The Challenge of Performing Ontology-driven Entity Extraction in Real-world Unstructured Textual Data from the Domain of Dementia

Sumaiya Suravee[1], Carsten Oliver Schmidt[2], Kristina Yordanova[1]
[1]Institute of Data Science, University of Greifswald, Germany
[2]Institute of Community Medicine, University Medicine of Greifswald, Germany
{sumaiya.suravee, carsten.schmidt, kristina.yordanova}@uni-greifswald.de

Abstract

Named entity recognition allows the automated extraction of structured domain-related information from unstructured textual data. Our study explores the task of ontology-driven entity recognition, a sequence labelling process for custom named entity recognition for the domain of dementia, specifically from unstructured forum texts where unprofessional caregivers of people with dementia discuss the challenges they face related to agitation. The targeted corpus is loosely structured, contains ambiguous sentences and vocabulary that does not match the agitation-related medical vocabulary. To address the above challenges, we propose a pipeline that involves the following steps: 1) development of an annotation codebook; 2) annotation of a textual corpus collected from dementia forums, consisting of 45,216 sentences (775 questions and 5571 answers); 3) data augmentation to reduce the imbalance in the corpus; 4) training of a bidirectional LSTM model and a transformer model; 5) comparison of the results with those from few shot- and zero-shot based prompt engineering techniques using a pretrained large language model (LLaMa 3). The results showed that LLaMa 3 was more robust than traditional neural networks and transformer models in detecting underrepresented entities. Furthermore, the study demonstrates that data augmentation improves the entity recognition task when fine-tuning deep learning models. The paper illustrates the challenges of ontology-driven entity recognition in real-world datasets and proposes a roadmap to addressing them that is potentially transferable to other real-world domains.

1 Introduction

People diagnosed with dementia experience cognitive impairments, such as memory, thinking, and reasoning, which interfere with their daily functioning (Nichols et al., 2022). People with dementia (PwD) may often encounter difficulty with language and communication, poor judgment, mood swings, and a decline in the ability to execute essential day-to-day tasks. Addressing dementia often involves handling the challenging behaviours and their causes. PwD can exhibit agitation, aggression, wandering, repetitive questioning, and other behaviours that are distressing for both the individual and caregivers (Vithanage et al., 2024). These behaviours are reported in clinical notes, caregiver logs, and qualitative reports, which usually use informal language and have an unstructured form. By tagging challenging behaviors, potential trigger factors, and identifying informal caregivers along with the PwD in text, a domain-specific entity recognition system can transform raw descriptions into structured data, which supports better research and understanding PwD's caregiver interaction.

This study investigates the challenges of performing ontology-driven entity recognition in unstructured texts collected from dementia-related forums where entity has been derived from the eDEM-Connect: Ontology of Dementia-related Agitation and Relationship between Informal Caregivers and Persons with Dementia (EDEM-CONNECTONTO). We conduct a comparison study on the entity recognition task using a traditionally trained Bi-directional long short term memory (Bi-LSTM) with a conditional random field (CRF) classifier, variants of masked-based language models such as BERT with the CRF classifier and then prompting a large language model (LLM) in a zero-shot setting and a few-shot setting, particularly LLaMa 3 with 70 billion parameters. This allows us to practically investigate existing limitations in LLMs, such as a hallucinations that discourage their usage in the context of specialised applications such as domain-specific entity recognition. The contributions of the paper are as follows: 1) we propose an approach for extracting ontology-

Proceedings of Recent Advances in Natural Language Processing,pages 1205–1214
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-139

driven entities from informal texts in the domain of dementia; 2) we introduce a dataset for the domain of agitation of PwD; 3) we present results from experiments that use traditional neural networks, transformer architectures and LLMs with zero shot and few shot prompting strategy. The paper is structured as follows. In Section 2, we discuss related works and present the proposed approach, along with the evaluation strategy, in Section 3. Section 4 introduces our experimental setup, and Section 5 describes the results from the evaluation and critically analyses the results from our evaluation. We illustrate the planned future works shortly in Section 6.

2 Related Work

Named entity recognition (NER) has proven highly useful for supporting biomedical applications. Studies using clinical data have applied rule-based, machine learning, and deep learning methods to NER. In particular, deep learning has shown clear improvements over earlier approaches in training NER models (Dang et al., 2018). With the availability of annotated biomedical corpora, several supervised approaches have been applied, such as support vector machine (SVM) (Yang et al., 2010) and CRF (Settles, 2004), which were used for the NER task on the GENIAFootnote2 and BioCreativeFootnote3 corpora. In (Yang et al., 2010), the authors presented a SVM-based approach, BioPPISVMExtractor, to specify protein-protein interactions in biomedical text. A Bi-LSTM-based model was used for biomedical NER in (Saad et al., 2020), where the authors evaluated and compared several models on six different datasets to identify biomedical named entities, including chemicals, diseases, drugs, species, and genes/proteins. Their proposed Bi-LSTM model, which utilises both word and character-level embeddings, vastly outperforms CRF and Bi-LSTM models that use only word-level embeddings. In (Fritsch et al., 2019), neural network language models with long short term memory cells were trained and evaluated on the DementiaBank corpus, which contains audio recordings from 194 PwD, and 98 healthy speakers serve as a control group. Besides, the conversational transcript was used in (Di Palo and Parde, 2019), where the authors employed a neural model based on a CNN-LSTM architecture that comprehends to detect alzheimer disease (AD) and related dementia using targeted and implicitly learned features from conversational transcripts. Their approach launched the new state of the art in the DementiaBank dataset, reaching an F1 score of 0.929 when classifying participants into AD and control groups.

LLM relies on transformer architecture and has been pre-trained on vast amounts of data, which has given rise to novel methods for the detection of AD. Pre-trained language models can be categorised into three large architectures: decoders (such as LLaMa, GPT, Claude), encoders (popular are masked language models, the BERT family), and encoder-decoder architectures that are typically used for language translation or speech recognition (such as Whisper) (Jurafsky and Martin, 2025). The authors of (Yuan et al., 2020) demonstrate that disfluencies and language problems in AD can be specified by fine-tuning a transformer-based pre-trained language models such as BERT and ERNIE, and 89.6% accuracy was obtained on the test set of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge. In (Li et al., 2022), the authors fine-tuned and systematically scrutinised BioBERT, BlueBERT, PubmedBERT, and SciBERT, which are the variants of the BERT model for the NER task on clinical trial eligibility criteria. The results revealed that domain-specific transformer models performed better than general transformer models using ten-fold cross-validation. Pre-trained LLM in conjunction with prompting strategies was analysed in (Zheng et al., 2024) following multiple rounds of training and a 10-fold cross-validation. The LLaMa2 model with the prompt tuning strategy showed 81.31% accuracy, which denotes 4.46% gain over the control group using the BERT model. In (Lu et al., 2024), LLMs such as LLaMa2, ChatGPT-4, Meditron and ChatGPT-3.5 were used for the token-level clinical NER task using the RareDis-v1 dataset (Martínez-deMiguel et al., 2022), the National Organisation for Rare Disorders database. Experiments involving zero-shot prompting, few-shot prompting, retrieval-augmented generation (RAG) were performed to specify five key named entity (NE) types: disease, rare disease, skin rare disease, symptoms, and signs. The study demonstrates the intrinsic challenges LLMs encounter in token-level NER, especially in rare diseases. This study focuses on developing an experimental pipeline tailored to the domain-specific ER task where the types of agitated behaviours and the causes behind

such challenging behaviour (e.g., screaming, wandering) are used as entities.

3 Proposed Approach

In this study, we first define our domain using the ontology: EDEM-CONNECTONTO in PwD, and then introduce an annotation codebook. The codebook is later used for annotating our dementia corpus, which consists of texts from dementia forums. As the distribution of entity class is imbalanced, we introduce a data augmentation strategy before processing the data. For the entity recognition task, we apply both classic neural-network architectures and masked-language models. We then benchmark their performance against an LLM using few-shot and zero-shot prompting. Figure 1 graphically presents our pipeline.

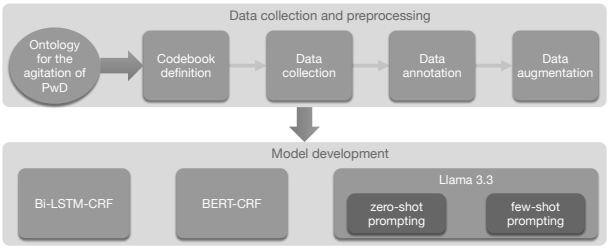

Figure 1: Proposed pipeline for the ontology-driven entity recognition in the domain of PwD.

3.1 Ontology and Codebook Development

We developed an ontology: EDEM-CONNECTONTO [1] describing agitation of the PwD and the relation of PwD with informal caregivers. The ontology incorporates domain knowledge about the different types of agitation seen in PwD along with non-pharmacological interventions, helping caregivers mitigate the bidirectional effects of agitation in PwD effectively. The ontology is developed based on expert domain knowledge, user questionnaires, and systematic literature review following the process proposed in (Yordanova et al., 2017). It has 241 concepts, 240 individuals and 10 relationship properties. Based on the ontology, we introduced an annotation scheme (codebook) (Suravee et al., 2022) that contains the eight most frequent concepts from the ontology. The codebook consists of the following entities, which are used as concepts in the EDEM-CONNECTONTO: **PwD, Family_Carer (FC), Cause (C), Agitation (A), Physical_aggressive (PA), Physical_nonaggressive (PNA), Verbal_aggressive (VA), Verbal_nonaggressive (VNA)**.

[1] https://bioportal.bioontology.org/ontologies/EDEM-CONNECTONTO

3.2 Dementia Dataset

We collected 45,216 informal, unstructured conversational sentences from an online dementia blog where users share personal experiences, challenges, and seek advice[2]. The dementia data set has 775 questions and 5571 answers. To address punctuation errors, extraneous whitespace, typos, misspellings, and grammatical inconsistencies, all sentences were preprocessed using the CleanText, TextBlob, and StanfordNLP Python packages. Each text file was then segmented so that each line contained a single sentence. Each entries were formatted into three components: "T: the title of the topic", "Q: dementia-related questions" and "A: the set of answers provided by the users corresponding to the question". The collected sentences include personal information related to the patient's history, sentences containing personal information were anonymised with the tags: `<name>`, `<location>`, `<age>`, `<time_period>`, `<distance>`, `<date>`, `<professional_practitioner>`, `<medicine>`.

3.3 Data Annotation

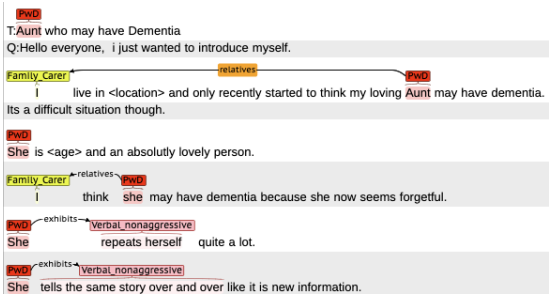

Figure 2: Example of an annotation with BRAT

We used the publicly available BRAT annotation tool (Stenetorp et al., 2012) and annotated entities based on the concepts selected in the codebook. In Figure 2 the word ***"Aunt"*** is annotated as "PwD" while the phrase ***"repeats herself"*** is annotated as "Verbal-nonaggressive". All the sentences in the file are not annotated. The raters annotated words or phrases in a sentence according to the definitions of the labels (Suravee et al., 2022). In the annotation process, three annotators were employed.

[2] https://www.healingwell.com/

The first annotator, hereafter referred to as the *expert annotator*, developed the annotation scheme and possesses advanced expertise in the domain of dementia; the *expert annotator* trained the other annotators, and their annotations were subsequently inspected, examined, and corrected by the expert. Because the dementia-related texts are often vague and context-dependent, it was challenging to distinguish whether a given passage referred to the PwD or to a family caregiver. To ensure consistency and accuracy, we conducted annotations in three iterative phases, each followed by an expert review and domain expert consultation, resulting in three consecutive annual revisions of the corpus. To train our domain-specific entity recognition model on dementia-related texts, we used the CoNLL-03 format (Sang and De Meulder, 2003), one of the most widely used NER format. The CoNLL format is a text file with one word per line with sentences separated by an empty line where the first word in a line should be the word and the last word should be the label.

3.4 Data Augmentation

In this experiment, 84 text files containing 9737 sentences were selected by considering the files with the most entity labels after the first and second phases of the annotation procedure (see (Suravee et al., 2022), (Suravee et al., 2024)). The resulting dementia corpus contains a total of 10303 annotations (see Table 1). It demonstrates a pronounced class imbalance dominated by PwD and FC labels. To mitigate this skew, we augmented the minority entity labels: A, C, VA, VNA, PA, PNA. We employed a masked language model-based approach (Kumar et al., 2020) to accomplish a more balanced distribution across all entity types using the pretrained Bidirectional Encoder Representations from Transformers (BERT) language model: "bert-base-uncased" (Devlin et al., 2019). BERT's pretrained masked language modeling (MLM) head provides a powerful mechanism for context-aware data augmentation. By randomly substituting a small subset of tokens in each sentence with the special [MASK] token and then using BERT to predict the most likely replacements, multiple semantically conceivable variants of the original text can be generated. In our case, each input document is tokenized, and a user-defined number of positions are chosen randomly for masking. The masked sequence is passed through the frozen BERT encoder and the MLM head, which leverages bidirectional self-attention to compute contextualised embeddings and output a probability distribution over the vocabulary at each masked index. Each mask was substituted with the highest-scoring prediction, yielding an augmented sentence that preserves grammaticality and domain relevance (Devlin et al., 2019). Repeating this process across multiple masks and iterations produced a diversified set of annotated examples—each inheriting the original BRAT labels – thereby enriching the training corpus with minimal manual effort and without drifting from the underlying semantics. In total, the augmentation process generated 38,280 sentences and 1302 BRAT formatted annotations. All annotations in the augmented dataset were manually reviewed, corrected, and then evaluated by the annotators. After merging the original (gold-standard) corpus with the augmented annotations, the resulting augmented dementia dataset comprises 48,017 sentences and 11,605 entity annotations (Table 1).

Table 1: No. of entity annotations from the original corpus and the augmented one for each entity type: PwD, Cause (C), Family-carer (FC), Agitation (A), Verbal-aggressive (VA), Verbal-nonaggressive (VNA), Physical-aggressive (PA), and Physical-nonaggressive (PNA)

	PwD	FC	C	A	PA	PNA	VA	VNA	N. ann.	N. sent.
Original	4869	4690	116	50	95	204	130	146	10303	9737
Augm.	4869	4690	395	152	277	364	408	450	11605	48017

3.5 Model Development

We employ a hybrid sequence labelling architecture that leverages a domain-adapted transformer-based BERT embedding with a recurrent-based Bi-LSTM with CRF classifier as the baseline model. We then compare the baseline with the performance of the variants of BERT models: BERT-CRF, PubmedBERT-CRF and Bio-Clinical BERT-CRF and the LLaMa 3 using a zero-shot and few-shot prompting strategy.

3.5.1 Bi-LSTM-CRF Model

We employ a Bi-LSTM-CRF architecture with the BERT embedding for the domain-specific entity recognition task. The architecture has three layers: (1) input embedding layer, (2) Bi-LSTM layer, which was introduced by Huang et al. (Huang et al., 2015). It can analyse each sequence forwards and backwards, grasping the context from both past and future tokens in a sentence, and (3) CRF layer, in-

troduced in (Lafferty et al., 2001), which is used to output the most likely sequence based on the output from the Bi-LSTM layer. Data processing begins by splitting each document into sentences. Then, each input sentence is tokenised into sub-tokens by a BERT tokeniser, incorporated from Hugging Face. Each sentence token is mapped to a vector representation sequence, where we use pre-trained BERT (Alsentzer et al., 2019) that provide context-sensitive embeddings for every sub-token. These embeddings are fed as contextual vectors into a Bi-LSTM layer (128 hidden units per direction). Finally, it is passed into the CRF layer, which delivers the most likely sequence of the expected labels based on the sequence of probability vectors from the previous layer. Combining BERT's deep contextual representations with Bi-LSTM's sequential modelling with the CRF's structured prediction allows for the extraction of rich lexical and syntactic patterns.

3.5.2 BERT-CRF Model

Using pre-trained language models (LM) within the biomedical and clinical domains has been very impressive in many different applications. In this study, we incorporate the following pre-trained LMs: BERT (Devlin et al., 2018), Bio-Clinical BERT (Alsentzer et al., 2019), and PubmedBERT (Gu et al., 2021) and add a CRF layer on top of the BERT frameworks so that the CRF layer can take the neighbouring tokens with their corresponding labels to predict the label of the token.

BERT (Devlin et al., 2018) follows the transformer architecture pretrained on a large corpus of raw english data in a self-supervised manner. The model was trained on four cloud TPUs in Pod configuration (16 TPU chips total) for one million steps where the sequence length was restricted to 128 tokens for 90% of the steps and 512 for the remaining 10%. **PubmedBERT** (Gu et al., 2021) and **Bio-Clinical BERT**(Alsentzer et al., 2019), are advanced LMs designed particularly for biomedical NLP, whereas PubmedBERT is trained on biomedical abstracts containing approximately 14 million abstracts in Pubmed and Bio-Clinical BERT is trained on all notes from MIMIC III (Johnson et al., 2016). Bio-ClinicalBERT is pre-trained on large biomedical corpora, MIMIC III (approximately 880M words) (Johnson et al., 2016) to initialise its word semantic features and tuned in a supervised method during training so that it comprehends the unique attributes of clinical language. PubmedBERT employs a custom vocabulary that is explicitly generated from biomedical texts, which allows the LM to comprehend biomedical terms better than the models that train on a general-purpose vocabulary derived from a broad range of texts.

3.5.3 LLaMa3 Model

We use the Meta's quantised LLaMa 3 (70B)model [3], a state-of-the-art open-source LLM incorporating the transformer architecture. It was trained with 70 billion parameters over approximately 15 trillion tokens from publicly available sources. The fine-tuning data comprises publicly available datasets and over 25 million synthetically produced examples. The quantised version was utilised to manage memory and computational limitations. All experiments were conducted with the default configuration on a system equipped with four NVIDIA GeForce RTX 4090 GPUs (NVIDIA Corporation, 2023).

4 Experimental Setup

All the entity recognition models were trained using the PyTorch framework. The longest length of a sentence can be set to 512 tokens. To grasp the entire context in the sentences, the excess part was split into another sentence once the length exceeded 512 tokens, until all the segmented sentences could satisfy the length constraint. The Bi-LSTM model was trained with a batch size of 8, where the hidden size of the LSTM layer was 128. All the experiments ran up to 100 epochs (see Table 2). Initially, we conducted experiments on the original dementia corpus, which contains 9737 sentences. The corpus was split into training, test, and development sets, which are 60%, 20% and 20% respectively. The models were trained on the training set and fine-tuned on the development set for parameter optimisation, whereas the test set was only used to evaluate the model's performance. We fine-tuned the hyperparameters on the development set using a random search and stopped the training when we achieved the best scores in the development set. Table 2 shows the parameter setting used for all the experiments. We ran the training and evaluation procedure three times on both the dementia dataset and augmented dataset and take the average scores of the three experiments.All entity recognition models were also trained and fine-tuned on the

[3]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

Table 2: Hyper-parameters for all experiments

Hyper-parameters	
LSTM hidden state size	128
Drop out	0.1
Learning rate	$0.00001 \rightarrow 0.00005$
Batch size	8
Optimizer	AdamW
Decay rate	0.01
Token length	512

augmented training set and the augmented development set, respectively. We use the same parameter settings and the same test set for all experiments.

Furthermore, we conduct experiments with LLaMa 3 on the original dementia test set. We adoped zero-shot and few-shot prompting strategy (Anthropic, 2023) and use XML tags to convey instructions. In the zero-shot prompting strategy, only the sentences in the dementia test set and specific task instructions were provided to the model, without any annotated examples. For the few-shot learning based approach, 73 annotated examples from the training set were included alongside test sentences to guide the model's predictions. Figure 3 showsthe **zero-shot** learning based prompt scheme that includes the contextual information of the entity recognition task, which describe the tagging rules and the entity labels. We provided test sets and the guidelines for output formatting to ensure a reliable evaluation. In the few-shot prompting strategy, besides providing contextual information and the tagging guideline, we fed annotated entities as training examples from the originial dementia training set and the augmented training set. We also provided the same output formatting instructions to the LLM as in the zero-shot prompting setting. If the model's response could not tag the entities according to the prompting guideline, we used a retry mechanism, permitting up to three attempts per test sentence to accomplish a valid response. As the model's predictions occasionally deviated from the expected tagging structure, especially with long sentences, the retry mechanism is needed to enhance the likelihood of getting a response that met all validation standards. Once responses were generated from the LLM, we checked alignment to verify that the words in the LLM's response matched the words in the test sentences. Finally, each predicted entity was compared and validated with the annotated entity from the dementia corpus using the F1-score.

We calculated macro F1 score for the token level

Zero-Shot Prompt Template

```
<context>
You are an expert Entity Recognizer. Your task is to identify the
entities in the given dementia-related text. Classify them
into their respective categories.
USE ONLY THE FOLLOWING LABELS:
"PwD", "Family_Carer", "Cause", "Agitation", "Verbal_aggressive",
"Verbal_nonaggressive", "Physical_aggressive",
"Physical_nonaggressive" and "O".
</context>
<description>
Each word in the text must be tagged with one of the allowed labels.
Read all the sentences carefully, understand the context. Assign the
correct labels to each word. Definitions of the labels are provided.
</description>
<task>
{Dementia texts}
</task>
<output_formatting>
<tagged_output>
Return ONLY the output in the exact format below with two columns
separated by a tab.
<pair>\\<word>WORD</word><pred\_tag>TAG</pred\_tag>\\</pair>\\
</tagged_output>\\
Ensure no extra text outside the tags.
</output_formatting>
```

Figure 3: Template of the zero-shot prompting

evaluation to measure the performance of the entity recognition models. F1-score is the harmonic mean of precision and recall (Sokolova and Lapalme, 2009).

5 Findings and Discussion

We report the token level evaluation results for the baseline Bi-LSTM with the CRF classifier and compare the baseline results with variants of transformer- based BERT model, using the original dementia and the augmented dementia corpus. We also provide token level evaluations on the test set of the augmented dementia corpus using LLaMa 3.

5.1 Token Level Evaluation on Dementia Dataset

The evaluation score of each model was calculated on the test set of the original dementia corpus (before augmentation). Table 3 shows the macro average (avg.) F1 score using the baseline Bi-LSTM-CRF model with BERT embedding and the BERT-Large uncased model. Regarding macro avg. evaluations, both models achieved poor performance, where the baseline Bi-LSTM with BERT embedding achieved a slightly better 0.23 F1-score than the BERT-Large uncased model. Notably, Bi-LSTM-CRF and BERT-Large uncased models exhibit similar F1 scores for classifying "PwD" in the dementia text, which were 0.89 and 0.87, respectively. Similar performance has been shown in identifying "FC" for both the models, with an F1 score of 0.83 and 0.84 F1 score, respectively.

Because of the underrepresented entity annotations in the original dementia corpus for the labels: A, C, PA, PNA, VA and VNA, the models were not able to reliably identify these entities in the text, resulting in poor macro F1 scores for both models.

5.2 Token Level Evaluation on Augmented Dataset

In comparison to the token-level evaluation on the original dataset and the augmented dataset, both, the Bi-LSTM-CRF and BERT-Large-uncased model's performance improved considerably from 0.23 to 0.58 and 0.22 to 0.60 respectively. Table 3 showcases the token-level evaluations on the Bi-LSTM-CRF model using BERT embedding, Bert-Large uncased with CRF, PubmedBERT with CRF and Bio-ClinicalBERT with CRF models. Compared to the outcomes on the original corpus, we observed that macro avg. F1-score for the entity labels: A, C, PA, PNA, VA and VNA were also improved for both the Bi-LSTM-CRF and BERT-Large-uncased models. Although PubmedBERT and Bio-ClinicalBERT are particularly designed for biomedical NLP, both models could not outperform the BERT-Large-uncased model, with an macro average F1 score of 0.60 (see Table 3). Using the BERT-Large-uncased model on the augmented dementia corpus yielded a substantial gain in detecting "PA" achieving an F1 score of 0.66. The Bio-ClinicalBERT model also improved performance on identifying the entity "PNA," which improved moderately from 0.38 to 0.61.

5.3 Token Level Evaluation Using LLaMa 3

The performance of the LLaMa 3 model is being evaluated using the prompt tuning method on the same dementia test set (see Table 3), where each experiment is carried out 3 times. With the LLaMa 3 model following a few-shot learning based prompting strategy, the F1 score for identifying entities improved compared to the baseline Bi-LSTM model from 0.23 to 0.30. Notably, the LLaMa 3 model showed significant improvement in identifying the "VA" entity, with F1-scores of 0.27 and 0.29, respectively, whereas the baseline Bi-LSTM-CRF model was unable to identify any "VA" entity in the test set. Additionally, it performed better in identifying "PA" entities, with F1 scores of 0.35 and 0.32, respectively (see Table 3). However, the LLM model was inferior to the baseline model in identifying "PwD" and "FC" entities, with an F1 scores of 0.76 and 0.66, respectively using the zero-shot prompting strategy. Hence, the predicted entities generated from the LLM were examined by the expert rater. Upon examining "PwD" and "FC" entities, we found that the LLM mistakenly categorized possessive pronouns as "FC". For instance, in the text "My mother has diagnosed with dementia", the LLM incorrectly tagged "My" as an "FC" despite the prompt instructing the model to label personal pronouns as "FC", this conflicted with the gold-standard annotation, leading to mismatches. Similarly, the tagged tokens "me", "his", and "her" by the LLaMa 3 did not match with the gold standard entities, causing a lower F1 score than the baseline model. The use of an augmented training set did not influence the model's performance. Interestingly, the LLM was inferior to recognise complex contextual information. With the augmented training examples, the model's performance degraded with an F1-score of 0.26 (Table 3) as the augmented training set contains more rare entity types than the training set of the original dementia corpus. The model tagged incorrect tokens with "C". For example, the LLM incorrectly classified the term "Hallucination" as "A" even though our prompts instructed it to classify "Hallucination" as "C". The model incorrectly tagged tokens as "C" with causal factors that trigger frustrations for caregivers, whereas "C" should be tagged with tokens in the text that refer to the causal factors of being agitated by the PwD. For instance, ***She doesn't remember things on a particular day and will ask repeatedly even if I told her***. Here, **doesn't remember** and **ask repeatedly** should be tagged as "C" and "VNA", respectively. Similarly, the model identified caregivers' agitated behaviour as "A" and "VA", which illustrates a gap in the LLM's ability to catch more nuanced forms of contextual texts.

In our experiments with Bi-LSTM, BERT, and LLaMa 3 for domain-specific entity recognition task, we encountered several key challenges, such as rare entity types (e.g. "Cause" or "Agitation") occur far less frequently than the "FC", "PwD", leading to models bias toward the majority class and struggle with rare entity types. Another challenge was converting the original BRAT-annotated data into CoNLL format while preserving the position of each token. Few-shot and zero-shot approaches involved extensive, time-consuming prompt tuning to achieve reliable entity tagging. The prompts had

Table 3: Macro avg. F1 scores on the original and augmented dataset using traditional neural networks, different variants of BERT model with the CRF and with LLaMa 3 model for each entity types: PwD, Cause (C), Family-carer (FC), Agitation (A), Verbal-aggressive (VA), Verbal-nonaggressive (VNA), Physical-aggressive (PA), and Physical-nonaggressive (PNA)

	Original				Augmented				
Models / NE labels	Bi-LSTM-CRF	BERT-Large	LLaMa Zero shot	LLaMa Few shot	Bi-LSTM-CRF	BERT-Large	Bio-Clinical BERT	Pubmed BERT	LLaMa Few shot
PwD	0.89	0.87	0.76	0.75	0.91	**0.96**	0.91	0.93	**0.65**
FC	0.83	0.84	0.66	0.67	0.87	**0.95**	0.87	0.92	**0.56**
C	0.0	0.0	0.0	0.0	**0.53**	0.37	0.41	0.39	0.0
A	0.0	0.0	0.06	0.06	**1.0**	0.45	**1.0**	0.66	0.07
VA	0.0	0.0	0.27	**0.29**	0.16	0.28	0.42	**0.45**	0.22
VNA	0.07	0.0	0.13	**0.17**	**0.71**	0.62	0.42	0.36	0.16
PA	0.0	0.0	**0.35**	0.32	0.10	**0.66**	0.06	0.0	0.32
PNA	0.05	0.0	0.12	**0.20**	0.38	0.48	**0.61**	0.26	0.10
Macro avg. F1	0.23	0.22	0.29	**0.30**	0.58	**0.60**	0.58	0.49	0.26

to be iteratively refined to ensure accurate token-level predictions in CoNLL format. Inference with pre-trained large language models (LLMs), particularly for few-shot prompting, requires substantial GPU resources. At the same time, contextual and loosely structured, conversational, real-world domain-specific texts require longer processing times due to their complex parsing and tokenisation structure. Additionally, following a few-shot prompting strategy, we were unable to provide more than 75 training examples due to the limited prompt window.

6 Conclusion and Future works

This study approaches to employing ontology-driven entity recognition using traditional neural networks, transformer architectures, and LLM with zero-shot and few-shot prompting strategies to real-world texts collected from an online dementia forum. As the corpus contains informal conversational texts with contextual and syntactical ambiguity that does not match the agitation-related medical vocabulary, we introduced an experimental pipeline that involves annotation codebook development, data augmentation to reduce the imbalance, ontology-driven entity recognition model training, and we compared the results with those from few shot- and zero-shot-based prompt engineering techniques using LLaMa 3. Our findings indicate that employing a transformer-based architecture, such as the BERT-Large model, which was trained on a large corpus of english raw text, can improve performance for the customized domain-specific entity recognition task. The LLaMa 3 model was able to detect entities from the underrepresented classes with an F1 score of 0.30, possibly because the input texts were descriptive, and the autoregressive character of the LLM allowed for a more vast interpretation of the text. In the future, we aim to fine-tune the LLaMa 3 model using optimized hyperparameters. We plan to investigate the integration of the developed ontology with retrieval augmented generation for improving the model performance. Furthermore, we also focus on better understanding the linguistic and syntactical characteristics of input texts that influence entity recognition performance. It would also be valuable to identify textual features that indicate the suitability of a corpus for the entity recognition task-specifically, its fitness for purpose in a given application context.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Anthropic. 2023. Prompt engineering overview. Accessed: 2024-10-29.

Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. 2018. D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. *arXiv preprint arXiv:1906.05483*.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Jianfu Li, Qiang Wei, Omid Ghiasvand, Miao Chen, Victor Lobanov, Chunhua Weng, and Hua Xu. 2022. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC medical informatics and decision making*, 22(Suppl 3):235.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2024. Large language models struggle in token-level clinical named entity recognition. *arXiv preprint arXiv:2407.00731*.

Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. 2022. The raredis corpus: a corpus annotated with rare diseases, their signs and symptoms. *Journal of biomedical informatics*, 125:103961.

Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. 2022. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2):e105–e125.

NVIDIA Corporation. 2023. Nvidia ada gpu architecture. Whitepaper, NVIDIA Corporation.

Farag Saad, Hidir Aras, and René Hackl-Sommer. 2020. Improving named entity recognition for biomedical and patent data using bi-lstm deep neural network models. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*, pages 25–36. Springer.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Sumaiya Suravee, Teodor Stoev, Sara Konow, and Kristina Yordanova. 2024. Assessing large language models for annotating data in dementia-related texts: A comparative study with human annotators. In *INFORMATIK 2024*, pages 487–498. Gesellschaft für Informatik eV.

Sumaiya Suravee, Teodor Stoev, David Schindler, Iris Hochgraeber, Christiane Pinkert, Bernhard Holle, Margareta Halek, Frank Krüger, and Kristina Yordanova. 2022. Annotation scheme for named entity recognition and relation extraction tasks in the domain of people with dementia. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 236–241. IEEE.

Dinithi Vithanage, Yunshu Zhu, Zhenyu Zhang, Chao Deng, Mengyang Yin, and Ping Yu. 2024. Extracting symptoms of agitation in dementia from free-text

nursing notes using advanced natural language processing. In *MEDINFO 2023—The Future Is Accessible*, pages 700–704. IOS Press.

Zhihao Yang, Hongfei Lin, and Yanpeng Li. 2010. Bioppisvmextractor: A protein–protein interaction extractor for biomedical literature using svm and rich feature sets. *Journal of biomedical informatics*, 43(1):88–96.

Kristina Yordanova, Philipp Koldrack, Christina Heine, Ron Henkel, Mike Martin, Stefan Teipel, and Thomas Kirste. 2017. Situation model for situation-aware assistance of dementia patients in outdoor mobility. *Journal of Alzheimer's Disease*, 60:1461–1478.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech*, volume 2020, pages 2162–6.

Tian Zheng, Xurong Xie, Xiaolan Peng, Hui Chen, and Feng Tian. 2024. Alzheimer's disease detection based on large language model prompt engineering. In *International Conference on Social Robotics*, pages 207–216. Springer.