and BNF** as our initial KB, which is further mapped to ICD and dm+d.

For data pre-processing, we chunk the sequence into a maximum of 128 words. If there is a full stop "." between the 100th and 128th word, it will be cut at the full stop. Regarding ensemble-learning strategy, we draw a InsightBuddy Ensemble figure (Figure 2) to explain in detail. Firstly the initial output of eight individual fine-tuned NER models is tokenised, i.e. at the **sub-word** level, due to the model learning strategy, e.g. "Para ##ce ##tam ##ol" instead of "Paracetamol". What we need to do at the first step is to **group** the sub-word tokens into words for both practical application and voting purposes. However, each sub-word is labeled with predefined labels and these labels often do not agree with each other within the same words. We designed **three group solutions**, i.e. first-token voting/selection, max-token voting, and average voting. The *first-token voting* is to assign a word the same label as its first sub-word piece. For example, using this strategy, the word "Paracetamol" will be labeled as "B-Drug" if its first sub-word "Para" is labeled as "B-Drug" regardless of other labels from the subsequent sub-words. The *max-token voting* will assign a word the label that has the highest sub-word logit, this indicates that the model is more confident in that prediction, the higher the logit is. The *average voting* solution calculates the average logits across all sub-words predictions and then samples from this to get the label for the entire word.

Regarding **word-level ensemble** learning, we investigate the classical **voting** strategy with modifications (two solutions). For the first solution ">=4 or O", if there are more than half of the mod-

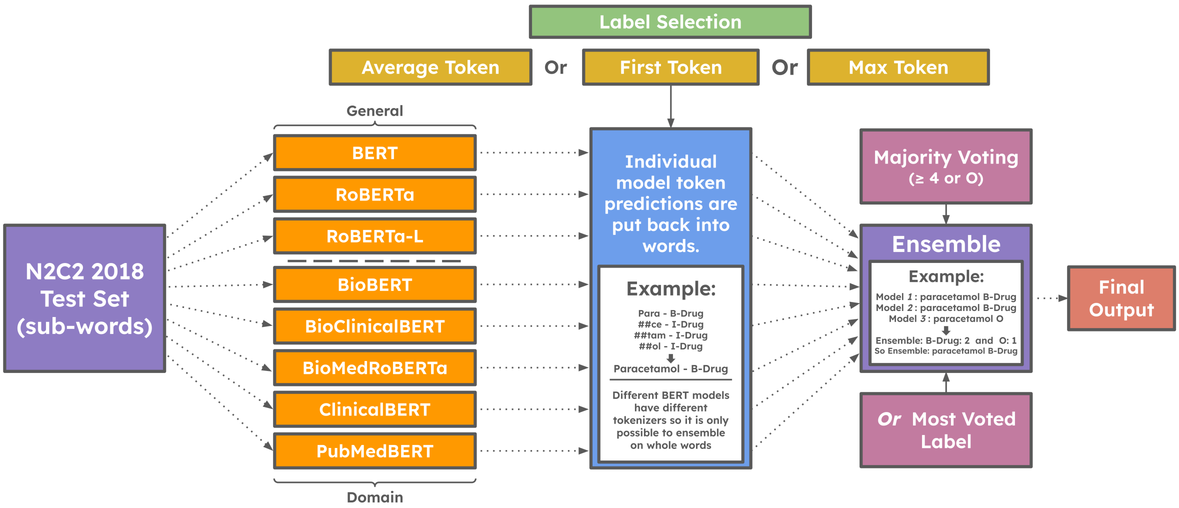

Label Selection
Average Token
Or
First Token
Or
Max Token
General
BERT
RoBERTa
RoBERTa-L
BioBERT
BioClinicalBERT
BioMedRoBERTa
ClinicalBERT
PubMedBERT
Domain
N2C2 2018 Test Set (sub-words)
Individual model token predictions are put back into words.
Example:
Para - B-Drug
##ce - I-Drug
##tam - I-Drug
##ol - I-Drug
Paracetamol - B-Drug
Different BERT models have different tokenizers so it is only possible to ensemble on whole words
Majority Voting (≥ 4 or O)
Ensemble
Example:
Model 1 : paracetamol B-Drug
Model 2 : paracetamol B-Drug
Model 3 : paracetamol O
Ensemble: B-Drug: 2 and O: 1
So Ensemble: paracetamol B-Drug
Or Most Voted Label
Final Output

Figure 2: INSIGHTBUDDY Voted Ensemble Pipeline: individual NER model fine-tuning outputs are at token/sub-word level. "Logits are the outputs of a neural network before the activation function is applied" first, we do the grouping of sub-words into words using three strategies: first token label, max token voting, or average voting (from our results, the first-token-lable selection gives higher Recall, while other two voting give higher precision, but they all end with the same F1 score, ref Table 4). || We take the best output from the first token label selection as the solution. For word-level ensemble on eight models, we have two solutions for voting, 1) either majority voting with >= 4 labels as the same then we pick it, otherwise choose default “O”, or 2) max voting with the most popular label whatever it is; for max voting, if there is a tie, e.g. (3,3,2), we tested both alphabetical pick-up, or random pick-up of tied labels. Our results show that “>=4, or O” performs similarly to “max + alphabetical”, while “max + random” slightly performs lower.

els agree on one label, we pick this label, i.e. >=4 such same labels. Otherwise, we assign the default “O” label to indicate it as context words, due to the models’ disagreement. For the second solution, we use max-voting, i.e. the most agreed label regardless of how many models they are, e.g. 2, 3, 4, or more. In this case, if there are ties, e.g. (3, 3, 2) two labels are voted both three times from six models, we need to decide on the tied labels. There are two solutions for the selection, 1) alphabetical, and 2) fully randomised.

We also draw the **STACKED**-ENSEMBLE in Figure 12 and 13, where the model training and one-hot encoded model predictions are illustrated. In the training phase, we cut the real data into 80% and 20% for the training and testing of the model. Model exports are conducted only if at least 2 models are predicting a label that is not “O”; otherwise “O” is the default option and the output is ignored and not included in the stacked training data. For training data collection, output logits for each model are converted into a one-hot encoded vector, concatenated and saved along with the real label for each token. There are 8 one-hot encoded vectors from 8 individual models and 1 label. So the model during training will see the value “1” eight times from the eight models, and the value “0” for the rest of the vector values. Overall, there are 8 vectors with each length of 19 digits. So there will be 8 (number of models) × 19 (number of labels) - 8 (eight 1s as there are 8 one hot encoded vectors so they have a single 1 each) = 144 “0” values for every training example. We use *one-hot encoding* instead of the output logits themselves to avoid the model *overfitting* because the model makes more confident predictions when running on the training set. As this is the data that it was originally trained on, it is very confident with it’s predictions. We can mitigate this by only feeding the one-hot encoded vectors to the stacked network.

4 Hyper Parameter Optimisations

We used a set of parameters for model fine-tuning and selected the better parameter set as below using the validation data. We tried different learning rates (0.0001, 0.0002, 0.00005) and batch sizes (16, 32).

- learning_rate: 0.00005
- train_batch_size: 32

Individual models max-logit grouping (word)			
Metric	**P**	**R**	**F1**
BERT			
accuracy	0.9773		
macro avg	0.7942	0.7965	0.7928
weighted avg	0.9784	0.9773	0.9775
RoBERTa			
accuracy	0.9780		
macro avg	0.8029	0.8201	0.8094
weighted avg	0.9795	0.9780	0.9784
RoBERTa-Large			
accuracy	0.9788		
macro avg	0.8091	0.8351	0.8202
weighted avg	0.9802	0.9788	0.9792
ClinicalBERT			
accuracy	0.9780		
macro avg	0.8087	0.7916	0.7964
weighted avg	0.9785	0.9780	0.9779
BioBERT			
accuracy	0.9776		
macro avg	0.7972	0.8131	0.8027
weighted avg	0.9787	0.9776	0.9779
BioClinicalBERT			
accuracy	0.9776		
macro avg	0.7999	0.8090	0.8017
weighted avg	0.9788	0.9776	0.9779
BioMedRoBERTa			
accuracy	0.9783		
macro avg	0.8065	0.8224	0.8122
weighted avg	0.9797	0.9783	0.9786
PubMedBERT			
accuracy	0.9784		
macro avg	0.8087	0.8292	0.8166
weighted avg	0.9800	0.9784	0.9788
Voting Max logit ensemble word level			
accuracy	0.9796		
macro avg	**0.8261**	0.8259	**0.8232**
weighted avg	0.9807	0.9796	0.9798

Table 1: Word-level individual model (grouping using max-logit) vs ensemble using max-logit, Eval on n2c2 2018 test data

- eval_batch_size: 32
- seed: 42
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- lr_scheduler_warmup_ratio: 0.1

Model	Macro P	Macro R	Macro F	Accuracy	Tokens
BERT	0.8336	0.8264	0.8283	0.9748	756798
ROBERTa	0.8423	0.8471	0.8434	0.9770	756014
ROBERTa-L	**0.8489**	**0.8606**	**0.8538**	**0.9782**	756014
PubMedBERT	0.8324	0.8381	0.8339	**0.9783**	681211
ClinicalBERT	**0.8482**	0.8245	0.8341	0.9753	*796313*
BioMedRoBERTa	**0.8482**	**0.8477**	**0.8468**	0.9775	756014
BioClinicalBERT	0.8440	0.8405	0.8406	0.9751	791743
BioBERT	0.8365	0.8444	0.8393	0.9750	791743

Table 2: INSIGHTBUDDY individual sub-word level model eval on n2c2-2018 test set. The first group: normal domain PLM; The second group: biomedical PLM. The different numbers of Support are due to the different tokenizers they used – ROBERTa and ROBERTa-L use the same tokenizers, BioClinicalBERT and BioBERT use the same tokenizers, and other models all use different tokenizers; PubMedBERT generated the least number of sub-words/tokens 681,211 while ClinicalBERT generated the largest number of tokens 796,313.

- num_epochs: 4
- mixed_precision_training: Native AMP

5 Experimental Evaluations

We use the n2c2-2018 shared task data on NER of adverse drug events and related medical attributes (Henry et al., 2020). The data is labeled with the following list of labels: ADE, Dosage, Drug, Duration, Form, Frequency, Reason, Route, and Strength in BIO format. So, overall, we have 19 labels, 2 (B/I) x 9 + 1 (O). The original training and testing sets are 303 and 202 letters respectively. We divided the original training set into two parts (9:1 ratio) for our model selection purposes: our new training and validation set, following the data split from recent work by (Belkadi et al., 2023).

We report Precision, Recall, and F1 score in two categories "macro" and "weighted", in addition to Accuracy. The "**macro**" category treats each label class the same weight regardless of their occurrence rates, while the "**weighted**" category" assigns each label class with a weight according to their occurrence in the data. We first report the individual model fine-tuning scores and compare them with related work (subword level); then we report the ensemble model evaluation with different ensemble solutions (word level).

5.1 Individual Models: sub-word level

The performance of individual models after fine-tuning is reported in Table 2 where it says that RoBERTa-L performs the best in the macro Precision (0.8489), Recall (0.8606) and F1 (0.8538) score across general domain models, also winning

Table 3: InsightBuddy integrated individual models and their Huggingface repositories.

Ensemble List	Link
BERT	https://huggingface.co/google-bert/bert-base-uncased
BioBERT	https://huggingface.co/dmis-lab/biobert-base-cased-v1.2
ClinicalBERT	https://huggingface.co/medicalai/ClinicalBERT
BioClinicalBERT	https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
PubMedBERT	https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext
BioMedRoBERTa	https://huggingface.co/allenai/biomed_roberta_base
RoBERTa	https://huggingface.co/FacebookAI/roberta-base
RoBERTa Large	https://huggingface.co/FacebookAI/roberta-large

domain-specific models. BioiMedRoBERTa wins the domain-specific category models producing macro Precision, Recall, and F1 scores (0.8482 0.8477 0.8468). In comparison to the NER work from (Belkadi et al., 2023), who's macro avg scores are: 0.842, 0.834, 0.837 from ClinicalBERT-Apt, our fine-tuned ClinicalBERT has similar performances (0.848, 0.825, 0.834), which shows our fine-tuning was successful. However, our best domain-specific model BioMedRoBERTa produces **higher** scores: macro P/R/F1 (0.8482 0.**8477** 0.**8468**) and weighted P/R/F1 (0.9782 0.9775 0.9776) and Accuracy 0.9775 as in Figure 6. Furthermore, the fine-tuned RoBERTa-L even achieved higher scores of (**0.8489 0.8606 0.8538**) for macro P/R/F1 and Acc 0.9782 in Figure 13. Both fine-tuned BioMedRoBERTa and RoBERTa-Large also *win the best models* reported by (Belkadi et al., 2023) which is their ClinicalBERT-CRF model, macro avg (0.85, 0.829, 0.837), Acc 0.976. Afterwards, in this paper, we emphasis on **word level** instead of sub-word, which was focused on by (Belkadi et al., 2023).

5.2 Ensemble: word-level grouping (logits)

We tried **first** logit voting, **max** voting, and **average** voting to group sub-words into words with corresponding labels. Their results are shown in Table 4, in the upper group. First logit voting produced a higher Recall 0.8260 while Max logit voting produced a higher Precision 0.8261 resulting in higher F1 0.8232, i.e. *Max* logit > *First* logit > *Average* logit with macro F1 (0.8232, 0.8229, 0.8227). However, overall, their performance scores are very close, so we chose the first-logit voting output for the afterward word-level ensemble due to computational convenience.

5.3 Individual vs Ensemble Models

The word-level performance comparisons from individual models and voting max-logit ensembles are presented in Table 1.

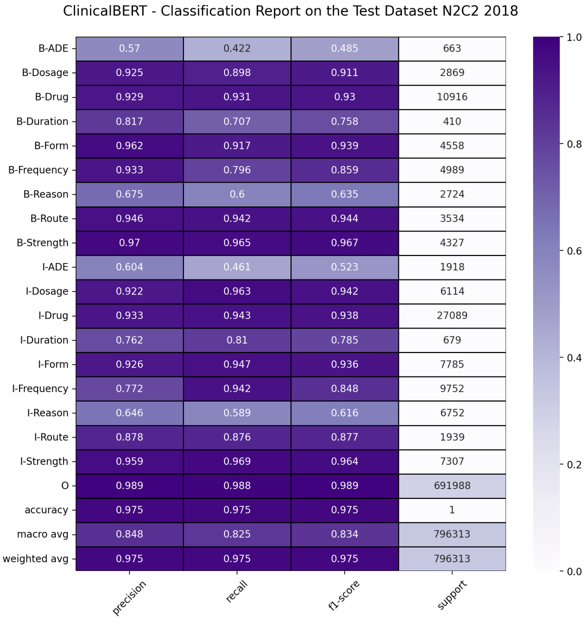


Figure 3: ClinicalBERT Eval at Sub-word Level. This score is similar (slightly winning R/F1) to (Belkadi et al., 2023) paper on ClinicalBERT-Apt whose macro: (85.3 81.0 82.5) and weighted: (0.974, 0.975, 0.974), which says our fine-tuning is successful. However, our best domain-specific model BioMedRoBERTa produces *better* score: macro P/R/F (0.8482 0.8477 0.8468) and weighted P/R/F (0.9782 0.9775 0.9776) and Accuracy 0.9775 as in Figure 8. Furthermore, the fine-tuned RoBERTa-L even achieved higher scores of (0.8489 0.8606 0.8538) for P/R/F1 and Acc 0.9782 in Table 1. Afterwards, in this paper, we emphasis on word level instead of sub-word, which was focused by Belkadi et al. (2023).

5.4 Ensemble: Voting vs Stacked (one-hot)

Regarding Stacked Ensemble using one-hot encoded vectors, as shown in the middle group in Table 4, it actually produced higher Precision score 0.8351 in comparisons to the highest Precision 0.8261 from Voting Ensembles. However, the Recall score on macro avg is 2 point lower than the voting ensemble, 0.8065 vs 0.8260, which means that the Stacked Ensemble *reduced the false positive errors* but also increased the false negative error prediction. This implies that it has stricter constraint on positive predictions.

5.5 Ensemble Models: BIO-span vs non-strict word-level

So far, we have been reporting the evaluation scores on the BIO-strict label categorization, i.e. we distinguish between the label's beginning or the inner part of the label. For instance, a B-Drug will be different from an I-Drug and it will be marked as wrong if they are different from the reference. However, we think in practice, there are situations when users do not need the BIO, especially B and I difference. In Table 4, we can see that, without considering the label difference of B and I, only focusing on the 9 label categories, word level ensemble model produced much higher Macro avg evaluations cores on Precision (0.8844) and Recall (0.8830) leading to higher F1 (0.8821), in comparison to BI-distinguished Macro F1 0.8232 (voting-max-logit) and F1 0.8156 (stacked-first-logit).

5.6 Word-level: voting ensembles vs individual fine-tuned

As in Table 1, BioMedRoBERTa individual word level max logit grouping scores macro avg P/R/F1 (0.8065 0.8224 0.8122 563329) vs max logit ensemble voting P/R/F1 (0.8261 0.8259 0.8232), we can see that ensemble boosted P (0.8261-0.8065)/0.8065= 2.43%, and F1 (0.8232-0.8122)/0.8122= 1.35% which says the ensemble voting is successful. By increasing the Precision score, the *ensembles reduce the **false positive** labels* in the system output, while keeping the Recall at the same level, i.e. the true positive labels.

5.7 Model Quantisation

To reduce the computational cost, we also carried out the quantisation on fine-tuned models. The quantised model can perform similar level of accurate scores in comparison to the original models but with 25% of the size. For instance, using BioMedRoberTa, the quantised model achieved (0.811, 0.821, 0.814) for macro(P, R, F1), which is very similar to the original size fine-tuned model scores (0.8065, 0.8224, 0.8122) as in Table 1, even achieving **slightly higher Precision and F1**. The reasons for this can be that 1) Block-wise Quantization: The Q4 implementation isn't just reducing precision uniformly - it uses sophisticated block-wise quantisation that preserves important patterns while simplifying others. 2) Calibrated Discretization: The extreme reduction to 4-bit (16 values) forces more decisive classification boundaries, which can be beneficial for NER tasks where *clear token boundaries* are important. 3) Optimisation Benefits: The transformers.js Q4 implementation includes specific optimisations for inference beyond simple precision reduction. Overall, this is fundamentally different from *naive quantization* - the transformers.js/GGML approach is carefully designed to maintain model performance while drastically reducing size. In some cases, this sophisticated quantisation can improve results by simplifying decision boundaries in beneficial ways.

The full model size is 497 MB and the 4 Bits Quantised model is 125 MB. The corresponding detailed evaluations on each entity type and the confusion matrix for quantised BioMedRoBERTa are presented in Figure 6 and 7 on word level with BIO.

Voting Average Ensemble word level (BIO)			
Metric	P	R	F1
accuracy	0.9796		
macro avg	0.8253	0.8256	0.8227 ± 0.0037
weighted avg	0.9807	0.9796	0.9798
Voting First logit Ensemble word level (BIO)			
Metric	P	R	F1
accuracy	0.9796		
macro avg	0.8255	**0.8260**	0.8229 ± 0.0034
weighted avg	0.9807	0.9796	0.9798
Voting Max logit Ensemble word level (BIO)			
Metric	P	R	F1
accuracy	0.9796		
macro avg	0.8261	0.8259	**0.8232** ± 0.0036
weighted avg	0.9807	0.9796	0.9798

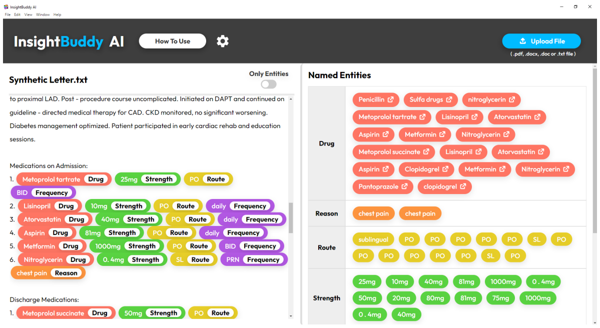
Stacked Ensemble first logit word level (BIO)			
Metric	P	R	F1
accuracy	0.9796		
macro avg	**0.8351**	0.8065	0.8156 ± 0.0037
weighted avg	0.9800	0.9796	0.9794

Non-BIO-only-word ensemble			
Metric	P	R	F1
accuracy	0.9839		
macro avg	0.8844	0.8830	0.8821 ± 0.0025
weighted avg	0.9840	0.9839	0.9838

Table 4: Word-level grouping ensemble voting evaluation with significance test. F1 score: max > first > average logit voting though they are very close scores. The **stacked** ensemble has the highest **Precision** scores, but the lowest Recall scores, which lead to lower F1. In the bottom cluster, it is the word-level evaluation without distinguishing B/I labels, evaluation on n2c2 2018 test data.

5.8 Significance Test

To assess the statistical significance of performance differences between ensemble methods and the strongest individual model (RoBERTa-Large with first token strategy), we conducted bootstrap resampling tests with 500 iterations. Our analysis revealed that the Non-BIO-only-word ensemble showed statistically significant improvement (p = 0.048) over the baseline. Interestingly, while the Stacked Ensemble first logit approach performed significantly worse in F1 score (p = 0.002), it achieved the highest precision (0.8351) among all methods, suggesting potential utility for precision-



Synthetic Letter.txt

to proximal LAD. Post - procedure course uncomplicated. Initiated on DAPT and continued on guideline - directed medical therapy for CAD. CKD monitored, no significant worsening. Diabetes management optimized. Patient participated in early cardiac rehab and education sessions.

Medications on Admission:

1. Metoprolol tartrate Drug 25mg Strength PO Route BID Frequency
2. Lisinopril Drug 10mg Strength PO Route daily Frequency
3. Atorvastatin Drug 40mg Strength PO Route daily Frequency
4. Aspirin Drug 81mg Strength PO Route daily Frequency
5. Metformin Drug 1000mg Strength PO Route BID Frequency
6. Nitroglycerin Drug 0. 4mg Strength SL Route PRN Frequency chest pain Reason

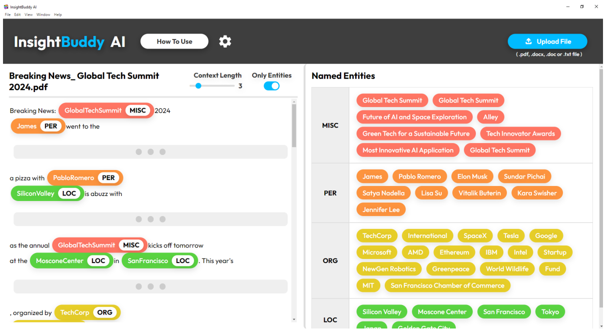
Discharge Medications:

1. Metoprolol succinate Drug 50mg Strength PO Route

Named Entities

Drug	Penicillin Sulfa drugs nitroglycerin Metoprolol tartrate Lisinopril Atorvastatin Aspirin Metformin Nitroglycerin Metoprolol succinate Lisinopril Atorvastatin Aspirin Clopidogrel Metformin Nitroglycerin Pantoprazole clopidogrel
Reason	chest pain chest pain
Route	sublingual PO PO PO PO PO SL PO PO PO PO PO PO SL PO
Strength	25mg 10mg 40mg 81mg 1000mg 0 . 4mg 50mg 20mg 80mg 81mg 75mg 1000mg 0 . 4mg 40mg

Figure 4: Demonstration of Clinical Events Outputs using A Synthetic Letter.


InsightBuddy AI
File Edit View Window Help
InsightBuddy AI
How To Use
Upload File
(.pdf, .docx, .doc or .txt file)

Breaking News_ Global Tech Summit 2024.pdf

Context Length 3

Only Entities

Breaking News: GlobalTechSummit MISC 2024 James PER went to the

a pizza with PabloRomero PER SiliconValley LOC is abuzz with

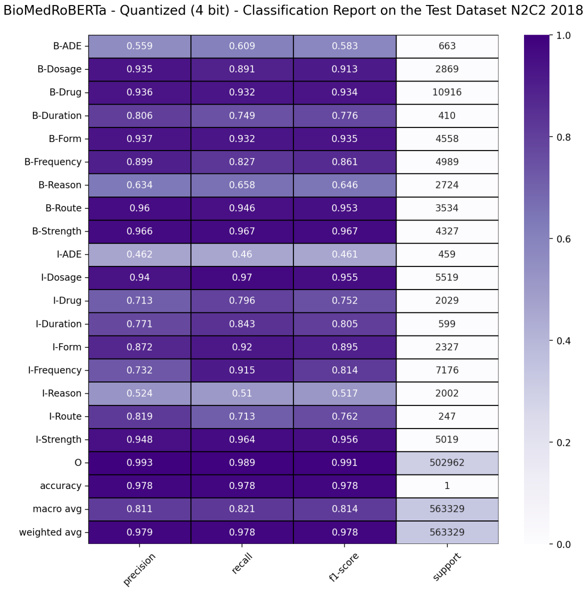
as the annual GlobalTechSummit MISC kicks off tomorrow at the MosconeCenter LOC in SanFrancisco LOC . This year's

, organized by TechCorp ORG

Named Entities

MISC	Global Tech Summit Global Tech Summit Future of AI and Space Exploration Alley Green Tech for a Sustainable Future Tech Innovator Awards Most Innovative AI Application Global Tech Summit
PER	James Pablo Romero Elon Musk Sundar Pichai Satya Nadella Lisa Su Vitalik Buterin Kara Swisher Jennifer Lee
ORG	TechCorp International SpaceX Tesla Google Microsoft AMD Ethereum IBM Intel Startup NewGen Robotics Greenpeace World Wildlife Fund MIT San Francisco Chamber of Commerce
LOC	Silicon Valley Moscone Center San Francisco Tokyo Japan Golden Gate City

Figure 5: Context-awareness Feature using Window Parameter around the Entity


BioMedRoBERTa - Quantized (4 bit) - Classification Report on the Test Dataset N2C2 2018

	precision	recall	f1-score	support
B-ADE	0.559	0.609	0.583	663
B-Dosage	0.935	0.891	0.913	2869
B-Drug	0.936	0.932	0.934	10916
B-Duration	0.806	0.749	0.776	410
B-Form	0.937	0.932	0.935	4558
B-Frequency	0.899	0.827	0.861	4989
B-Reason	0.634	0.658	0.646	2724
B-Route	0.96	0.946	0.953	3534
B-Strength	0.966	0.967	0.967	4327
I-ADE	0.462	0.46	0.461	459
I-Dosage	0.94	0.97	0.955	5519
I-Drug	0.713	0.796	0.752	2029
I-Duration	0.771	0.843	0.805	599
I-Form	0.872	0.92	0.895	2327
I-Frequency	0.732	0.915	0.814	7176
I-Reason	0.524	0.51	0.517	2002
I-Route	0.819	0.713	0.762	247
I-Strength	0.948	0.964	0.956	5019
O	0.993	0.989	0.991	502962
accuracy	0.978	0.978	0.978	1
macro avg	0.811	0.821	0.814	563329
weighted avg	0.979	0.978	0.978	563329

Figure 6: BioMedRoBERTa Quantised Model Eval.

focused applications. The three Voting ensemble approaches (Average, First logit, and Max logit) showed slight numerical improvements in F1 scores but these differences did not reach statistical significance ($p > 0.05$).

For robust evaluation, we calculated 95% confidence intervals using bootstrap resampling on the test dataset. This involved randomly sampling 95% of the sentences with replacement, calculating the F1 score for each resampled dataset, and repeating this process 200 times per model. The standard deviation across these iterations provides a measure of performance stability across different subsets of the data. These findings demonstrate that while some ensemble configurations can offer consistent improvements, performance gains are sensitive to both the specific ensemble strategy employed and the evaluation methodology. Our comprehensive comparison provides valuable insights for researchers applying ensemble approaches to clinical named entity recognition tasks.

6 Entity Linking: BNF and SNOMED

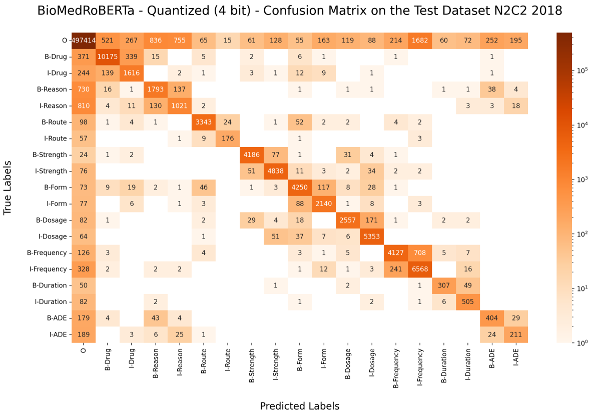
To map the identified named entities into the clinical knowledge base. We use the existing code mapping sheet from the British National Formulary (BNF) web between SNOMED-CT, BNF, dm+d, and ICD [2]. We preprocessed the SNOMED code from 377,834 to 10,804 to filter repeated examples between the mapping of SNOMED and BNF. We looked for non-drug words present in the text, then we filtered the drugs further by seeing if words like ['system', 'ostomy', 'bag', 'filter', 'piece', 'closure'] were present in the text, and if so, it was discarded.

For SNOMED CT mapping, we applied a fuzzy search to the cleaned mapping list with drug names. Then the SNOMED CT code will be added to the searching function on the SNOMED CT web, whenever there is a match. For BNF mapping, the linking function uses keyword search to retrieve the BNF website with corresponding drugs, due to its different searching features in comparison to the SNOMED-CT web page. Potential users can select whichever is suitable to their preferences between the two clinical knowledge bases (KBs), Figure 11.

7 InsightBuddy-AI Desktop Application

We illustrate the Desktop Applications of InsightBuddy-AI in Figure 4 and 14, for demonstration of clinical event recognition using a synthetic letter via 1) loading our pre-trained model and common NER categories via 2) loading a Huggingface NER model. There is also a **sliding window feature** called "context length" to allow flexible length of context around the entities visible to users, as in

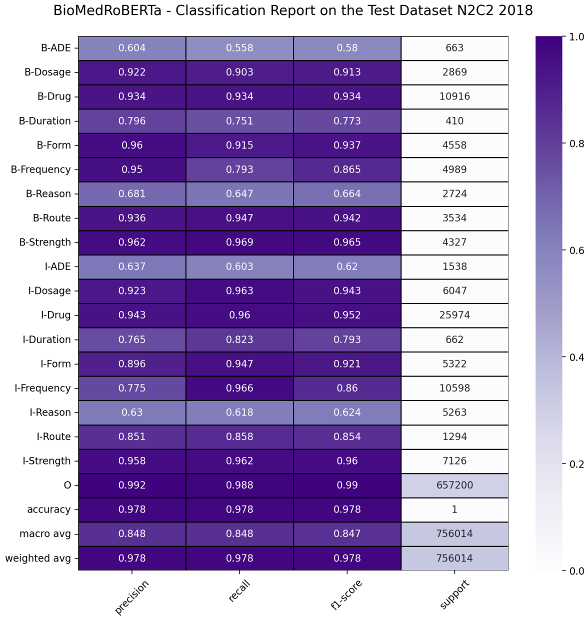
[2] https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping



BioMedRoBERTa - Quantized (4 bit) - Confusion Matrix on the Test Dataset N2C2 2018

True Labels \ Predicted Labels	O	B-Drug	I-Drug	B-Reason	I-Reason	B-Route	I-Route	B-Strength	I-Strength	B-Form	I-Form	B-Dosage	I-Dosage	B-Frequency	I-Frequency	B-Duration	I-Duration	B-ADE	I-ADE
O	497414	521	267	836	755	65	15	61	128	55	163	119	88	214	1682	60	72	252	195
B-Drug	371	10175	339	15		5		2		6	1			1				1	
I-Drug	244	139	1616		2	1		3	1	12	9		1					1	
B-Reason	730	16	1	1793	137					1		1	1			1	1	38	4
I-Reason	810	4	11	130	1021	2											3	3	18
B-Route	98	1	4	1		3343	24		1	52	2	2		4	2				
I-Route	57				1	9	176			1					3				
B-Strength	24	1	2					4186	77	1		31	4	1					
I-Strength	76							51	4838	11	3	2	34	2	2				
B-Form	73	9	19	2	1	46		1	3	4250	117	8	28	1					
I-Form	77		6		1	3				88	2140	1	8		3				
B-Dosage	82	1				2		29	4	18		2557	171	1		2	2		
I-Dosage	64					1			51	37	7	6	5353						
B-Frequency	126	3				4				3	1	5		4127	708	5	7		
I-Frequency	328	2		2	2					1	12	1	3	241	6568		16		
B-Duration	50								1			2			1	307	49		
I-Duration	82			2						1			2		1	6	505		
B-ADE	179	4		43	4													404	29
I-ADE	189		3	6	25	1												24	211

Figure 7: BioMedRoBERTa Quantised Eval Confusion Matrix.



BioMedRoBERTa - Classification Report on the Test Dataset N2C2 2018

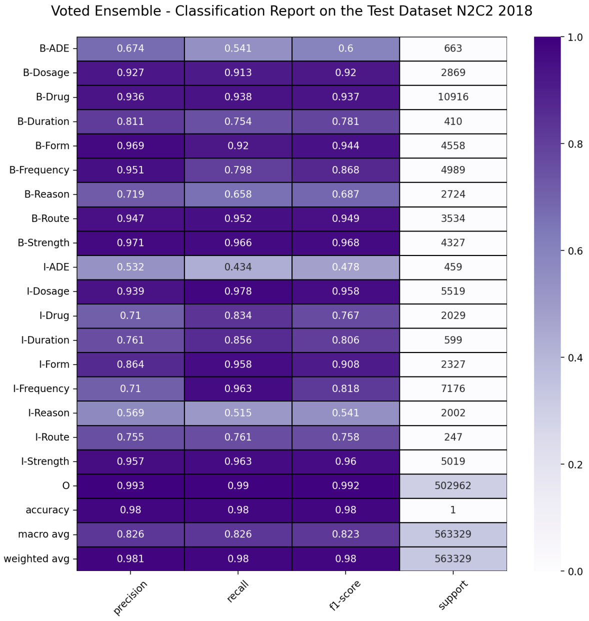
	precision	recall	f1-score	support
B-ADE	0.604	0.558	0.58	663
B-Dosage	0.922	0.903	0.913	2869
B-Drug	0.934	0.934	0.934	10916
B-Duration	0.796	0.751	0.773	410
B-Form	0.96	0.915	0.937	4558
B-Frequency	0.95	0.793	0.865	4989
B-Reason	0.681	0.647	0.664	2724
B-Route	0.936	0.947	0.942	3534
B-Strength	0.962	0.969	0.965	4327
I-ADE	0.637	0.603	0.62	1538
I-Dosage	0.923	0.963	0.943	6047
I-Drug	0.943	0.96	0.952	25974
I-Duration	0.765	0.823	0.793	662
I-Form	0.896	0.947	0.921	5322
I-Frequency	0.775	0.966	0.86	10598
I-Reason	0.63	0.618	0.624	5263
I-Route	0.851	0.858	0.854	1294
I-Strength	0.958	0.962	0.96	7126
O	0.992	0.988	0.99	657200
accuracy	0.978	0.978	0.978	1
macro avg	0.848	0.848	0.847	756014
weighted avg	0.978	0.978	0.978	756014

Figure 8: BioMedRoBERTa Eval at Sub-word Level on n2c2 2018 test data.

Figure 5. For **Clinical Coding** (entity linking) options, the desktop application can currently directly link the extracted entities to BNF and SNOMED-CT. The INSIGHTBUDDY-AI software supports both Mac and Windows systems.

8 Discussion and Conclusion

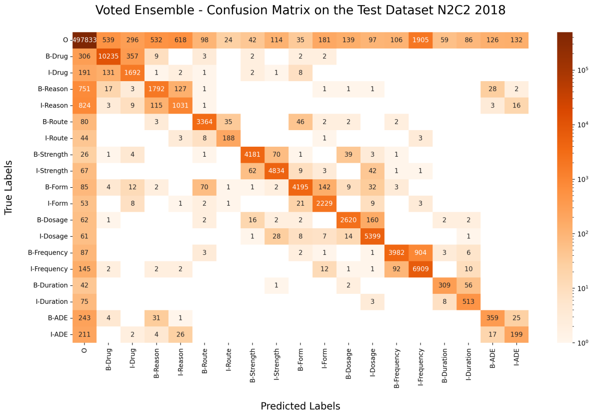
In this paper, we investigated **Stacked Ensemble** and **Voting Ensemble** on *medical named entity recognition* tasks using eight pretrained LMs from both general and biomed/clinical domains. Our experiments show that our fine-tuned best individual models outperformed the state-of-the-art on standard shared task data n2c2-2018. The two ensemble strategies using output logits and one-hot encoding further improved the model performances. We carried out model quantisation and again improved the model performances, especially on Precision scores, while reducing the model size by 75%. We carried out **statistical significance** testing and the results show that the word-level MER ensemble significantly improved over the baseline model (p=0.048). We offer desktop applications and user interfaces for individual fine-tuned models where we added the entity linking/normalisation function to BNF and SNOMED CT clinical knowledge base. We call the package INSIGHTBUDDY-AI, which is released publicly for free research use.



Voted Ensemble - Classification Report on the Test Dataset N2C2 2018

	precision	recall	f1-score	support
B-ADE	0.674	0.541	0.6	663
B-Dosage	0.927	0.913	0.92	2869
B-Drug	0.936	0.938	0.937	10916
B-Duration	0.811	0.754	0.781	410
B-Form	0.969	0.92	0.944	4558
B-Frequency	0.951	0.798	0.868	4989
B-Reason	0.719	0.658	0.687	2724
B-Route	0.947	0.952	0.949	3534
B-Strength	0.971	0.966	0.968	4327
I-ADE	0.532	0.434	0.478	459
I-Dosage	0.939	0.978	0.958	5519
I-Drug	0.71	0.834	0.767	2029
I-Duration	0.761	0.856	0.806	599
I-Form	0.864	0.958	0.908	2327
I-Frequency	0.71	0.963	0.818	7176
I-Reason	0.569	0.515	0.541	2002
I-Route	0.755	0.761	0.758	247
I-Strength	0.957	0.963	0.96	5019
O	0.993	0.99	0.992	502962
accuracy	0.98	0.98	0.98	1
macro avg	0.826	0.826	0.823	563329
weighted avg	0.981	0.98	0.98	563329

Figure 9: word-level grouping ensemble, max logit voting Eval on n2c2 2018 test data.



Voted Ensemble - Confusion Matrix on the Test Dataset N2C2 2018

True Labels \ Predicted Labels	O	B-Drug	I-Drug	B-Reason	I-Reason	B-Route	I-Route	B-Strength	I-Strength	B-Form	I-Form	B-Dosage	I-Dosage	B-Frequency	I-Frequency	B-Duration	I-Duration	B-ADE	I-ADE
O	497833	539	296	532	618	98	24	42	114	35	181	139	97	106	1905	59	86	126	132
B-Drug	306	10235	357	9		3		2		2	2								
I-Drug	191	131	1692	1	2	1		2	1	8									
B-Reason	751	17	3	1792	127	1					1	1	1					28	2
I-Reason	824	3	9	115	1031	1												3	16
B-Route	80			3		3364	35			46	2	2		2					
I-Route	44				3	8	188				1				3				
B-Strength	26	1	4			1		4181	70	1		39	3	1					
I-Strength	67							62	4834	9	3		42	1	1				
B-Form	85	4	12	2		70	1	1	2	4195	142	9	32	3					
I-Form	53		8		1	2	1			21	2229		9		3				
B-Dosage	62	1				2		16	2	2		2620	160			2	2		
I-Dosage	61							1	28	8	7	14	5399				1		
B-Frequency	87					3				2	1		1	3982	904	3	6		
I-Frequency	145	2		2	2						12	1	1	92	6909		10		
B-Duration	42								1			2				309	56		
I-Duration	75												3			8	513		
B-ADE	243	4		31	1													359	25
I-ADE	211		2	4	26													17	199

Figure 10: word-level ensemble max-logit voting Eval confusion matrix on n2c2 2018 test data.

Limitations

The affiliated entity linking / clinical coding part of our software InsightBuddyAI was manually verified by ourselves qualitatively with some sampled medical terms, especially drug names. It would be more accurate to 1) quantitatively evaluate such entity linking result, as well as 2) a systematic qualitative assessment such as by multiple annotators (clinical coders) with the measurement of agreement levels. For option 2), it is costly to carry out such an experiments. For option 1), we are still looking for any publicly available data set for such purposes.

At the publication stage, we are informed of the related software implementation in this domain from Johnsnowlabs [3] on Clinical NER. While this is a commercialised company developing NLP packages for healthcare, it is worthy in the future to carry out some comparisons on experimental performances using the same shared task data. On the other hand, it is also possible that they already integrated the shared task data into their system pre-trainings.

In addition, a more detailed *error analysis*, particularly for specific entity types or challenging cases, would help determine whether improvements are consistent across all medication attributes. The current study does not compare ensemble models with *decoder-only large language models* (LLMs), such as GPT-4 or BioMistral, demonstrating strong zero-shot and fine-tuned performance. It is useful to integrate such comparisons in the future, even though this is already an extended investigation with more findings based on our initial software release IndightBuddy-AI (Romero et al., 2025).

Acknowledgements

We thank the reviewers' valuable comments, which have made our work much better. LH and GN are grateful for the support from the grant "Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease", and the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC). LH is grateful for the 4D Picture EU project (https://4dpicture.eu/) on supporting the journey of cancer patients. PR was supported by the University of Manchester Urenco Internships grant as part of an outreach programme.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Samuel Belkadi, Lifeng Han, Yuping Wu, and Goran Nenadic. 2023. Exploring the value of pre-trained language models for clinical named entity recognition. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3660–3669.

Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Funda Güneş, Russ Wolfinger, and Pei-Yi Tan. 2017. Stacked ensemble models for improved prediction accuracy. In *Proc. Static Anal. Symp*, pages 1–19.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

[3] https://demo.johnsnowlabs.com/healthcare/NER_CLINICAL/

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Meizhi Ju, Nhung TH Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. *Journal of the American Medical Informatics Association*, 27(1):22–30.

Youngjun Kim and Stéphane M Meystre. 2020. Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, 27(1):31–38.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jinsu Lee, Sang-Kwang Lee, and Seong-Il Yang. 2018. An ensemble method of cnn models for object detection. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 898–901.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.

Ammar Mohammed and Rania Kora. 2022. An effective ensemble deep learning framework for text classification. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part A):8825–8837.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nona Naderi, Julien Knafou, Jenny Copara, Patrick Ruch, and Douglas Teodoro. 2021. Ensemble of deep masked language models for effective named entity recognition in health and life science corpora. *Frontiers in research metrics and analytics*, 6:689803.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Pablo Romero, Lifeng Han, and Goran Nenadic. 2025. Insightbuddy-ai: Medication extraction and entity linking using pre-trained language models and ensemble learning. In *NAACL-SRW, Forthcoming*, New Mexico, USA. ACL.

Hager Saleh, Sherif Mostafa, Lubna Abdelkareim Gabralla, Ahmad O. Aseeri, and Shaker El-Sappagh. 2022. Enhanced arabic sentiment analysis using a novel stacking ensemble of hybrid and deep learning models. *Applied Sciences*, 12(18).

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

A Diagrams on System Details

More details on Stacked Ensemble are listed in Figure 12 and 13 on training strategy and one-hot encoding. Figure 11 shows the entity linking / coding diagram.

B Further Analysis on Models and Scores

B.1 Word-level vs Sub-word Level scores

From word-level ensemble result in Figure 9, it says that the ensembled model can achieve word-level evaluation scores 0.826, 0.826, and 0.823 for macro P/R/F1, which is close to sub-word level best model 0.847 F1. We can see that at word-level evaluation, there are 563,329 support tokens in Figure 9, vs sub-word level 756,014 tokens in Figure 8.

Word-level ensemble voting, max-logit voting > first-logit > average-logit, as shown in Table 4, with Macro F1 scores (0.8232, 0.8229, 0.8227) respectively, which are very close though. They have the same weighted average F1 and Accuracy scores (0.9798, 0.9796) respectively.

B.2 Ensemble: Stacked using output logits (non one-hot)

When we used the ‘output logits’ instead of ‘one-hot encoding’ for stacked ensemble, as we dis-

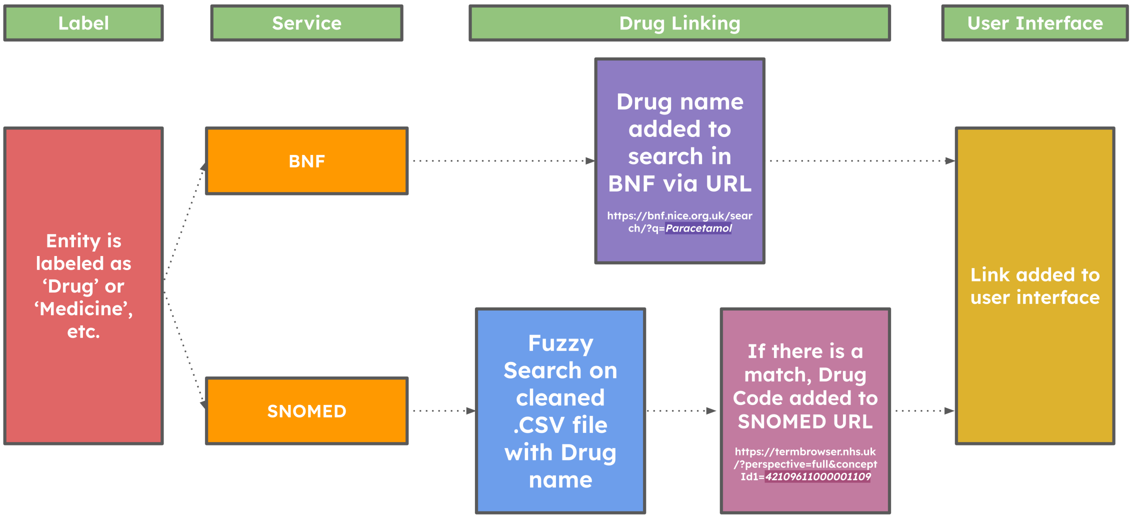

Label
Service
Drug Linking
User Interface
Entity is labeled as 'Drug' or 'Medicine', etc.
BNF
SNOMED
Drug name added to search in BNF via URL
https://bnf.nice.org.uk/search/?q=Paracetamol
Fuzzy Search on cleaned .CSV file with Drug name
If there is a match, Drug Code added to SNOMED URL
https://termbrowser.nhs.uk/?perspective=full&conceptId1=421096110000011O9
Link added to user interface

Figure 11: ENTITYLINKING: function illustration for mapping to both BNF and SNOMED-CT

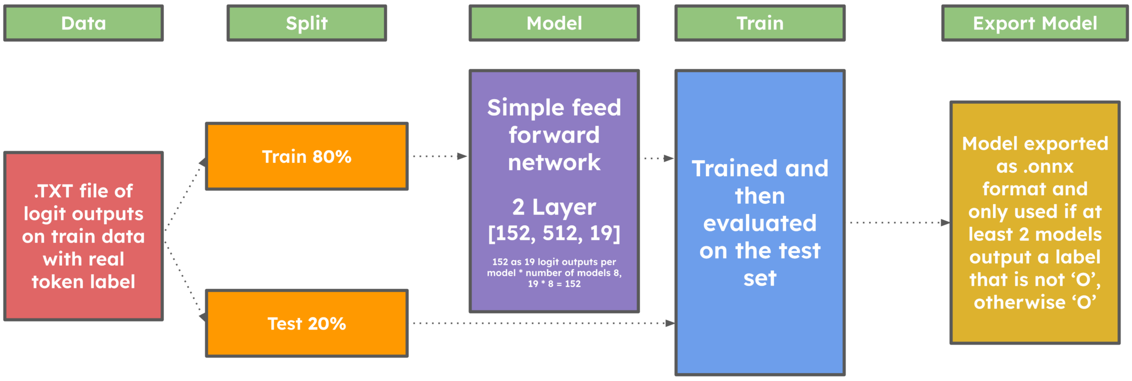

Data
Split
Model
Train
Export Model
.TXT file of logit outputs on train data with real token label
Train 80%
Test 20%
Simple feed forward network
2 Layer [152, 512, 19]
152 as 19 logit outputs per model * number of models 8, 19 * 8 = 152
Trained and then evaluated on the test set
Model exported as .onnx format and only used if at least 2 models output a label that is not 'O', otherwise 'O'

Figure 12: STACKEDENSEMBLE: training strategy.

cussed in the methodology section, it will lead to overfitting issues. We use the Max logit stacked ensemble as an example, which shows that the Stacked Ensemble using output logits produced much lower evaluation scores macro avg (0.6863, 0.7339, 0.6592) than the voting mechanism macro avg (0.8261, 0.8259, 0.8232) for (P, R, F1).

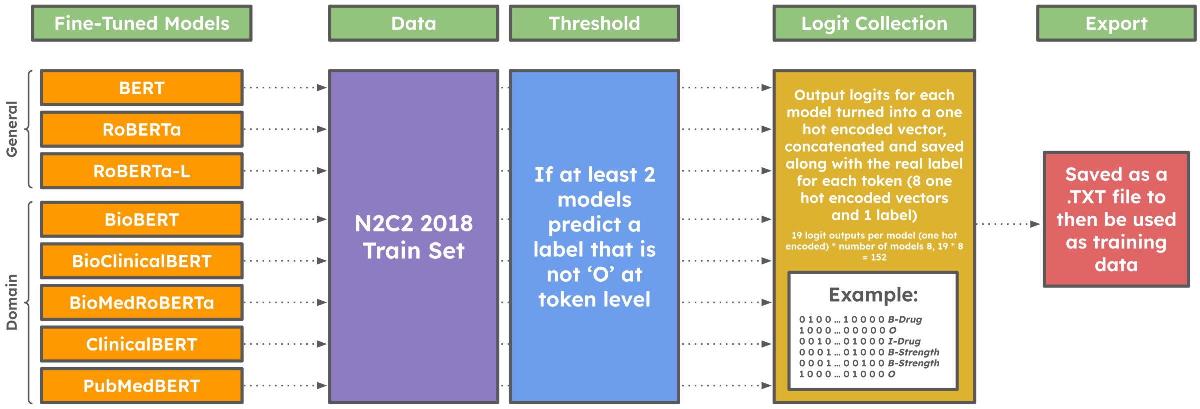
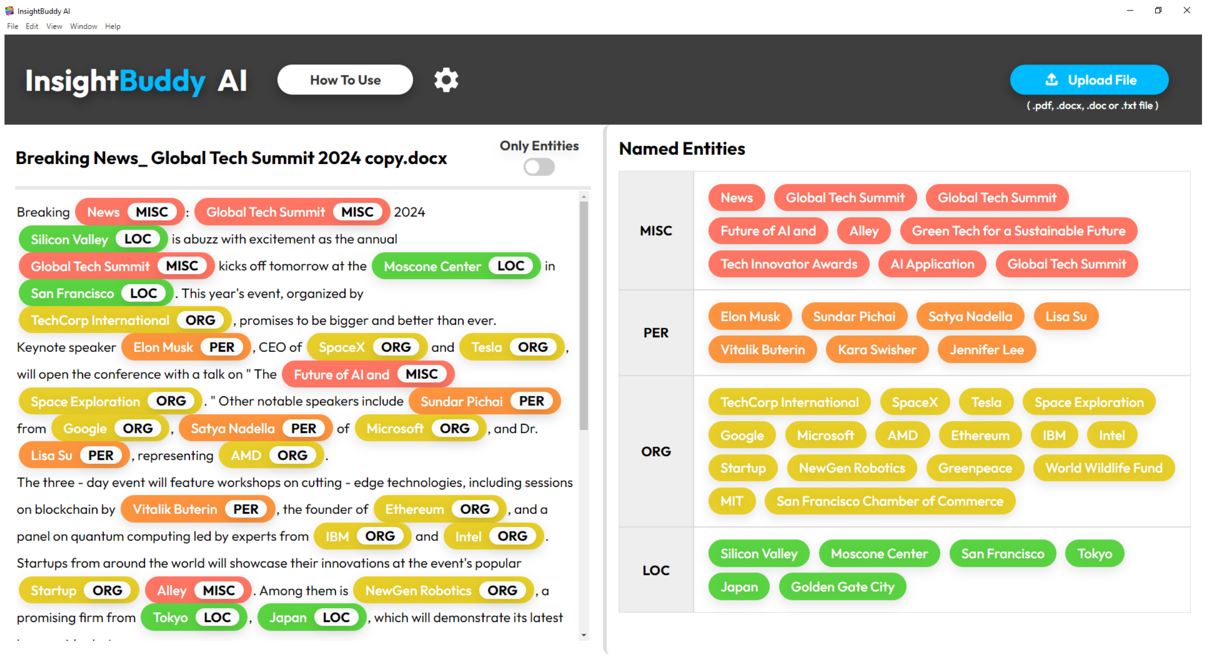

Fine-Tuned Models
Data
Threshold
Logit Collection
Export
General
Domain
BERT
RoBERTa
RoBERTa-L
BioBERT
BioClinicalBERT
BioMedRoBERTa
ClinicalBERT
PubMedBERT
N2C2 2018 Train Set
If at least 2 models predict a label that is not 'O' at token level
Output logits for each model turned into a one hot encoded vector, concatenated and saved along with the real label for each token (8 one hot encoded vectors and 1 label)
19 logit outputs per model (one hot encoded) * number of models 8, 19 * 8 = 152
Example:
0 1 0 0 ... 1 0 0 0 0 B-Drug
1 0 0 0 ... 0 0 0 0 0 O
0 0 1 0 ... 0 1 0 0 0 I-Drug
0 0 0 1 ... 0 1 0 0 0 B-Strength
0 0 0 1 ... 0 0 1 0 0 B-Strength
1 0 0 0 ... 0 1 0 0 0 O
Saved as a .TXT file to then be used as training data

Figure 13: STACKEDENSEMBLE: one-hot encoding data.



Breaking News_ Global Tech Summit 2024 copy.docx

Breaking News MISC : Global Tech Summit MISC 2024 Silicon Valley LOC is abuzz with excitement as the annual Global Tech Summit MISC kicks off tomorrow at the Moscone Center LOC in San Francisco LOC . This year's event, organized by TechCorp International ORG , promises to be bigger and better than ever. Keynote speaker Elon Musk PER , CEO of SpaceX ORG and Tesla ORG , will open the conference with a talk on " The Future of AI and MISC Space Exploration ORG . " Other notable speakers include Sundar Pichai PER from Google ORG , Satya Nadella PER of Microsoft ORG , and Dr. Lisa Su PER , representing AMD ORG .

The three - day event will feature workshops on cutting - edge technologies, including sessions on blockchain by Vitalik Buterin PER , the founder of Ethereum ORG , and a panel on quantum computing led by experts from IBM ORG and Intel ORG .

Startups from around the world will showcase their innovations at the event's popular Startup ORG Alley MISC . Among them is NewGen Robotics ORG , a promising firm from Tokyo LOC , Japan LOC , which will demonstrate its latest

Named Entities

MISC	News, Global Tech Summit, Global Tech Summit, Future of AI and, Alley, Green Tech for a Sustainable Future, Tech Innovator Awards, AI Application, Global Tech Summit
PER	Elon Musk, Sundar Pichai, Satya Nadella, Lisa Su, Vitalik Buterin, Kara Swisher, Jennifer Lee
ORG	TechCorp International, SpaceX, Tesla, Space Exploration, Google, Microsoft, AMD, Ethereum, IBM, Intel, Startup, NewGen Robotics, Greenpeace, World Wildlife Fund, MIT, San Francisco Chamber of Commerce
LOC	Silicon Valley, Moscone Center, San Francisco, Tokyo, Japan, Golden Gate City

Figure 14: Loading Any Huggingface NER model: example outcome with typical (PER, LOC, ORG, MISC) label set