INSIGHTBUDDY-AI: Medication Extraction and Entity Linking using Pre-Trained Language Models and Ensemble Learning

Pablo Romero
MMU
Oxford Rd
Greater Manchester, UK
pablo2004romero@gmail.com

Lifeng Han*
LIACS & LUMC
Leiden University, NL
University of Manchester, UK
l.han@lumc.nl

Goran Nenadic
University of Manchester
Oxford Rd
Greater Manchester, UK
g.nenadic@manchester.ac.uk

* *corresponding author*

Abstract

This paper presents our system, INSIGHTBUDDY-AI, designed for extracting medication mentions and their associated attributes, and for linking these entities to established clinical terminology resources, including SNOMED-CT, the British National Formulary (BNF), ICD, and the Dictionary of Medicines and Devices (dm+d). To perform medication extraction, we investigated various ensemble learning approaches, including stacked and voting ensembles (using first, average, and max voting methods) built upon eight pre-trained language models (PLMs). These models include general-domain PLMs—BERT, RoBERTa, and RoBERTa-Large—as well as domain-specific models such as BioBERT, BioClinicalBERT, BioMedRoBERTa, ClinicalBERT, and PubMedBERT. The system targets the extraction of drug-related attributes such as adverse drug effects (ADEs), dosage, duration, form, frequency, reason, route, and strength. Experiments conducted on the n2c2-2018 shared task dataset demonstrate that ensemble learning methods outperformed individually fine-tuned models, with notable improvements of 2.43% in Precision and 1.35% in F1-score. We have also developed cross-platform desktop applications for both entity recognition and entity linking, available for Windows and macOS. The INSIGHTBUDDY-AI application is freely accessible for research use at https://github.com/HECTA-UoM/InsightBuddy-AI.

1 Introduction

Extracting information about medications and their associated attributes is a crucial task in natural language processing (NLP) for the clinical domain, particularly to enhance digital healthcare solutions. Traditionally, clinicians and healthcare professionals have manually performed clinical coding to translate medical events—such as diseases, medications, and treatments—into standardised terminologies like ICD and SNOMED. This manual process is often labour-intensive and prone to human error, potentially compromising accuracy. Automating the extraction of medication-related information paves the way for automatic mapping of these terms to existing medical terminologies, enabling automated clinical coding. Given the potential of this approach, numerous NLP models have been applied in recent years to tasks such as medication mining and clinical coding—though typically in isolation. In this study, we unify these tasks by 1) developing a pipeline that integrates medication and attribute extraction (including dosage, route, strength, adverse effects, frequency, duration, form, and reason) with automated clinical coding. Furthermore, 2) we explore ensemble learning techniques—specifically Stacking and Voting—across a diverse set of NLP models fine-tuned for named entity recognition (NER). These include general-domain models like BERT, RoBERTa, and RoBERTa-L, as well as clinical-domain models such as BioBERT, BioClinicalBERT, BioMedRoBERTa, ClinicalBERT, and PubMedBERT. Our approach allows practitioners to bypass the challenge of selecting individual models for clinical NER tasks; instead, they can incorporate newer models into the ensemble framework to evaluate their effectiveness.

2 Literature Review and Related Work

Named Entity Recognition (**NER**) plays a vital role in extracting essential information from unstructured texts, such as medical correspondence. The inherent complexity and context sensitivity of medical language make accurate entity extraction particularly challenging. Traditional NER methods, including rule-based approaches, have had limited success in capturing the rich contextual details required for clinical applications (Nadeau and Sekine, 2007). The introduction of deep learning methods, notably Long Short-Term Memory (LSTM) net-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 18–27

works, led to considerable improvements in NER performance (Graves and Schmidhuber, 2005), particularly through their capacity to model long-range dependencies in text. Nevertheless, these models continued to face difficulties with infrequent entities and intricate contextual relationships commonly found in **clinical notes**. The emergence of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) brought a major breakthrough across multiple NLP tasks, including NER. BERT leverages masked language modelling on extensive corpora to learn rich token-level representations, which can then be fine-tuned with an added classification layer for token-level predictions. However, since BERT is pre-trained on general-domain corpora (Wikipedia and books), its effectiveness on specialised medical texts has been constrained. This limitation has spurred the development of **domain-specific** BERT variants. Examples include BioBERT (Lee et al., 2019), trained on large biomedical datasets; ClinicalBERT (Wang et al., 2023), fine-tuned on electronic health records from three million patients following pre-training on 1.2 billion words across various disease contexts; and Med-BERT (Rasmy et al., 2021), all of which have shown improved results for medical NER tasks due to their focused training in the healthcare domain. Other notable versions of ClinicalBERT include (Huang et al., 2019) and (Alsentzer et al., 2019), both trained on data from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset (Johnson et al., 2016).

Despite these advancements, **single-model solutions** still encounter obstacles due to the inherent variability and complexity in clinical language, as demonstrated in the comparative evaluation in (Belkadi et al., 2023), which tested models including BERT, ClinicalBERT, BioBERT, and custom-trained Transformers. To mitigate these limitations, **ensemble** techniques have gained traction. Successfully applied in other areas such as computer vision (Lee et al., 2018), ensemble methods combine multiple models to exploit their complementary strengths and reduce their individual shortcomings. In the NER domain, ensembling has led to improved outcomes, as evidenced by (Naderi et al., 2021), who demonstrated significant performance gains by applying ensemble strategies to health and life sciences corpora. Naderi et al. (2021) employed max voting across models for word-level data in biology, chemistry, and medicine. However, their work focused on French for the clinical/medical NER domain using the DEFT benchmark, while English data were only utilised for the biology and chemistry domains. Among ensemble methods, two of the most widely adopted are voting and stacked ensembles: 1) **Maximum voting**, where each model has equal influence on the final decision—as used in (Naderi et al., 2021)—selects the label with the most votes. 2) **Stacking**, a more advanced method introduced by Wolpert (1992), involves training a meta-model on the outputs of base models to learn complex relationships between predictions. For instance, (Saleh et al., 2022) showed that stacking, when implemented with a support vector machine (SVM), improved sentiment analysis performance. In our work, we opt for a simple feed-forward network that maps the ensemble outputs to final predictions. Additional examples of stacking can be found in (Mohammed and Kora, 2022; Güneş et al., 2017). While ensemble strategies have shown promise across various NER applications, their applicability to clinical NER—especially with complex datasets like n2c2 2018 (Henry et al., 2020)—has yet to be thoroughly explored. This study **seeks to bridge that gap** by examining whether ensemble approaches, particularly stacking and voting, can enhance NER performance on clinical texts and help overcome the challenges associated with individual model limitations.

3 Methodologies

The overall architecture of INSIGHTBUDDY is illustrated in Figure 1, which outlines the base models used from both general and clinical domains. From the general domain, we included 1) BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and RoBERTa-Large; and from biomedical/clinical domains, 2) 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). All eight models were fine-tuned using the same hyperparameters and training set from the n2c2-2018 shared task, following data pre-processing. The performance of each model was first evaluated individually using the n2c2-2018 test set, providing a baseline comparison. Subsequently, ensemble learning was applied to the outputs of all models. We then introduced an **entity linking** component to map the extracted medical entities into standardised clinical terminologies. Initially, we

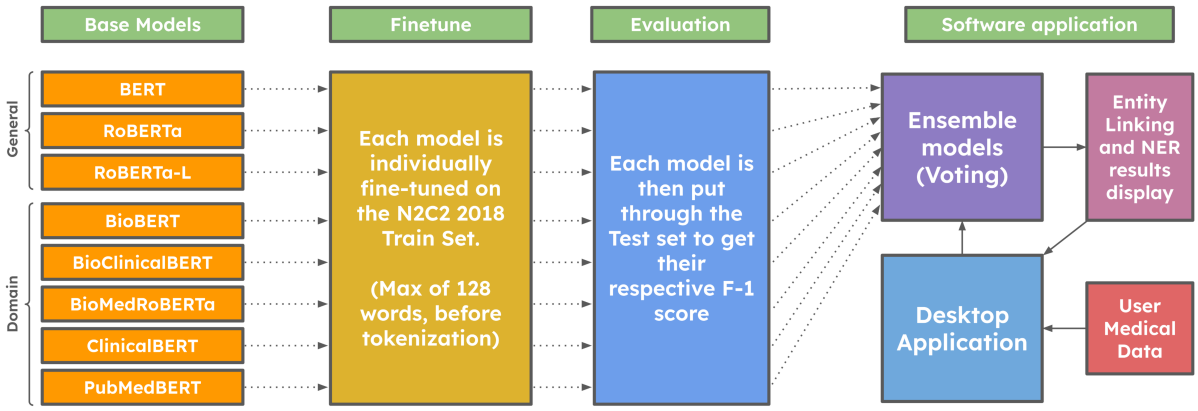

Figure 1: INSIGHTBUDDY Framework Pipeline: This diagram illustrates the full pipeline, including individual NER model fine-tuning, ensemble integration, entity linking, and desktop applications in both Windows and Mac systems. The base models are drawn from two domains: general and biomedical. Data pre-processing involves splitting the input sequence either at the first full stop (".") occurring after the 100th word or, if none is found, truncating at 128 words. Fine-tuning is carried out using identical hyperparameter settings across all eight models. Ensembling is performed using various strategies, which are detailed in Figure 3. Entity linking connects extracted entities to clinical knowledge bases (KBs), specifically BNF and SNOMED CT.

used **SNOMED-CT and BNF** as our knowledge bases (KB), which were further aligned with ICD and dm+d.

For pre-processing, the input text was segmented into chunks of up to 128 tokens. If a full stop (".") appeared between the 100th and 128th word, the chunk was cut at that punctuation mark. To explain our ensemble-learning approach, we present the InsightBuddy ensemble diagram in Figure 3. The initial outputs from each of the eight fine-tuned NER models are in sub-word format, as per their tokenisation strategy. For example, the word "Paracetamol" may be tokenised as "Para ##ce ##tam ##ol". Therefore, our first step is to **reconstruct** words from sub-word tokens for practical usage and voting. However, since each sub-word receives a potentially different label, discrepancies often occur within the same word. To resolve this, we implemented three grouping strategies: first-token voting, max-token voting, and average voting. In the *first-token voting* method, the label of the first sub-word is applied to the entire word. For instance, if "Para" is labelled as "B-Drug", then "Paracetamol" will be assigned the same label, regardless of labels on subsequent sub-words. In the *max-token voting* method, the label with the highest logit score among the sub-words is assigned to the word—reflecting the model's highest confidence in that prediction. The *average voting* approach computes the mean of logits across all sub-words, from which the label for the full word is derived. Regarding **word-level ensemble** learning, we explore a classical **voting** approach with two specific strategies: The ">=4 or O" strategy assigns the majority label if at least four models agree. If no majority exists, the label "O" (non-entity) is used by default to signify context words. The max-voting strategy selects the most frequently predicted label, regardless of how many models it came from (e.g. 2, 3, or 4 votes). In cases of a tie (e.g. two labels each receiving three votes from six models), we resolve it either alphabetically or randomly.

We also depict the STACKED-ENSEMBLE approach in Figure 2. During training, the data is split into 80% for training and 20% for testing the ensemble model. Output data from base models is only used when at least two models assign a label other than "O"; otherwise, "O" is kept and the token is excluded from stacked training data. For the stacked model's input, we convert each model's output logits into one-hot encoded vectors, then concatenate them alongside the true label of each token. As we use eight models, the training input consists of eight one-hot vectors and one label. Each vector is of length 19 (representing 19 possible labels), containing a single '1' at the predicted label's index and '0' elsewhere. As a result, each training sample contains 8 vectors × 19 values = 152 values, with exactly eight '1's and the remaining 144 being '0's. We choose to use

Voting Average Ensemble word level (BIO)			
Metric	P	R	F1
accuracy		0.9796	
macro avg	0.8253	0.8256	0.8227
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
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**
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
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***
weighted avg	0.9840	0.9839	0.9838

Table 1: Word-level ensemble grouping results: While all three logit aggregation methods—max, first, and average—produce similar scores, max-logit voting slightly outperforms the others. The **stacked** ensemble achieves the highest **Precision**, but at the cost of lower Recall, resulting in a reduced F1 score overall. The lower section of the table presents word-level evaluation results without differentiating between B- and I-labels, based on the n2c2 2018 test dataset.

one-hot encoding instead of raw logits to reduce the risk of *overfitting*, since models often produce highly confident predictions on the data they were trained on. We provide evaluation outcomes when using "raw logits" for stacked-ensemble in Figure 7 (evaluation scores) and 8 (confusion matrix) using word-level grouping ensemble using max logit, stacked ensemble, non-one-hot encoding, where they showed lower performances. One-hot vectors help regularise training by removing this overconfidence and ensuring a more generalisable stacked model.

4 Experimental Evaluations

We employed the dataset from the n2c2-2018 shared task, which focuses on named entity recognition (NER) of adverse drug events and associated medical attributes (Henry et al., 2020). The data includes annotated labels such as ADE, Dosage, Drug, Duration, Form, Frequency, Reason, Route, and Strength in BIO tagging format, resulting in a total of 19 possible tags: 2 (B/I) for each of the 9 classes, plus 1 (O). The original dataset comprises 303 training letters and 202 testing letters. Following the data split approach by Belkadi et al. (2023), we divided the training set into a 9:1 ratio for training and validation purposes. We evaluate the models using Precision, Recall, and F1-score under both "macro" and "weighted" averaging schemes, along with overall Accuracy. The "**macro**" average gives equal importance to each class, regardless of how often it appears in the dataset, whereas the "**weighted**" average scales scores according to label frequency. We begin by reporting the results from individual fine-tuned models (sub-word level), followed by evaluations of ensemble models using various strategies (word level).

4.1 Individual Models: sub-word level

The performance of individual models post fine-tuning is presented in Table 2. Among general-domain models, RoBERTa-Large achieved the highest macro Precision (0.8489), Recall (0.8606), and F1-score (0.8538), even outperforming domain-specific models. BioMedRoBERTa emerged as the top performer among domain-specific models, with macro Precision, Recall, and F1 scores of 0.8482, 0.8477, and 0.8468, respectively. When compared to the results reported by Belkadi et al. (2023), whose ClinicalBERT-Apt model achieved macro averages of 0.842, 0.834, and 0.837, our fine-tuned ClinicalBERT model delivered comparable results (0.848, 0.825, 0.834), validating the effectiveness of our fine-tuning. Notably, our BioMedRoBERTa model outperforms theirs with macro scores. Furthermore, RoBERTa-Large achieved even higher macro scores and Accuracy of 0.9782 (Figure 4). Both BioMedRoBERTa and RoBERTa-Large thus surpass the best-performing model reported in Belkadi et al. (2023), namely ClinicalBERT-CRF, which scored 0.85, 0.829, and 0.837 with Accuracy of 0.976. Building on this, our work transitions to a focus on **word-level** evaluation, which contrasts with the sub-word emphasis seen in Belkadi et al. (2023).

4.2 Ensemble: word-level grouping (logits)

We evaluated three strategies for aggregating sub-word predictions into word-level labels: **first** logit voting, **max** logit voting, and **average** logit voting. Their results are displayed in the upper section of Table 1. The first-logit method produced a higher Recall (0.8260), while max-logit voting yielded the highest Precision (0.8261) and F1-score (0.8232),

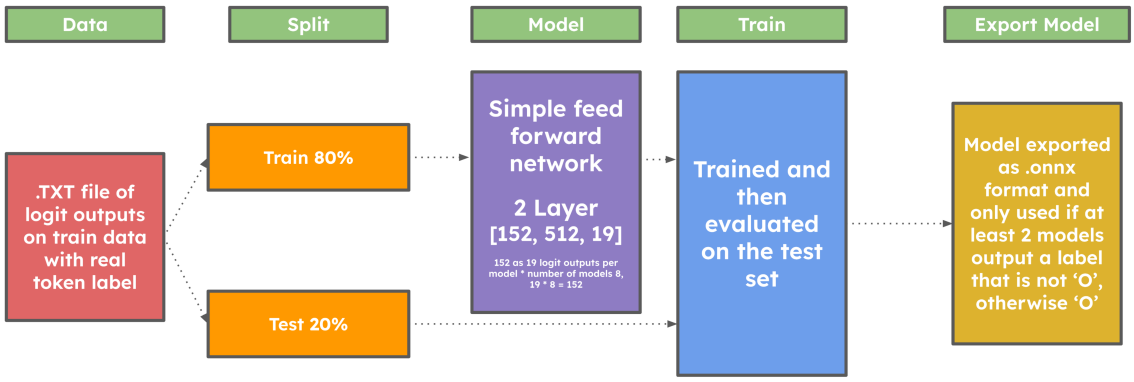

Figure 2: StackedEnsemble: training strategy.

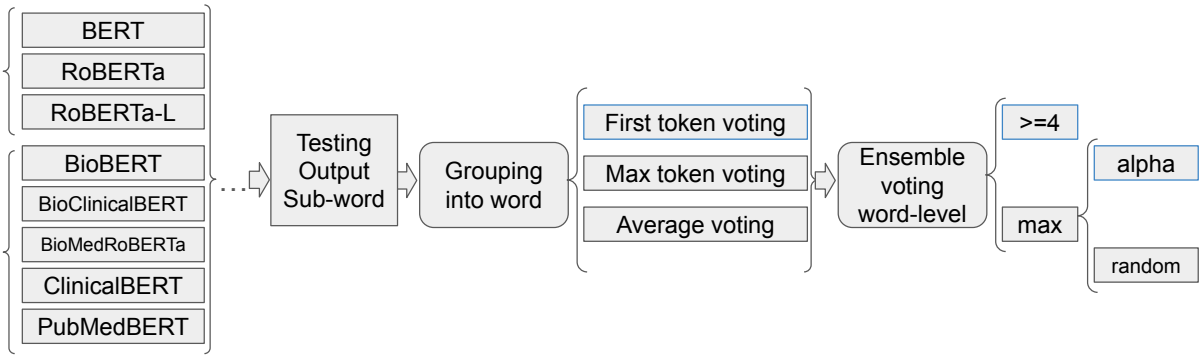

Figure 3: InsightBuddy Voted Ensemble Pipeline: Each individual NER model is fine-tuned to produce predictions at the token or sub-word level. (Note: "Logits" refer to the neural network outputs prior to applying the activation function.) The first step involves aggregating sub-word tokens into complete words using one of three strategies: selecting the label of the first sub-word, applying max-token voting, or averaging logits across sub-words. According to our results (see Table 1), the first-token approach yields higher Recall, while the other two methods slightly favour Precision. However, all three produce nearly identical F1 scores. Based on these findings, we adopt the first-token label method for further processing. For the word-level ensemble across all eight models, two voting strategies are explored: 1) majority voting—if four or more models assign the same label, it is selected; otherwise, the label defaults to "O", and 2) max voting—selecting the most frequently predicted label, regardless of count. In the case of ties (e.g. 3,3,2), we experimented with resolving ties either alphabetically or randomly. Our findings indicate that the ">=4 or O" strategy performs comparably to max + alphabetical", while "max + random" shows slightly reduced performance.

following the trend: *Max* > *First* > *Average*, based on macro F1 (0.8232, 0.8229, 0.8227). Given the marginal performance differences, we selected the first-logit voting output for the next ensemble step for its computational efficiency.

4.3 Ensemble: Voting vs Stacked (one-hot)

The Stacked Ensemble approach, which uses one-hot encoded vectors, is shown in the middle part of Table 1. It achieved a higher Precision (0.8351) compared to the best from voting ensembles (0.8261). However, its macro Recall dropped to 0.8065, whereas voting ensembles reached 0.8260. This suggests that while stacking reduced false positives, it also increased false negatives—indicating a more conservative prediction style when identifying positive cases.

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

Up to this point, evaluations have been based on strict BIO tagging—treating labels like B-Drug and I-Drug as distinct, with mismatches considered incorrect. However, in practice, the distinction between B and I tags may not be necessary for all use cases. As shown in Table 1, when we ignore the B/I prefix and evaluate based on the 9 core label types, the ensemble model at the word level significantly

Model	Macro P	Macro R	Macro F	Accuracy	Tokens(sub-words)
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.

improves. Macro Precision reaches 0.8844, Recall 0.8830, and F1 0.8821—well above the macro F1 of 0.8232 (voting-max-logit) and 0.8156 (stacked-first-logit) under strict BIO conditions.

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

As reported in Table 3, the BioMedRoBERTa model, when evaluated individually using max-logit grouping, achieved macro averages of P/R/F1 (0.8065, 0.8224, 0.8122). In contrast, the max-voting ensemble delivered (0.8261, 0.8259, 0.8232). This represents an improvement of 2.43% in Precision and 1.35% in F1-score. These gains confirm the success of ensemble voting, which enhances Precision—thus reducing the number of *false positive* predictions—while maintaining Recall, thereby preserving true positive detections.

5 Entity Linking: BNF and SNOMED

To integrate the recognised named entities with a clinical knowledge base, we utilised the existing mapping resources provided by the British National Formulary (BNF), which establish links between SNOMED-CT, BNF, dm+d, and ICD codes (available at https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping). We began by reducing the full set of 377,834 SNOMED codes to 10,804 entries through pre-processing, eliminating duplicate mappings between SNOMED and BNF. Additionally, we filtered out non-drug terms found in the text. This included removing items that contained words such as ['system', 'ostomy', 'bag', 'filter', 'piece', 'closure'], as these typically refer to medical equipment rather than pharmaceuticals. For mapping to SNOMED CT, we applied a fuzzy string-matching technique on the refined list, using drug names as search queries. When a match was found, the associated SNOMED CT code was appended and used to generate a direct link to the SNOMED CT online portal. In contrast, the BNF mapping process relied on a keyword-based search to retrieve matching entries from the BNF website. This approach was necessary due to differences in how the BNF site handles search queries compared to the SNOMED CT platform. Depending on their needs or preferences, users can choose to utilise either of these two clinical knowledge bases (KBs), as illustrated in Figure 9.

6 Discussion and Conclusion

This paper presented a pilot investigation into the application of Stacked and Voting Ensemble techniques for medical named entity recognition, utilising eight pre-trained language models (PLMs) drawn from both general-purpose and biomedical/clinical domains. Our experimental results demonstrate that the best-performing fine-tuned individual models surpassed the state-of-the-art results on the standard n2c2-2018 shared task dataset. Moreover, by incorporating ensemble approaches—specifically using output logits and one-hot encoded vectors—we achieved further performance gains, with a 2.43% improvement in Precision and a 1.35% increase in F1-score. In addition, we developed a desktop tool and user interface for our fine-tuned models, which includes an entity linking and normalisation feature that maps recognised entities to the BNF and SNOMED CT clinical knowledge bases. This tool, named INSIGHTBUDDY-AI, is publicly accessible at https://github.com/HECTA-UoM/InsightBuddy-AI.

Limitations

Ensemble approaches—particularly those involving large-scale models—can be demanding in terms of computational resources. During both training and inference, we encountered challenges related to hardware limitations. Future directions include reducing the computational load associated with ensemble learning, investigating alternative ensemble strategies, model *quantisation*, model output *significance* testing, and extending the approach to additional datasets. At present, the desktop applications support the deployment of all individual fine-tuned NER models, including any Hugging Face-compatible models. However, ensemble-based models are not yet integrated. Future work may focus on embedding ensemble learning directly into the application workflow, rather than requiring it as a separate, manual process.

Ethics

To use the n2c2 shared task data, the authors have carried out CITI training (https://physionet.org/settings/credentialing/) and gained the access to the data with user agreement.

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.

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.

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.

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.

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.

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.

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 InsightBuddy-AI Desktop Application

For **Clinical Coding** (entity linking) options, the desktop application can currently directly link the extracted entities to BNF and SNOMED-CT, as in Figure 9 from the screenshots. The INSIGHTBUDDY-AI software supports both Mac and Windows systems.

B Diagrams and Scoring Tables

B.1 Sub-word level diagrams

Sub-word level BioMedRoBERTa confusion matrix, RoBERTa-L evaluation and confusion matrix are shown in Figure 6, 4, and 5.

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

When we used the 'output logits' instead of 'one-hot encoding' for stacked ensemble, as we discussioned in the methodology section, it will lead to overfitting issues. We use the Max logit stacked ensemble as an example, in figure 7, 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). The corresponding confusion matrix from the stacked ensemble using the max logit is shown in Figure 8 with more errors spread in the image, the coloured numbers outside the diagonal line.

B.3 Individual vs Ensemble Models

The word-level performance comparisons from individual models and voting max-logit ensembles are presented in Table 3.

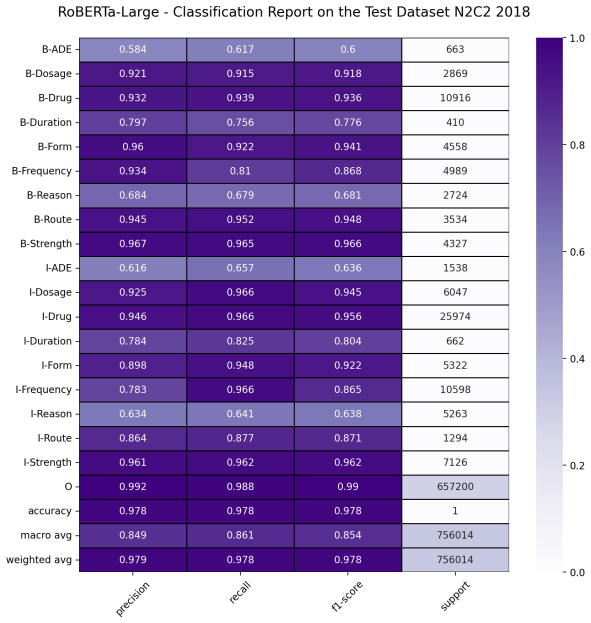

Figure 4: RoBERTa-L Eval at Sub-word Level on n2c2 2018 test data.

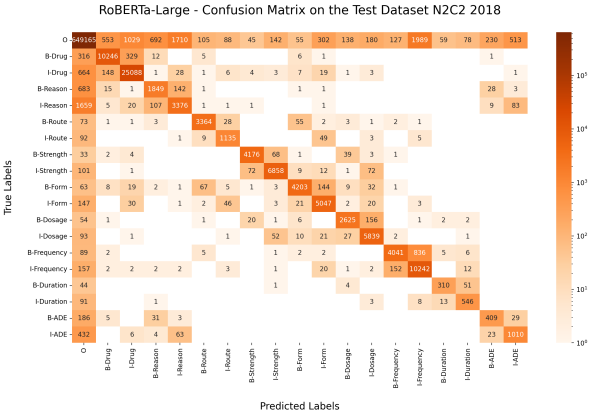

Figure 5: RoBERTa-L Eval Confusion Matrix at Sub-word Level on n2c2 2018 test data.

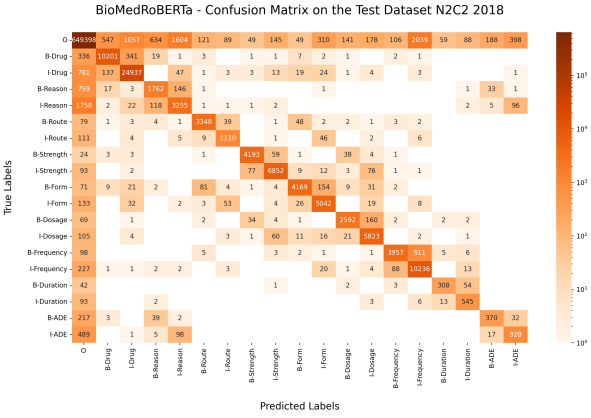

Figure 6: BioMedRoBERTa Eval Confusion Matrix at Sub-word Level on n2c2 2018 test data.

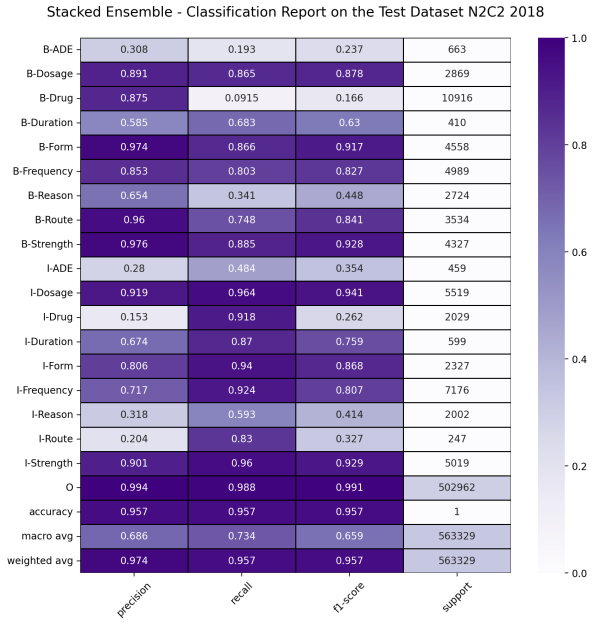

Figure 7: word-level grouping ensemble, max logit (logits, non-one-hot): stacked ensemble Eval on n2c2 2018 test data, which is much lower than the max voting.

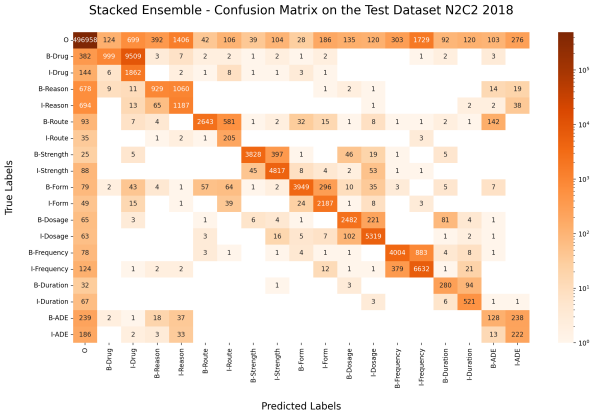

Figure 8: word-level grouping ensemble, max logit: stacked ensemble confusion matrix Eval on n2c2 2018 test data, which is much worse than the max voting.

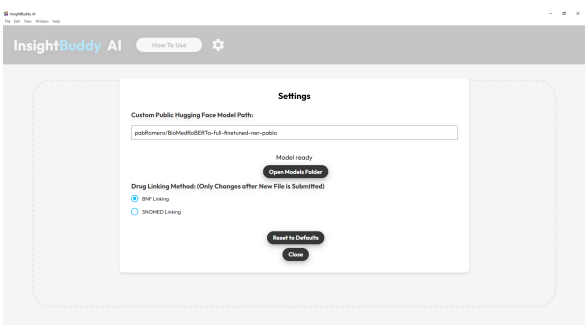

Figure 9: INSIGHTBUDDY-AI coding: Choice of BNF and SNOMED-CT Linking

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