Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Biomedical Entity Characterization to Optimize Named Entity Recognition

Gregory Desouza
Desouza Analytics
Andrea Corradini
andrea.corradini@mci.edu

MCI

Research Article

Keywords: Biomedical Entity Analysis, Linguistic Feature Analysis, Text Min-ing, NLP, NER

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9602153/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Biomedical Entity Characterization to Optimize Named Entity Recognition

Gregory Desouza[1] and Andrea Corradini[2]

[1] Desouza Analytics, TW5 9LR, London, England
g_m_desouza@hotmail.com
[2] MCI, 6020 Innsbruck, Austria
andrea.corradini@mci.edu

Abstract

Named entity recognition (NER) techniques still struggle to identify entities of greater syntactic and semantic complexity, and although most studies recognize that complexity impacts machine learning (ML) model performance, they do so only after model evaluation. This study takes a unique approach by evaluating whether initial entity complexity characterization using quantitative methods aligns with downstream model performance.

The results showed lower F1-scores for entities characterized as higher-complexity than for those characterized as lower-complexity, indicating that quantitative characterization aligns in model performance. By establishing a measurable link between entity characteristics and performance variability, this study offers a novel foundation for complexity-aware optimization strategies. Future research could investigate whether the methodology can categorize entities by complexity across broader biomedical corpora and evaluate its applicability to emerging learning methods, including prompt-based LLMs.

Keywords: Biomedical Entity Analysis, Linguistic Feature Analysis, Text Min-ing, NLP, NER

1. Introduction

Machine Learning (ML) methods can provide a high-capacity, scalable solution for processing unstructured biomedical text. For instance, Bidirectional Encoder Representation From Transformers (BERT), a widely used natural language understanding model, has been shown to recognize diseases, medications, and company names, attaining an overall F1-score of 70.62% [1]. However, it struggled to classify the lines of therapy (LoT) and the stage of cancer (SoC), achieving accuracies of 61% and 56%, respectively [2]. This performance difference could be caused by ambiguity, abbreviations, word quantity and variation, and sentence structures of biomedical entities [3], resulting in BioNE Complexity.

Previous studies have examined how entity complexity affects ML model performance and how such challenges may be mitigated. Standard practice includes improving data quality and quantity, selecting an appropriate model, and implementing HPO [3]. Other advanced techniques include debiasing to improve generalization and joint labelling to handle nested entities [4,5]. However, despite numerous methods to address complexity and improve model performance, most studies did not assess entity complexity before implementing models, making it difficult to identify the entity features that influence model performance.

Given this gap, this study investigates whether biomedical entities can be organized by lexical complexity prior to model evaluation, whether such characterization is reflected in NER performance, and whether strategies can be used to improve the identification of the lagging entity.

1.1. Research goals

The central objective is to analyze syntactic and semantic characteristics to assess whether biomedical entities can be systematically organized by complexity, then evaluate an ML model to determine whether complexity characterization is reflected in downstream NER performance. To do so, this study applies quantitative descriptive statistical analyses to characterize entity-related features in biomedical text, including word and sentence length, phrase structure, lexical diversity, part-of-speech distribution, entity repetition, common word usage, and entity dependency patterns.

By investigating linguistic complexity before model evaluation, the study aims to be independent of any specific ML architecture and tokenizer. This data-first approach enables complexity-informed model evaluation and optimization across different NER systems, with the selected ML model serving as an empirical case study.

To support this objective, the study further examines whether commonly cited descriptions of entity complexity in the literature are empirically reflected in observable linguistic features and whether differences in recognition performance persist under varying model configurations and optimization conditions.

Research questions

The supporting research questions of this work are as follows:

i. Is there a difference between the ML's ability to identify entities, and is this due to lexical features?
ii. Do the methods characterizing entities by complexity reflect in the performance of the ML algorithm?
iii. How do optimization strategies impact the ML algorithm's performance in detecting entities as characterized by quantitative methods?
iv. Is an entity-complexity perspective valuable for interpreting NER performance and guiding model optimization?

2. Literature review and research gap

2.1. NER performance

Machine Learning models have shown respectable performance in identifying biomedical entities across different media. BERT had commendable performance when identifying entities from news articles for NER achieving an accuracy of 90% when identifying company names [6]. Similarly, BERT model variants were fine-tuned for a biomedical NER task using ophthalmology news articles to identify diseases, medications, and company names where the best performing model attained commendable performance [1].

Although ML models generally show good performance in identifying a range of biomedical entities. A study using the Oncology Drug Approval NEWS (ODAN) dataset, highlighted the differences in ML algorithms to identify biomedical concepts. For instance, they reported strong performance identifying clinical trial names (F1-score of 87%) in an NER task, but comparatively lower performance (61% accuracy) in the line of therapy (LoT) classification task [2]. While these tasks differ in formulation, both involve the classification of biomedical entities and rely on the model's ability to capture contextual and structural patterns in text, suggesting that intrinsic entity characteristics may influence model performance.

2.2. Machine Learning models

Prior research has highlighted the importance of model capacity and domain alignment in biomedical NER. For example, using ophthalmology corpora, it was demonstrated that RoBERTa, a transformer-based model with greater parameter capacity, better captured nuanced biomedical patterns [1]. Similarly, using electronic health records (EHRs), a study demonstrated that biomedical-specific language models trained with bespoke vocabularies outperformed generic counterparts under such constraints [7].

Recent advances in biomedical NLP have focused on models with increasingly sophisticated architectures, resulting in state-of-the-art performance. For instance, a hybrid architecture with BERT embeddings has been reported to outperform comparable models without contextual embeddings in a data-constrained biomedical setting [8,9].

2.3. Data quantity and quality

Data quantity alone does not fully explain performance variation. Research examining annotation practices has shown that annotation quality can exert a greater influence than corpus size, particularly for complex entity types. In one study, model performance declined as corpus size increased due to inconsistent annotation introduced after expert-annotated subsets were exhausted, exacerbating boundary identification errors [10]. These findings suggest that entities with greater syntactic diversity and lower lexical repetition may require more careful annotation to support reliable model learning.

Unstructured biomedical and health-related news articles often contain extraneous information, which can introduce noise into the learning process. Prior work has demonstrated that selecting relevant text segments from news articles can improve entity recognition by promoting clearer syntactic and semantic patterns [6]. Such observations further highlight the importance of understanding how entity characteristics interact with both data curation and model performance.

2.4. Advanced techniques to improve performance

Beyond model selection and data-centric strategies, prior work has explored advanced techniques to mitigate entity-level challenges in biomedical NER, including debiasing methods, customized vocabularies, and the integration of external knowledge sources. Debiasing aims to reduce a model's reliance on memorization by promoting synonym and concept performance, thereby encouraging more contextual learning [5]. Similarly, the use of bespoke vocabulary has been shown to reduce misinterpretation by improving entity representation in model embeddings [11]. Knowledge graphs (KGs) have also been investigated and shown to aid ML models by providing structural relation information for discerning valuable patterns in entities [12].

2.5. Summary of limitations and research gap in existing literature

Lexical, syntactic and semantic complexities consistently affect models across the biomedical domain, contributing to performance variation across entity types. These complexities include

the range of vocabulary, abbreviations, token length, and sentence structure [3]. For instance, entities such as medications are often explicitly stated and composed of a few tokens and a limited set of parts-of-speech (PoS) patterns. Alternatively, entities can be concepts composed of many tokens and exhibit greater contextual variability, such as LoT.

Although existing studies have explored strategies to mitigate entity-level challenges through model selection, data refinement, and optimization techniques, none have explicitly examined entity complexity prior to model evaluation. This gap limits understanding of how intrinsic linguistic properties of entities influence downstream NER performance and motivates the need for systematic, pre-model complexity assessment.

3. Conceptual framework

3.1. Entity types and complexity definitions

Prior work observed differences in ML models' ability to identify entities and investigated techniques to improve model performance by aiding syntactic and semantic learning, without accounting for each entity's complexity before model evaluation. While the literature lacks explicit definitions of simple and complex biomedical entities, it does describe a set of linguistic attributes that consistently contribute to complexity.

Syntactic complexity, which affects semantic interpretation by models, arises from ambiguity, abbreviations, word quantity and variation, and sentence structure in biomedical text [3]. Accordingly, these attributes provide a theoretical basis for examining relative differences in entity complexity (Table 1), forming the conceptual foundation for the subsequent complexity analysis.

Table 1. Characteristics associated with relative entity complexity (synthesized from Pakhale, 2023)

Lower Complexity	Explicitly stated, low ambiguity, minimal context inference and low variability in spelling, abbreviations, length and vocabulary. They appear consistently in simple sentence structures and similar locations, have distinct patterns and are easy to label.
Higher Complexity	Implicitly stated, high ambiguity, significant context inference, and high variability in spelling, abbreviations, length and vocabulary. They have an inconsistent presentation in complex sentence structures and varied locations, have fuzzy patterns, and require an advanced understanding of labels.

3.2. Complexity dimensions

Supporting this characterization of complexity, prior research has shown that variability in token spans and PoS composition influences boundary ambiguity and performance variability in biomedical NER [3,13,14,15]. Additionally, entity frequency and surrounding words have also been shown to affect model generalization, especially for context-dependent ML models [16].

Accordingly, quantitative analyses can be used to examine whether entities exhibit differing complexity-related characteristics. For instance, entities with higher-complexity characteristics may exhibit relatively greater skewed distributions in their word and vocabulary counts and entity length, compared to lower-complexity entities. If these characteristics do impact ML models, as supported by the literature, this should be reflected in their performance.

4. Data and methods

4.1. Data

An ODAN corpus was created and used as the experimental domain. ODAN articles announce the approval of oncology drugs by regulatory bodies and typically include the regulatory agency that approved the drug, the sponsoring pharmaceutical company, the approved drug, clinical trials, and patient population data. The latter can include LoT, SoC, disease, and genetic information [17]. Hence, ODAN contains extensive biomedical information and serves as a source of authoritative information on the current drug approval landscape, impacting patients and drug development.

Consequently, some ODAN entities can be explicit, while others are more implicit [2]. For instance, company names are composed of a few explicitly stated tokens and have relatively straightforward patterns. However, entities like LoT, which may not be explicitly stated, depend on other entities, such as a sequence of treatments. Hence, complex entities exhibit syntactic and semantic interdependencies, where their meaning depends on the relationships with other biomedical attributes, generating nuanced patterns.

Convenience sampling was employed, ensuring a dataset comparable to that of Gendrin et al [2], who achieved commendable results in their NER task of recognizing clinical trial names. If limited data affects ML models, simple entities should be less affected than complex ones, suggesting that models are sensitive to entity complexity.

Not all data was needed for the investigation, so only segments of the text associated with a drug approval were extracted. The extraction process was done manually by reading and copying the appropriate approval paragraphs, usually self-contained, into a spreadsheet, where each row was a distinct approval. Therefore, key contextual patterns associated with the chosen entities were preserved [1,6].

To prevent the risk of data duplication and leakage arising from overlapping announcements across regulatory sources, records were filtered by date, regulatory body, drug name, and company name, and any overlaps were manually verified.

Preprocessing revealed that the dataset exhibited right-skewed distributions in token, word, and sentence counts, with high stop word frequency and relatively balanced entity classes. Symbols that were inconsistently used, unrelated to entities, or lacking grammatical significance were removed, while stop words were preserved, as these words may provide useful information about syntactic features aiding entity identification. The data were split into training (64%), validation (16%), and test (20%) sets, with Table 2 showing the prevalence of company names and LoT.

Table 2. Entity representation in the training data

Perspective	Company names	LoT
Record wise	53%	47%

4.2. Entity complexity classification framework

Building on the conceptual definition of complexity outlined in section four, this study operationalizes these attributes into measurable linguistic dimensions. To categorize biomedical named entities by syntactic and semantic complexity prior to model evaluation, this study introduces Biomedical Named Entity Complexity Analysis (BioNE CA).

BioNE CA comprises a set of quantitative descriptive analyses designed to analyze syntactic and semantic properties that contribute to entity complexity, as identified in the literature using the following measures:

- Token and word count distributions, were used to assess the span of entities and distributional skewness.
- Corrected type-token ratio (CTTR), a length-normalized indicator of vocabulary variation to measure lexical diversity

- PoS distributions to capture structural variability across entity spans.
- Entity repetition frequency, to evaluate pattern consistency across records.
- Contextual word overlap to examine the consistency of surrounding lexical environments.

Company names and LoT were preselected from the literature to determine whether they exhibited distinct linguistic profiles. BioNE CA was then applied to a subset of unrefined training data selected via proportional sampling to assess whether there were differences in syntactic and semantic profiles. Based on their quantitative properties entities were treated as representing relatively lower or higher complexity, forming the basis for subsequent model evaluation.

4.3. Named entity recognition framework

NER was selected as the experimental task due to its direct token-entity alignment and reduced reliance on external biomedical knowledge during corpus curation. This allowed the investigation to focus explicitly on the relationship between entity complexity and model performance.

The datasets were tagged using the following schema: "B" (beginning) token is at the start of an entity; "I" (inside) token is within an entity, and "O" (outside) token does not belong to an entity, collectively called BIO. Each label will be suffixed with either “Co” for the company name or “LoT” to signify the entity to which they belong. This tagging strategy helps BERT understand the structure of the entities [10]. Consistent BIO sequences reinforce entity boundaries, facilitating model generalization.

4.4. Configuration strategies (ConS)

To assess whether optimization strategies mitigate performance differences between simple and complex entities, a series of Configuration Strategies (ConS) were applied. These strategies reflect common practices in biomedical NLP and include:

- Model type: domain vs non-domain pre-training
- Data quality: unrefined vs refined annotation protocols
- Data quantity: baseline vs expanded datasets
- Hyperparameter optimization (HPO): applied vs not applied

4.4.1. Chosen ML model

BERT was selected as the experimental test bed to evaluate the study objectives, as it has demonstrated strong performance in biomedical named entity recognition (BioNE) tasks, achieving high F1-scores for identifying a range of biomedical entities [1,6]. It employs a bidirectional transformer architecture to create contextual representations of tokens by analyzing both their left and right contexts, identifying local and long-range patterns associated with entities [18]. This is particularly important for complex biomedical entities, as their meaning often depends on the surrounding text.

BERT models were selected for their commendable performance in identifying a range of biomedical entities, as they can model contextual patterns in unstructured text [19], and provide a platform for systematic comparison without introducing architectural differences. Specifically, the study evaluates DistilBERT and BioDistilBERT, which share the same underlying transformer architecture but differ in their pre-training corpora, with the latter being trained on biomedical text [20,21]. This design enables investigation of whether differences exist in the identification of relatively lower and higher complexity entities and whether these differences persist across varying configuration strategies.

4.4.2. Data quality and quantity

To assess the impact of data quality on the performance of the models, three corpora were created: an unrefined corpus without a formal annotation protocol, a refined corpus annotated using a predefined tagging protocol, and a refined corpus augmented with additional data. The tagging protocol was applied to the refined dataset only, with the intention of standardizing tagging decisions, entity boundaries, repetition, and contextual relevance, reducing annotation noise and inconsistency in line with established annotation guideline principles [22]. Additional records were incorporated into the refined corpus to evaluate the effect of increased dataset size on model performance.

4.4.3. Hyperparameter optimization

Hyperparameter optimization was conducted using Bayesian optimization with cross-validation, enabling efficient exploration of the hyperparameter space. This approach prioritizes generalizability and stability over exhaustive search, which is particularly important given the limited dataset size [23].

4.5. Evaluation strategy

Model performance was evaluated using the F1-score, which balances precision and recall and is widely adopted in NER tasks [24]. To enable entity-specific analysis, F1-scores were calculated separately for company names and LoT by excluding non-relevant entity labels during evaluation.

To test whether model performance differs significantly between simple and complex entities, the Mann-Whitney U (MWU) test was applied. This non-parametric test was selected due to the unknown distribution of F1-scores and the unpaired nature of the samples.

For each configuration strategy (ConS), models were trained and evaluated over five independent runs to account for stochastic variation in training. Performance metrics were averaged across runs to obtain a single score per entity type for each configuration. These aggregated scores were then compared between simple and complex entities using the Mann-Whitney U test.

4.6. Statistical analysis

Statistical testing was conducted using Python (Pandas, NumPy and Mathplotlib) libraries. Python was used for BioNE CA, a collection of quantitative descriptive metrics designed to characterize entities according to syntactic and semantic complexity. To determine whether there was a significant difference in model performance across entities, F1-scores were computed and compared using the MWU test. This non-parametric test was selected due to the non-normal distribution of the data and the unpaired nature of the samples. Statistical significance was assessed at $\alpha = 0.05$.

5. Results

5.1. Lexical length and distribution

BioNE CA results revealed a linguistic difference between the company name and LoT across multiple syntactic and semantic features. Figure 1 shows that LoT exhibited longer token spans, as evidence by a greater spread and a right-skewed distribution compared to company names, which were tightly distributed, reflecting shorter and more consistent lexical forms. This pattern indicates greater boundary ambiguity for LoT entities.

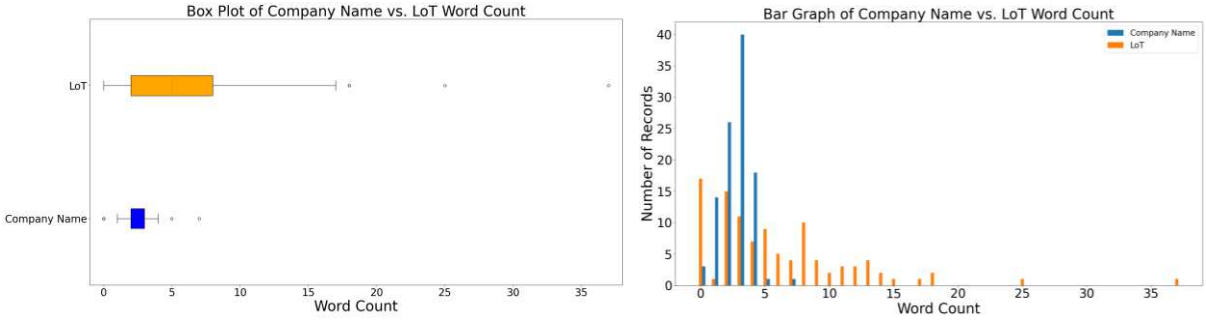


Figure 1: word count comparison between company names and LoT.

5.2. Lexical and contextual diversity distribution

As seen in Figures 2 and 3, LoT has a greater variety of common words, as reflected in a wider distribution. When entities are compared within their surrounding context, company names tend to exhibit a greater variety of common words and frequencies, with broader distributions compared to LoT. All distributions are right-tailed, and repetition-frequency analysis shows that zero repetition is the dominant category for both entity types, occurring in approximately 81% of company name records and 94% of LoT records.

These results suggest that when isolated company names exhibit more consistent lexical patterns and when considered in context, the entity is surrounded by a richer and more repetitive set of common words, providing robust entity-level and contextual cues. In contrast, LoT entities have a greater variety of words and are surrounded by fewer stable contextual cues.

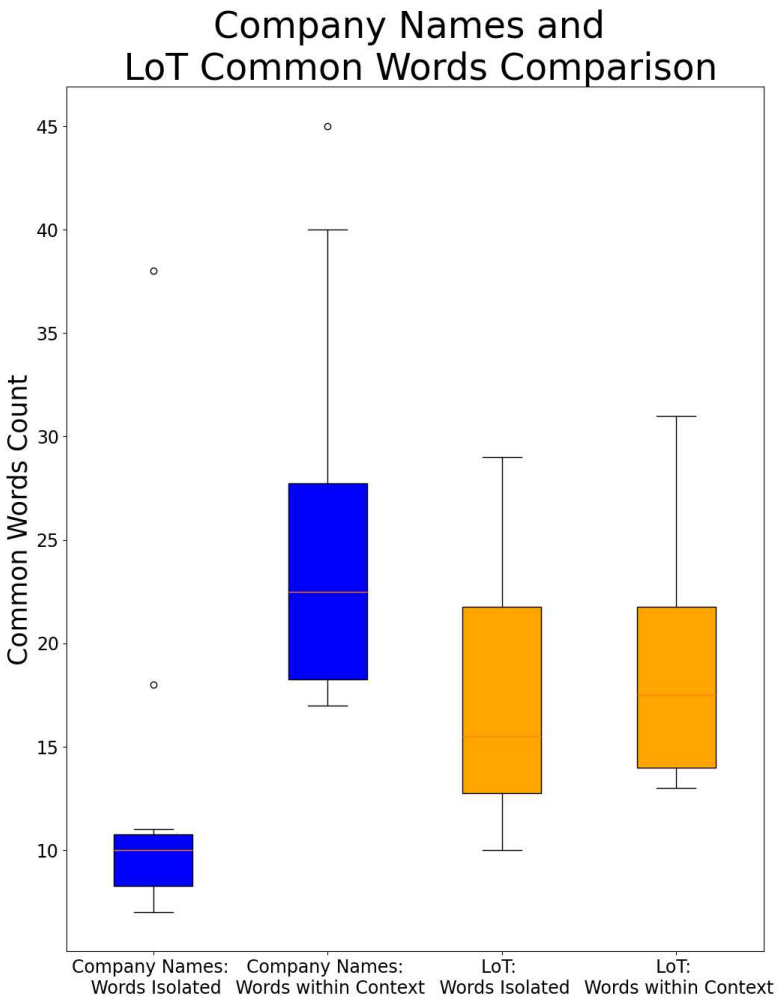


Figure 2. Common word comparison between company names and LoT

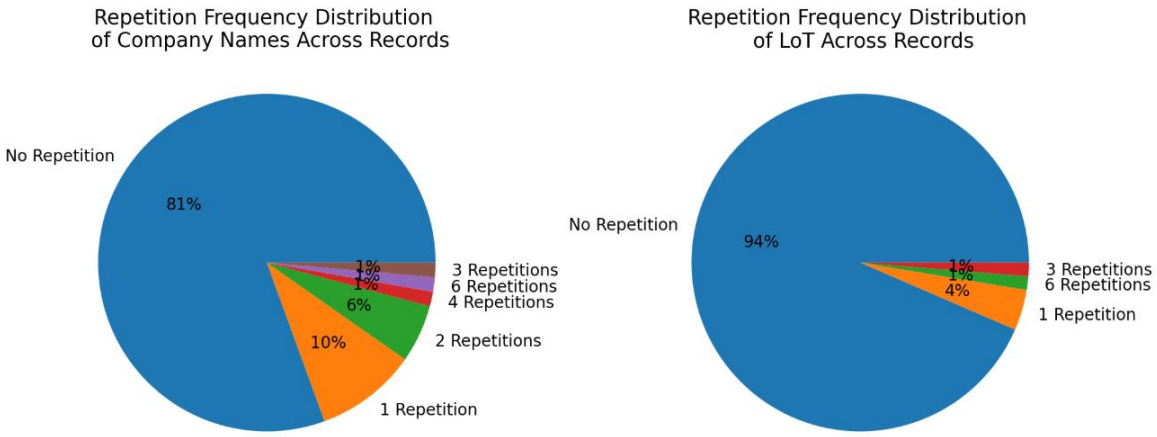


Figure 3. NE repetition counts

Analyzing the common-phrase count (Figure 4) showed that bigrams have notable representation, while trigrams and higher-order n-grams exhibit minimal frequency across entity types. When analyzed in isolation, company name bigrams are less frequent than LoT bigrams. In a contextual setting, company name bigrams showed higher frequency and a broader distribution, indicating greater repetition and greater variability in common phrases. In contrast, LoT bigrams within context are less frequent and maintain a tighter distribution, displaying tighter repetition pattern.

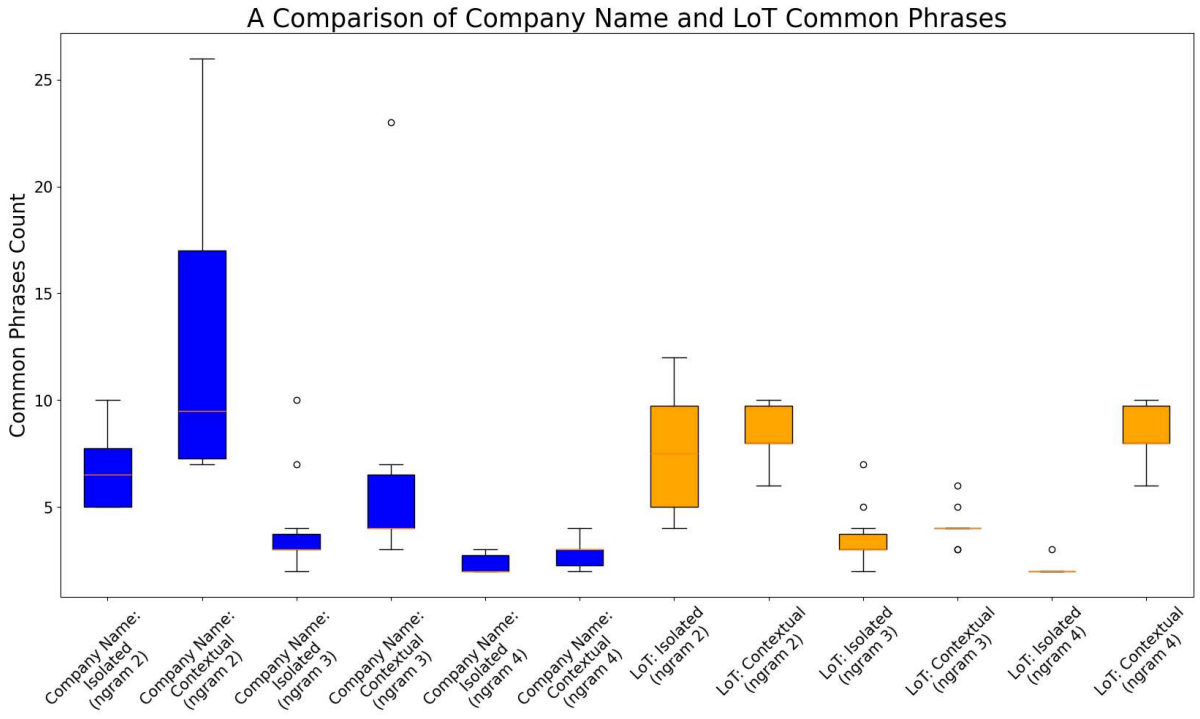


Figure 4. Common phrases, NE isolated and within context settings

5.3. PoS and corrected type-token ratio

PoS and CTTR (Figures 5 and 6, and Table 3), corroborated the lexical patterns observed in the earlier observations. Company names exhibited simpler grammatical structure, reflected in lower median total PoS counts and a narrower interquartile range (IQR). In contrast, LoT entities exhibit a greater syntactic variability, evidenced by higher median PoS counts and increased dispersion.

CTTR showed that the distributions of complexity overlap between the entities. However, LoT has a greater variability, particularly in the upper tail and across the IQR. Additionally, LoT exhibited a larger lexical diversity score across all measures of central tendency.

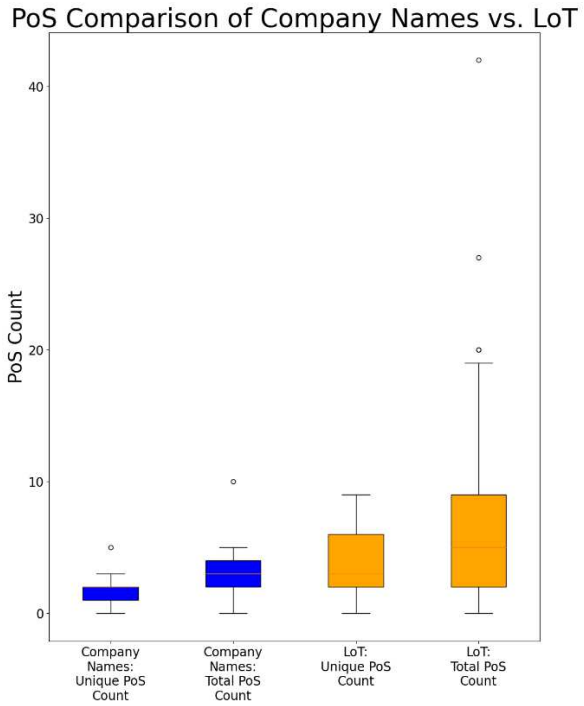


Figure 5. PoS analysis

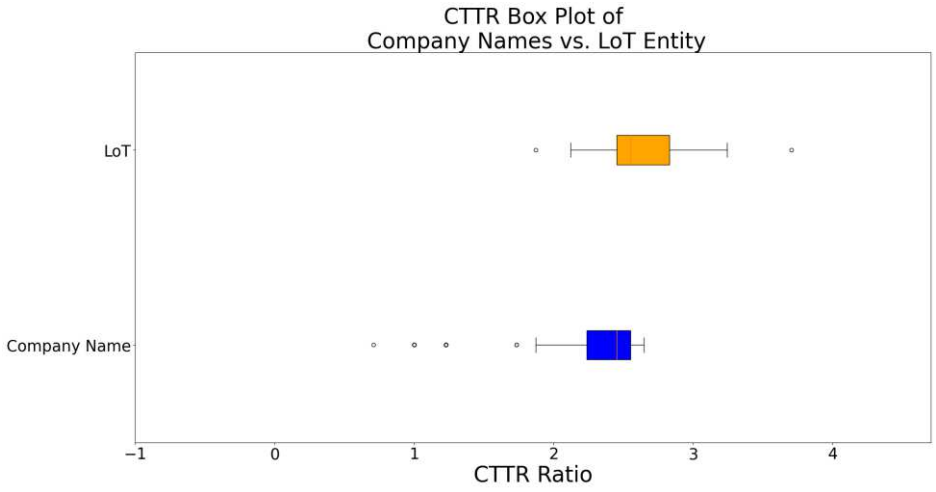


Figure 6. CTTR distribution.

Table 3. CTTR central tendency using record-wise results.

Measures	Company names	LoT
Median	2.35	2.46
Mean	2.24	2.52
Mode	2.35	2.45

5.4. Entity complexity characterization

Across all analytical measures, including word counts, common-word repetition, contextual analysis, PoS, and CTTR, BioNE CA demonstrated that LoT entities comprise longer token spans, exhibit lower repetition across records, and have more diverse PoS, demonstrating higher lexical diversity. Accordingly, company names were designated as lower complexity, and LoT were designated as higher complexity, forming the basis for subsequent performance evaluation.

6. Discussion

The central aim of this study is to evaluate whether entities can be characterized based on lexical features and if the model's performance aligns with the characterization. Supporting this aim is the evaluation of ConS and its impact on BERT's ability to identify entities based on complexity.

6.1. Hypothesis testing

The null hypothesis was rejected and the alternative hypothesis accepted, as the p-value was below the α value across all ConS (Table 4). Hence, there is a significant difference in BERT's ability to detect entities associated with differing lexical complexity, providing empirical support for BioNE CA's methods.

Specifically, across all runs, the model's F1-score for LoT was lower than for company names (Table 5). This pattern is consistent with the structural characteristics of LoT, which often contain multiple words, including stop words, that are split into segments throughout the text and exhibit inconsistent context.

Table 4. Hypothesis results

Model	Dataset and HPO	p-value	α	Null Hypothesis (Accept/Reject)
BioDistilBERT	Unrefined	0.007937	0.05	Reject
	Unrefined + HPO	0.007937		Reject
	Refined	0.007937		Reject
	Refined+ HPO	0.007937		Reject
	Refined + more data	0.007937		Reject
	Refined + more data + HPO	0.011925		Reject
DistilBERT	Unrefined	0.007937		Reject
	Unrefined + HPO	0.007937		Reject
	Refined	0.007937		Reject
	Refined+ HPO	0.007937		Reject
	Refined + more data	0.007937		Reject
	Refined + more data + HPO	0.007937		Reject

Table 5. Result summary. The F1-score is consistently less than the company names.

Model	BioNE	No HPO			HPO		
		Dataset groups			Dataset groups		
		Unrefined	Refined	Refined + more data	Unrefined	Refined	Refined + more data
DistilBERT	**Company Name**	0.935	0.912	0.916	0.949	0.939	0.931
	LoT	0.590	0.522	0.539	0.630	0.582	0.604
BioDistilBERT	**Company Name**	0.934	0.927	0.926	0.950	0.939	0.937
	LoT	0.565	0.502	0.520	0.647	0.581	0.689
* Averaged scores over five runs.							

The qualitative analysis further supports the influence of lexical characteristics on BERT's ability to identify certain entities (Figures 7 and 8). Company names have a higher frequency of repeating words that are strongly associated with their entities. These included "inc", "inc.", "pharmaceutical" and the company brand name. Alternatively, LoT is often implicitly expressed and embedded in a mixture of words that can describe other ODAN entities, such as disease references or stage of cancer (SoC).

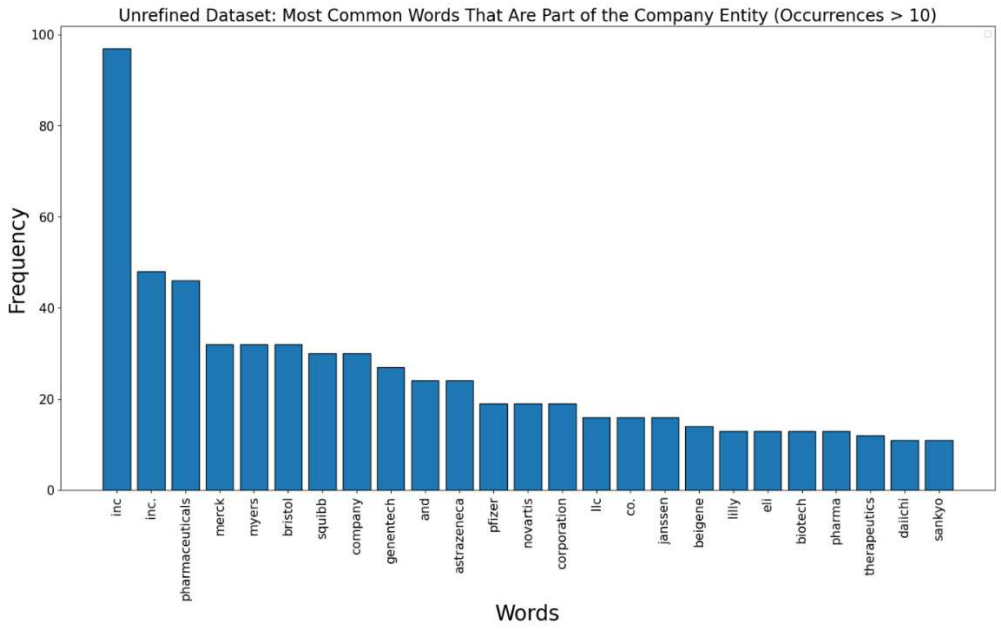


Figure 7. Frequency of common words in Company Names

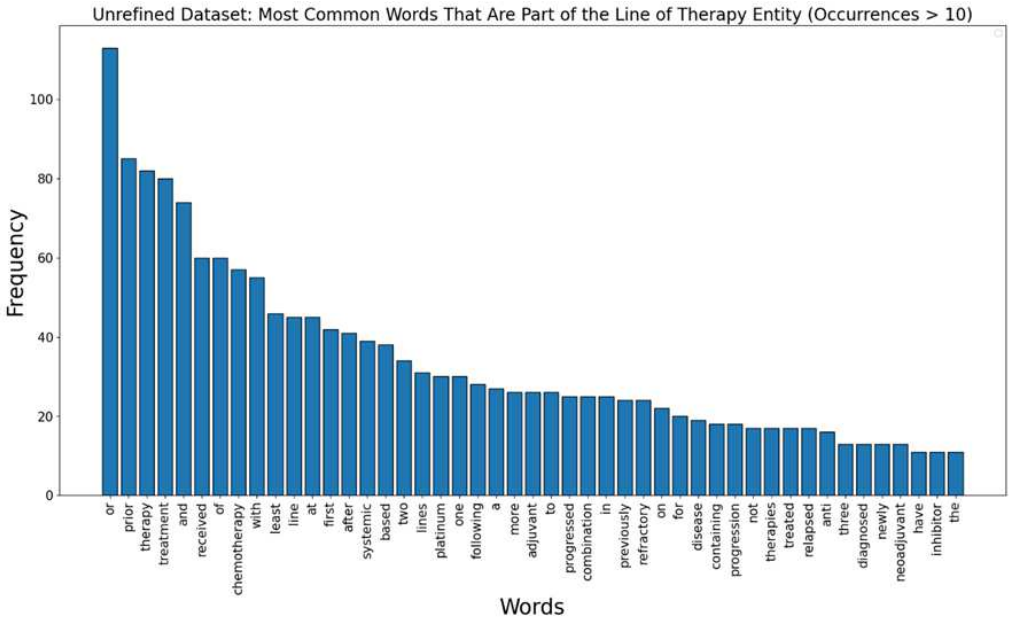


Figure 8. Frequency of common words in LoT

Additionally, there is a difference in the prevalence of stop words and in the consistency of context presentation (Tables 6 and 7, and Figures 9 and 10). LoT has a significantly higher frequency of stop words, probably contributing to noise. In contrast, company names were favoured by far fewer stopwords and a high frequency of preceding common phrases. These characteristics further add to the syntactic and semantic complexity differences, aligning with the BioNE CA findings.

Table 6. The top common stop words in Company Names and LoT.

Company	Word count	LoT	Word count
and	24	or	113
a	2	and	74
â€™s	2	with	55
's	4	more	26
us	8	three	13
one	1	an	8
in	1	if	1

Table 7. Top n-grams in ODAN

text ngram2	count	text ngram3	count	text ngram4	count
Food Drug	171	Food Drug Administration	171	Food Drug Administration approve	99
Drug Administration	171	Drug Administration approve	99	small cell lung cancer	61
FDA approve	158	grant accelerated approval	85	non small cell lung	57

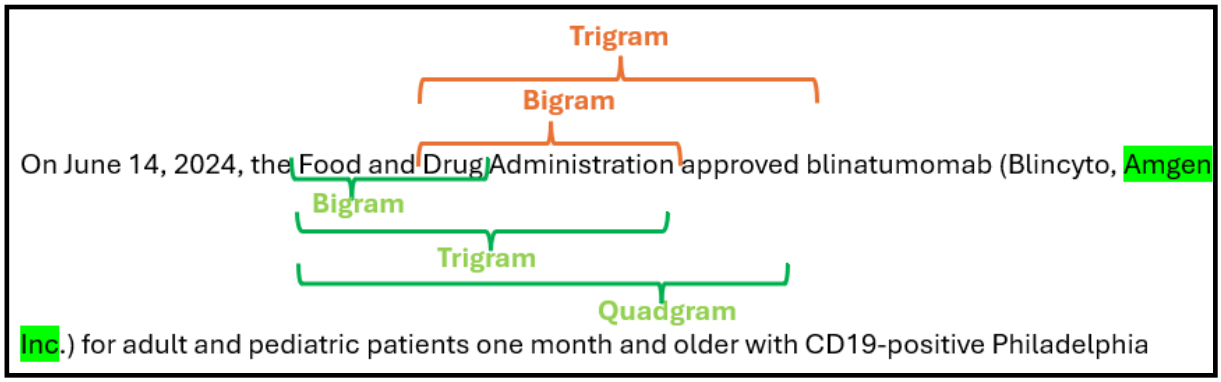


Figure 9. The locations of the common (green) and second (orange) most common n-grams in an ODAN relative to a company name (green highlighted).

who have received at least one prior anti-CD20-based regimen; Adult patients with relapsed or refractory follicular lymphoma (FL) who have received at least three prior lines of systemic therapy.

Figure 10. LoT can be in segments.

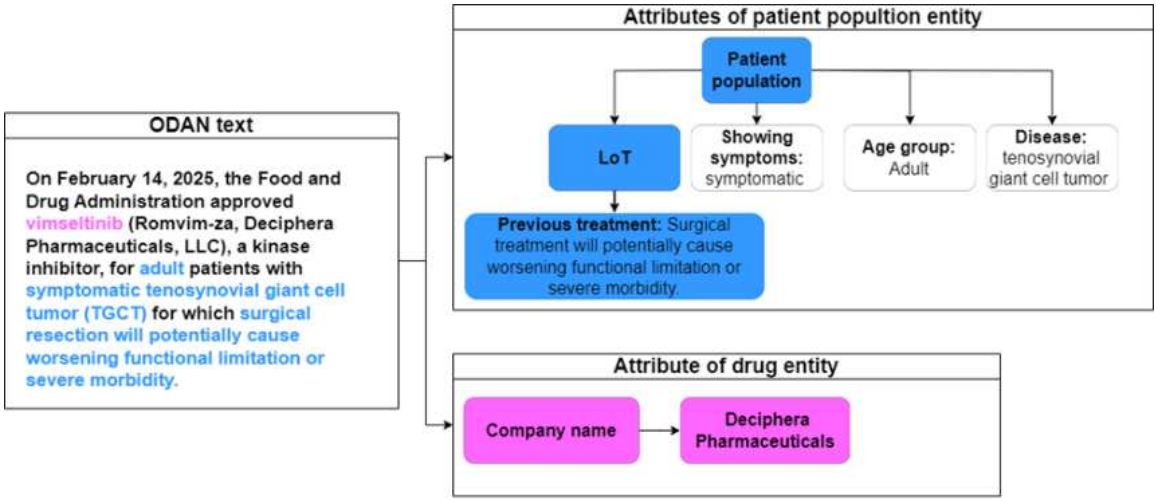


Figure 11. Example ODAN to demonstrate the difference in entity dependencies. The drug is pink, and the patient population and LoT are blue.

Furthermore, the syntactic structure of ODANs shows a difference in the surface-level presentation of the two entities (Figure 11). The regulatory agency and publisher are typically placed at the beginning, followed by the approved drug, the pharmaceutical company, and then patient information such as LoT and SoC. The presence and position of LoT within the patient information hierarchy are important for accurately communicating important medical information, but this results in a comparatively greater syntactic structure complexity than that of company names. Hence, although the placement of company names and LoT was not directly evaluated, the structure of ODANs is consistent with the trends seen in the BioNE CA, where LoT tends to be more complex than company names.

This organizational structure could create a nested or overlapping NER challenge for ML models when dealing with entities that exhibit this structural syntactic complexity. A nested entity occurs when an entity is contained within another entity, and an overlapping entity is one that shares parts of the text but has different boundaries, leading to issues such as entity boundary ambiguity, annotation complexity, modelling complexity, and context loss [25]. Therefore, unlike company names, LoT is more likely to exhibit span overlap and nesting, which adds to its complexity.

The models' performance for each entity, along with supporting qualitative analysis, suggests that the methods used by BioNE CA to quantify syntactic features demonstrate utility for differentiating entities based on linguistic characteristics. Certain biomedical entities capture detailed clinical information, such as medical terminology, treatment histories, and patient

descriptors, which require a diverse vocabulary, less repetition, more conjunctions, and varied context to connect concepts and ideas [26]. Consequently, the accurate identification of higher-complexity entities relies more heavily on contextual representation learning to create robust embeddings, whereas for simpler lexical entities, such as company names, straightforward lexical cues are sufficient [5]. However, consideration should be given that only two entities were chosen per category, and so the methods presented may be most applicable to entities with lexical profiles similar to those examined in this study.

6.2. Configuration strategies

This section primarily focuses on LoT, as the F1-scores for identifying company names were above 0.90 across all ConS permutations, with the worst-performing model being Distil-BERT on the refined dataset with no HPO.

6.2.1. Model comparison results

Table 8. Averaged F1-score percentage change of BioDistilBERT (presented) compared to DistilBERT.

Model being compared	Entities	Data Configuration Strategies					
		Unrefined	Unrefined+ HPO	Refined Strategy	Unrefined+ HPO	Refined + More Data	Refined+ MoreData +HPO
BioDistilBERT	Company Names	▼0.09%	▲0.18%	▲1.64%	▼0.02%	▲1.13%	▼0.65%
	LoT	▼04.36%	▲2.70%	▼3.84%	▼0.06%	▼3.64%	▲13.99%
			BEST for Company Name				**BEST for LoT**

▲: Improvement, ▼: Decline.

The best and worst models for identifying LoT were the domain models trained with and without HPO (Table 8). The primary difference between BioDistilBERT and DistilBERT is the data used for pre-training, which influences how well the models generalize to fine-tuning tasks [27]. The performance of BioDistilBERT without HPO, combined with its substantially larger gains after optimization, indicates that it is comparatively more dependent on fine-tuning configurations and dataset characteristics than DistilBERT, which may have contributed to the lower F1-scores.

Consequently, while the 2023 study, which used an ODAN corpus, found that a smaller non-domain model outperformed a larger domain model [2], this investigation, which controls for BERT model size, shows alignment between BERT pre-trained on biomedical data and ODAN. These observations could reflect differences in model capacity across datasets [28].

Table 9. Averaged F1-score percentage change of HPO models (presented) compared to non-HPO models.

Model	F1-score	Unrefined strategy	Refined strategy	Refined+more data
DistilBERT	Company	▲1.43%	▲2.99%	▲1.66%
	LoT	▲6.73%	▲11.48%	▲12.01%
BioDistilBERT	Company	▲1.70%	▲1.31%	▲1.18%
	LoT	**▲14.60%**	**▲15.86%**	**▲32.51%**

▲: Improvement, ▼: Decline.

HPO models performed better than their non-optimized counterparts regardless of entity (Table 9). The uniform improvement in BERT's performance suggests that HPO balances performance gains, as it reduces the need to compromise NER performance by identifying an optimal region within the hyperparameter space that enhances the model's generalisability [23]. A notable point is that BioDistilBERT benefited greatly from HPO, with relatively larger performance gains, particularly for LoT. This pattern further supports the idea that BioDistilBERT has a stronger dependence on fine-tuning configuration and optimization than DistilBERT.

Although HPO improved identification performance across all entities, distinct models performed best for each entity (Table 8). This observation, as supported by BioNE CA, is likely due to differences in the lexical characteristics of company names and LoT. Subsequently, the HPO method searched the hyperparameter space and identified configurations that maximized a model's overall performance given the specifications of each ConS permutation.

Table 10. Averaged F1-score percentage change between refined (presented) and unrefined datasets.

Model	F1-score	Refined	Refined+HPO

DistilBERT	**Company F1**	▼2.50%	▼1.00%
	LoT F1	▼11.63%	▼7.70%
BioDistilBERT	**Company F1**	▼0.81%	▼1.19%
	LoT F1	▼11.16%	▼10.18%
▲: Improvement,▼: Decline.			

The refined dataset generated using the tagging protocol reduced BERT's performance at identifying both entities (Table 10). Examination of the distribution patterns for LoT shows that the tagging protocol decreased the regularity of the first words tagged as "B-LoT" in the refined test dataset (Figure 12). This suggests a blurring of "B" (beginning) tag patterns, which the models use to anticipate "I" (inside) tag sequences [29]. Therefore, the weakened conditional relationship between B-LoT →I-LoT produced fuzzier patterns, undermining the predictive power for both classes and exacerbating the NER boundary problem.

The unrefined test dataset may therefore have preserved simpler, more consistent tagging patterns, either because of the data-splitting method or the effects of more aggressive tagging. Nevertheless, given the evidence, the unrefined dataset was inadvertently more refined.

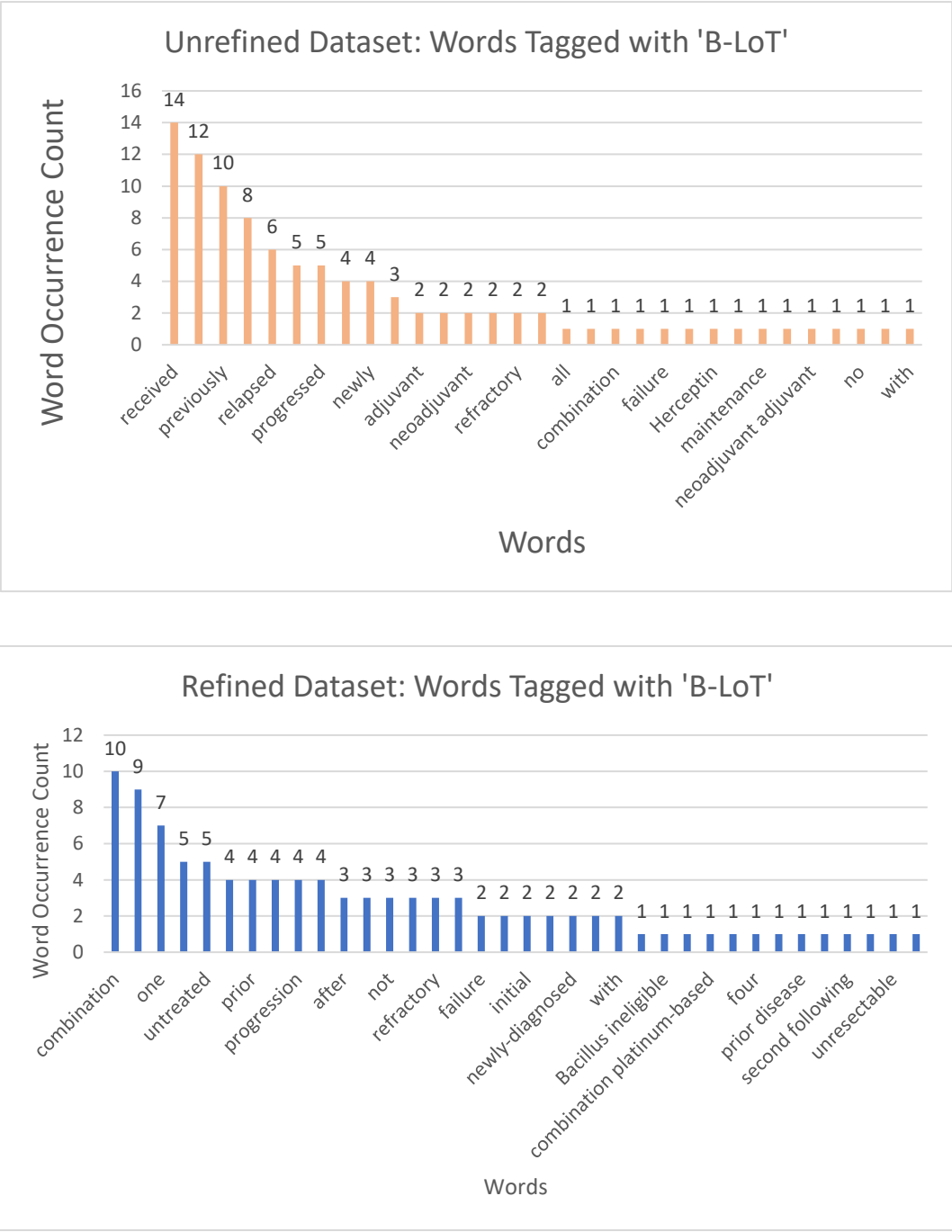


Figure 12. The refined dataset (bottom) has less first-word repetition than the unrefined dataset (top).

Table 11. F1-score percentage change between the refined + more data (presented) and the refined datasets

Model	F1-score	Refined + more data	HPO + refined + more data
DistilBERT	Company F1	▲0.44%	▼0.86%
	LoT F1	▲3.34%	▲3.84%
BioDistilBERT	Company F1	▼0.07%	▼0.20%
	LoT F1	▲3.56%	▲18.44%
▲: Improvement, ▼: Decline.			

When the models were exposed to more data, there was a slight increase in all ConS permutations concerning LoT performance (Table 11). However, BioDistilBERT underwent the most significant improvement, suggesting that to leverage its biomedical pre-training, increasing the number of additional records helped, specifically in conjunction with HPO.

Table 12. Averaged F1-score percentage improvement between the best performing ConS permutation (presented) and the worst.

Company Name	LoT
▲4.00%	**▲27.14%**
▲: Improvement, ▼: Decline.	

Table 12 showed the overall performance improvement between the worst and best performing ConS. BERT's performance increased substantially for LoT compared to company names, suggesting that the model's capability for finding additional useful lexical patterns is approaching saturation, while greater scope for LoT. Therefore, these results indicated that when utilizing appropriate optimization strategies, BERT's ability to identify LoT can be meaningfully improved.

Across all ConS, statistical significance remained consistent, indicating that performance differences in entity identification were robust to ConS permutations. However, in the systematic comparison of ConS, BERT's F1-scores for LoT increased proportionally more than for company names. This pattern indicates that the disparity between entity types can be

influenced by model configuration and data characteristics. While the results still confirm a consistent difference in BERT's entity detection performance across lexical complexity levels, they also demonstrated that optimization strategies can reduce the performance gap.

Although it is challenging to draw conclusions, the data suggested that if ML models can effectively identify useful patterns associated with entities, thereby alleviating the boundary problem, their overall entity recognition performance, especially for lagging entities, improves. This research showed that performance can be enhanced through informed model selection based on data alignment, increasing the quantity and quality of the data, exposing the model to the true distribution of the data, appropriate tagging, and optimizing hyperparameters to enhance the model's ability to extract patterns.

Beyond these optimization strategies, more advanced techniques could be used to further enhance NER performance. These include integrating knowledge graphs that provide structured relationships between entities, supporting disambiguation and reinforcing relevant lexical and semantic patterns; debiasing techniques to improve generalization by reducing the reliance on superficial patterns and encouraging models to learn contextual cues; and joint labelling approaches to manage nested or overlapping entities. Beyond traditional NER, complexity analysis may also offer a conceptual basis for future exploration of prompt design in large language models by characterizing the lexical and structural features associated with different entity types.

8. Conclusions

ML methods can identify a range of biomedical entities from text, although their performance often varies depending on the syntactic and semantic complexity of those entities. Based on this limitation, this study investigated whether initial entity characterization by lexical complexity is reflected in BERT's performance. Therefore, unlike prior studies that infer complexity from model performance, this investigation operationalized entity complexity prior to model evaluation, thereby demonstrating alignment between linguistic features and NER outcomes, which could potentially aid in informing targeted optimization strategies.

The results showed a significant difference in the model's ability to identify entities based on syntactic and semantic complexity, demonstrating that quantitative analyses can reveal the lexical complexity of biomedical entities. Compared to company names, LoT exhibited a greater word count, stronger right skew, fewer common surrounding words, less repetition, and greater POS and CTTR.

Regarding the implementation of optimization strategies, BERT's ability to identify the lagging entity improved, such that the domain model, combined with increased training data and HPO, attained the highest F1-scores for LoT. While the refined dataset did not improve model performance, the unrefined dataset with more consistent annotations saw comparatively better performance. Consequently, in the context of ML models and biomedical data, optimization strategies could help alleviate performance disparities between entity types, and perhaps more advanced techniques may further enhance performance, potentially mitigating lexical disparities between entities.

Limitations included the small number of models per ConS and potential overlap in contextual features, which may affect reproducibility and statistical rigour. The MWU test assumes independence between entities, an assumption that may be weakened in contextual language models where overlapping lexical patterns are common in biomedical text. Therefore, a statistical test that accounts for correlation within the groups, such as generalized linear models (GLMs), may have been more appropriate [30].

Another consideration is the potential introduction of bias into the interpretation of BioNE CA results. This stemmed from the reliance on the literature to select entities, which may have compromised BioNE CA's impartiality. Hence, although the statistical patterns supported the entity categorization, future applications should adopt a blinded technique to enhance objectivity.

Consequently, enhancing designation integrity in a blind setting requires improving supporting approaches, including the addition of qualitative and improved quantitative methods to further reveal the context of entities. For instance, including stop word analysis and understanding common phrase placements, with support from entity frequency and clear empirical boundaries, would have provided more contextual understanding and objectivity, improving interpretability. Meanwhile, removing the use of literature to identify potential entities would mitigate the introduction of bias. Future research could expand on BioNE CA beyond ODAN, evaluating whether the methodology can characterize entities by complexity across diverse biomedical corpora. Additionally, BioNE CA could be assessed for its integration into active learning strategies where samples of greater complexity are selectively used for fine-tuning prompt-based LLMs. Such approaches could further enhance adaptability, generalization, and interpretability in biomedical NER systems.

9. Declarations

9.1. Ethics approval

N/A

9.2. Consent for participation and publication

N/A

9.3. Availability of data and materials

Data and materials will be made available upon requests to the authors.

9.4. Competing interests

N/A

9.5. Funding

N/A

9.6. Author contributions

Conceptualization: Gregory Desouza and Andrea Corradini

Project Administration: Gregory Desouza and Andrea Corradini

Supervision: Andrea Corradini

Methodology: Gregory Desouza

Data Curation: Gregory Desouza and Andrea Corradini

Investigation: Gregory Desouza

Formal Analysis: Gregory Desouza

Writing, review and editing: Gregory Desouza and Andrea Corradini

9.7. Acknowledgments

N/A

9.8. Clinical Trial Registration

N/A

References

1. Le B, Pham TQ, Van Hoang AT, Nguyen BT. OphNER: Named Entity Recognition for Ophthalmology Newspapers. 2023 2023:1-6.

2. Gendrin A, Souliotis L, Loudon-Griffiths J, et al. Identifying Patient Populations in Texts Describing Drug Approvals Through Deep Learning–Based Information Extraction:

Development of a Natural Language Processing Algorithm. *JMIR Formative Research*. 2023;7:e44876.

3. Pakhale K. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:230914084*. 2023;

4. Agrawal A, Tripathi S, Vardhan M, Sihag V, Choudhary G, Dragoni N. BERT-based transfer-learning approach for nested named-entity recognition using joint labeling. *Applied Sciences*. 2022;12(3):976.

5. Kim H, Kang J. How do your biomedical named entity recognition models generalize to novel entities? *Ieee Access*. 2022;10:31513-31523.

6. Chen K, Cosgro B, Domfeh O, Stern A, Korkmaz G, Kattampallil NA. Leveraging Google BERT to detect and measure innovation discussed in news articles. 2021 2021:1-6.

7. Niero LHP, de Souza JVA, da Silva LMG, et al. Challenges and Issues on Extracting named Entities from Oncology Clinical Notes. *Journal of Health Informatics*. 2023;15(Especial)

8. Pooja H, Jagadeesh MPP. A Deep Learning Based Approach for Biomedical Named Entity Recognition Using Multitasking Transfer Learning with BiLSTM, BERT and CRF. *SN Computer Science*. 2024;5(5):482.

9. Ezen-Can A. A Comparison of LSTM and BERT for Small Corpus. *arXiv preprint arXiv:200905451*. 2020;

10. Park Y-J, Lee M-a, Yang G-J, Park SJ, Sohn C-B. Biomedical text NER tagging tool with web interface for generating BERT-based fine-tuning dataset. *Applied Sciences*. 2022;12(23):12012.

11. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*. 2022;29(7):1208-1216.

12. Mulang IO, Singh K, Prabhu C, Nadgeri A, Hoffart J, Lehmann J. Evaluating the impact of knowledge graph context on entity disambiguation models. 2020 2020:2157-2160.

13. Sang ETK, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. 2003:142-147.

14. Malvern D, Richards B, Chipere N, Durán P. *Lexical diversity and language development*. Springer; 2004.

15. Reviriego P, Conde J, Merino-Gómez E, Martínez G, Hernández JA. Playing with words: Comparing the vocabulary and lexical richness of ChatGPT and humans. *arXiv preprint arXiv:230807462*. 2023;

16. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:200410964*. 2020;

17. Mitra-Majumdar M, Gunter SJ, Kesselheim AS, et al. Analysis of supportive evidence for US Food and Drug Administration approvals of novel drugs in 2020. *JAMA Network Open*. 2022;5(5):e2212454-e2212454.

18. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. 2018;

19. Cho HN, Jun TJ, Kim Y-H, et al. Task-specific transformer-based language models in health care: scoping review. *JMIR Medical Informatics*. 2024;12:e49724.

20. Sanh V. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:191001108*. 2019;

21. Rohanian O, Nouriborji M, Kouchaki S, Clifton DA. On the effectiveness of compact biomedical transformers. *Bioinformatics*. 2023;39(3):btad103.

22. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Computational linguistics*. 2008;34(4):555-596.

23. Bischl B, Binder M, Lang M, et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2023;13(2):e1484.

24. Vajjala S, Balasubramaniam R. What do we Really Know about State of the Art NER? *arXiv preprint arXiv:220500034*. 2022;

25. Hu Z, Hou W, Liu X. Deep learning for named entity recognition: a survey. *Neural Computing and Applications*. 2024;36(16):8995-9022.

26. Jolly A, Pandey V, Singh I, Sharma N. Exploring biomedical named entity recognition via SciSpaCy and BioBERT models. *The Open Biomedical Engineering Journal*. 2024;18(1)

27. Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018:328-339.

28. Zhong R, Ghosh D, Klein D, Steinhardt J. Are larger pretrained language models uniformly better? comparing performance at the instance level. 2021:3813-3827.

29. Ramshaw L, Marcus M. Text chunking using transformation-based learning.

30. Bolker BM. Linear and generalized linear mixed models. *Ecological statistics: contemporary theory and application*. 2015;2015:309-333.