Journal of Healthcare Informatics Research (2023) 7:501–526
https://doi.org/10.1007/s41666-023-00146-1

RESEARCH ARTICLE

Sequence Labeling for Disambiguating Medical Abbreviations

Mucahit Cevik[1] · **Sanaz Mohammad Jafari**[1] · **Mitchell Myers**[1] · **Savas Yildirim**[1,2]

Received: 2 October 2022 / Revised: 2 June 2023 / Accepted: 29 August 2023 /
Published online: 14 September 2023

Abstract

Abbreviations are unavoidable yet critical parts of the medical text. Using abbreviations, especially in clinical patient notes, can save time and space, protect sensitive information, and help avoid repetitions. However, most abbreviations might have multiple senses, and the lack of a standardized mapping system makes disambiguating abbreviations a difficult and time-consuming task. The main objective of this study is to examine the feasibility of sequence labeling methods for medical abbreviation disambiguation. Specifically, we explore the capability of sequence labeling methods to deal with multiple unique abbreviations in a single text. We use two public datasets to compare and contrast the performance of several transformer models pretrained on different scientific and medical corpora. Our proposed sequence labeling approach outperforms the more commonly used text classification models for the abbreviation disambiguation task. In particular, the SciBERT model shows a strong performance for both sequence labeling and text classification tasks over the two considered datasets. Furthermore, we find that abbreviation disambiguation performance for the text classification models becomes comparable to that of sequence labeling only when postprocessing is applied to their predictions, which involves filtering possible labels for an abbreviation based on the training data.

Keywords Abbreviation disambiguation · Medical text · Sequence labeling · Transformers models

✉ Mucahit Cevik
mcevik@torontomu.ca

[1] Toronto Metropolitan University, 44 Gerrard St E, Toronto M5B 1G3, Ontario, Canada

[2] Faculty of Engineering and Natural Sciences, Istanbul Bilgi Univeristy, Eyüpsultan 34060, İstanbul, Turkey

1 Introduction

Word-sense disambiguation (WSD) is the task of identifying the correct sense of an ambiguous word with several possible meanings given its context [1]. For example, the term *well* can be associated with health, i.e., *I am feeling well* or a source of water, i.e., *You can drink from the well*. Humans have the innate ability to do this and often differentiate between word senses subconsciously. However, automated WSD is considered to be one of the most challenging tasks in natural language processing (NLP) [2]. Abbreviation disambiguation (AD) is a sub-task of WSD that focuses on accurately expanding or decoding ambiguous abbreviations (ABV) in text data (e.g., ABI for *Acquired Brain Injury*). While humans are able to interpret common ABVs to a degree, full-word WSD is a much more difficult task as they are usually domain-specific and require prior knowledge. In this paper, we focus on AD in medical texts, and in this case, disambiguation can be in the form of providing the actual long-form version of the ABV or outputting key additional context words associated with it so that any confusion about the sense is removed, assuming the reader is a medical professional. Table 1 provides examples of these two forms of AD.

In the medical domain, especially in clinical notes, abbreviations are extremely common due to their ability to increase efficiency and protect patients' privacy [3, 4]. One study found that an estimated 30–50% of clinical notes were made up of abbreviations [5]. However, while there are benefits to ABVs, they can sometimes be obscure to their actual sense making them un-intuitive and more difficult to decipher without additional information. Furthermore, ABVs can have multiple meanings (i.e., long-form versions), and the actual sense depends heavily on the context. Moreover, new ABVs are regularly being created without a standardized mapping system. This makes AD a difficult and time-consuming task that can delay the extremely important flow of communication in medical operations. These challenges in AD point to the necessity for building reliable alternatives using WSD techniques.

WSD is a heavily researched area in NLP, and the majority of proposed methods can be grouped into three categories: knowledge-based [6], unsupervised [7], and supervised [8]. Knowledge-based methods rely mainly on dictionaries, sense inventories, and hand-crafted rules to predict the correct sense of an ambiguous word. Unsupervised methods do not need sense inventories and, instead, mainly employ clustering methods to differentiate between different senses and contexts. Finally, supervised methods constitute the most commonly employed techniques for WSD; they typically require annotated data, and classifiers trained over this data are used to detect the correct sense of the abbreviation. Common supervised WSD methods include decision trees (DT) [9], support vector machines (SVM) [10, 11], naive Bayes (NB) [12], and neural networks [13].

Table 1 Examples of AD outputs

Output type	ABV	Label for prediction task	Actual ABV sense
Long form version	MBF	Myocardial blood flow	Myocardial blood flow
Key context word	IF	Staining	Immunofluorescence staining

Most research in AD looks at the task through text classification, where, given a piece of text with at least one ABV present, a label is assigned to the whole text that decodes the ABV. Multiple ABVs in a text can complicate the text classification approaches, as it is possible to confuse the labels that are associated with the ABVs. Therefore, in this study, we propose to frame the WSD problem as a sequence labeling task, which can automatically handle multiple ABVs occurring in the same text. We employ recent state-of-the-art methods to find the most effective solution strategy. Well-known NER methods include probabilistic conditional random fields (CRF) [14] and transformer models such as BERT [8].

We conduct a detailed empirical analysis to investigate the feasibility of sequence labeling methods for the WSD task with the objective of providing a more practical means to automatically expand or disambiguate medical ABVs in text. We particularly focus on pre-trained BERT models to compare text classification and sequence labeling for the WSD task. We compare these models against a CRF-based approach as well. The main contributions of our study can be summarized as follows:

- We apply sequence labeling methods for medical AD tasks. Our work differs from previous studies in that they assume the abbreviations are known as a predefined list and they apply WSD to determine the correct senses. In contrast, our study has two steps. In the first step, the model identifies the abbreviation since it is unknown and then applies sequence labeling methods to determine the semantic category. Our findings suggest that the sequence labeling methods can outperform the text classification approaches for these AD tasks.
- We propose a postprocessing strategy for ABV prediction, which filters the candidate set of labels for any individual ABV based on the corresponding occurrences in the training data. We find that text classification methods benefit significantly from postprocessing. On the other hand, sequence labeling methods, thanks to their ability to examine the data at a more granular level, do not require postprocessing to achieve a high-performance level.
- We conduct an extensive numerical study with two unique medical AD datasets and several text classification and sequence labeling models. In this regard, our study contributes to a better understanding of the capabilities of these state-of-the-art methods for medical AD tasks.

The rest of the paper is structured as follows. Section 2 explores previous literature on NER as well as medical AD and WSD tasks. Section 3 provides an exploratory analysis of our datasets and details the proposed methodologies, experimental design, and evaluation metrics. Section 4 presents results from our detailed numerical study. Finally, Sect. 5 concludes the paper with a summary of our main findings and a discussion on future research directions.

2 Related Work

We briefly review the previous works on medical AD and NER and discuss the performance of different models and techniques in these tasks. AD and WSD research has become fairly popular in recent years, and while they have been applied to several

domains, we focus our review on WSD in the medical domain. A summary of the most closely related studies to our work is provided in Table 2.

Pakhomov et al. [9] were among the first to investigate medical acronym disambiguation and looked at the feasibility of semi-supervised learning for this task. By generating training data for each sense based on their context found across three large external sources including the World Wide Web, they were able to show the utility of leveraging massive publicly available data for medical AD. Xu et al. [7] built on this work, and they not only included more external data sources for training data generation but also incorporated sense frequency information as an additional feature. While semi-supervised learning methods have shown encouraging performance when there is a lack of annotated corpora, supervised learning is a much more popular method for AD. Joshi et al. [10] and Moon et al. [11] compared different supervised learning approaches, including NB, SVM, and DT for the WSD task. Joshi et al. [10] extracted several key features from the AD datasets and found that all three models (NB, SVM, DT) performed fairly equally, but overall, performance was maximized when all features were included in the training. Moon et al. [11] focused on minimizing training time and used these models to examine the impact of different window sizes around the target ABV as well as finding the minimum training samples required for each label to achieve reasonable performance.

The use of word embeddings for model training has become increasingly common in many NLP tasks, including text classification, information retrieval, and language translation. Wu et al. [16] explored the impact of neural word embeddings trained on an extremely large medical corpus in the medical WSD task. On top of the traditional WSD feature set, by adding two unique embedding-based features to their SVM classifier, one that took the max score of each embedding dimension from all surrounding words and another one that took the sum of the embedding vectors from the surrounding words, their model was able to achieve state-of-the-art results on their test datasets [16]. Jaber and Martínez [12] conducted a similar study and compared different word embedding strategies on two different models: SVM and NB. Their results showed that SVM outperformed NB overall, and the best performance was achieved when using word embeddings generated from both medical and general data sources.

Deep learning-based approaches have been popular for AD tasks in recent years. Li et al. [17] proposed a neural topic attention model for the medical AD task where they took a few-shot-learning approach that combined topic attention and contextualized word embeddings learned from ELMo [24]. Applying the topic information and word embeddings to a long-short-term-memory (LSTM) model yielded the best results in their experiments. Jin et al. [19] proposed the DEep Contextualized Biomedical ABV Expansion (DECBAE) model that utilizes BioELMo [25] word embeddings, a domain-specific version of ELMo [24], and a fine-tuned Bidirectional-LSTM (BiLSTM) model to achieve state-of-the-art performance. Similar to ELMo and its descendent BioELMo, the pre-trained BERT_{base} [8] models offer complex word embeddings that apply to a wide array of topics. However, minimal or no exposure to key biomedical terms limits the benefits of transfer learning in this domain. Accordingly, Lee et al. [13] proposed BioBERT, a pre-trained model for biomedical text mining. Jaber and Martínez [4] explored different BERT variants for the medical AD task and found that the variations pre-trained on medical text such as BioBERT outperformed BERT_{base}.

Table 2 Summary of the relevant papers

Study	Model	Methodology	Datasets
Pakhomov et al. [9]	C5.0 decision tree [15]	Leverage additional context found online to disambiguate ABV	Clinical notes of Mayo Clinic[a]
Xu et al. [7]	Profile-based method	Combine sense frequency information with a profile-based method	NYPH[b] discharge summary corpus and physician-typed hospital admission notes
Joshi et al. [10]	NB, DT, and SVM	Include POS tags, unigrams, and bigrams to improve the accuracy	Clinical notes of Mayo Clinic[a]
Moon et al. [11]	NB, DT, and SVM	Optimize the window size for each orientation and determine the minimum training sample size	Clinical notes of Fairview Health Services[c]
Wu et al. [16]	SVM	Investigate three different embedding methods to improve the feature set of SVM	Annotated ABV of the Vanderbilt University Hospital's admission notes and clinical notes of the University of Minnesota (UMN)
Jaber and Martínez [12]	NB, SVM	Investigate four strategies to use pre-trained word embedding as features	Clinical notes of the UMN
Li et al. [17]	Traditional ML models, NN models, and proposed ELMo+Topic model	Incorporate topic attention on ELMo word representation	UMN, PubMed[d] and MIMIC-III [18]
Jin et al. [19]	DECBAE	Utilize BioELMo to extract features and pass them to BiLSTM	PubMed
Lee et al. [13]	BioBERT	Pre-train BERT on medical corpora	NCBI Disease [20], GAD [21], BioASQ 6b-factoid [22]
Jaber and Martínez [4]	One-fits-all classifier	Added simple neural network structure on top of the pre-trained BERT structure	UMN
Our study	Adaptation of various BERT models	Incorporate sequence labeling for AD task	MeDAL [23] and UMN

[a]from https://www.mayoclinic.org/
[b]from https://www.nyp.org/
[c]from https://www.fairview.org/
[d]from https://pubmed.ncbi.nlm.nih.gov/

While many studies on medical AD referenced above consider the same objective as in our study, they mainly focus on single- or multi-label classification. We set out to investigate the potential of NER methods for the WSD task. NER is a sequence labeling technique that identifies and categorizes key information and entities in unstructured text. There are three traditional approaches to address the NER problem: rule-based [26, 27], unsupervised learning [28], and feature-based supervised learning [29, 30]. The majority of NER studies follow the supervised learning approach, and accordingly, several classifiers have been explored for NER over the years. However, the CRF model has been shown to be highly effective for NER tasks due to its ability to capture the context around the target word or entity. McCallum [14] proposed a feature induction method for CRF in NER. By iteratively adding features and only focusing on the ones that maximize the log-likelihood, they were able to greatly improve the NER performance over a fixed feature set approach, while also controlling the training time. More recently, transfer learning has become popular for NER with the advent of ELMo and BERT models [8, 24]. Souza et al. [31] proposed a BERT-CRF model in a Portuguese NER task. Additionally, biomedical named entity recognition (BioNER) was developed to identify entities such as genes, proteins, diseases, chemicals, and species, in medical texts and clinical notes. For instance, Liu et al. [32] proposed K-BERT, a BERT model that can be injected with domain-specific knowledge more efficiently than pre-training. In several domain-specific NER tasks, including a medical NER task, K-BERT was shown to outperform regular BERT.

The methods discussed above are highly relevant to our study, particularly the work by Jaber and Martínez [4]. Therefore, these relevant techniques are implemented as baselines in our numerical study. We note that the key difference between our work and the previous literature is that we apply NER methods for a medical WSD task. Not only does our model have to predict whether a token is an ABV, but it also needs to output the correct label depending on the context. While this is similar to regular BioNER, in BioNER, there are usually only a few (e.g., < 10) possible NER labels, whereas, in AD/WSD tasks, we are required to deal with more than 1000 labels. We also note that based on these works and our empirical observations, transformers-based methods perform well in sequential labeling tasks such as NER and POS due to their transfer learning capabilities and ease of fine-tuning. Since our task resembles a sequential labeling problem, we mainly develop our solution strategy based on transformer architectures. Particularly, for the classification problems, we consider the BERT architecture.

3 Methodology

In this section, the sequence labeling methods employed in our analysis are explained, and the structure of the classification models and datasets are reviewed. Furthermore, the details of the experimental setup, hyperparameter tuning, and evaluation metrics are provided.

Recent studies on WSD mainly focus on text classification where, given a piece of text with an abbreviation present, the entire text is labeled with the correct sense of the ABV. However, since datasets can have multiple unique abbreviations in each

instance, we take an alternative approach to solving the WSD task. While multi-label classification methods can also be considered for this task, the challenge would then become to ensure that each predicted label is correctly assigned to its respective ABV. Further complications might also arise when the number of labels predicted does not match the number of target ABVs. Sequence labeling, on the other hand, allows us to assign a single label to each token in the text, and therefore, it enables assigning each abbreviation to its corresponding sense. As such, it allows easy interpretation of the predictions and adoption of these methods beyond these experiments.

3.1 Classification Models

We provide the details of each of the seven models used in our experiments below. CRF and BiLSTM are considered baseline models, and we employ five BERT variants for our main comparative analysis: DistilBERT, BioBERT, BlueBERT, MS-BERT, and SciBERT.

3.1.1 CRF

CRF is a model that uses a probabilistic approach in modeling sequential data [33]. CRFs are popular in part-of-speech (POS) tagging, NER, and other token classification tasks due to their ability to learn sequential contexts in text and utilize domain and data-specific handcrafted features to predict the label for each token. For our experiments, several features are designed for the CRF model including word components such as prefix and suffix, capitalization, and a check flag to determine whether each token is an abbreviation or not. We also capture contextual features that examine the nearest neighbor on either side of the current word (i.e., with window size $[-1, 1]$). While this model is commonly used for sequence labeling tasks, one drawback is its computational complexity as training time drastically increases with the sequence length and the number of labels. Accordingly, we consider the CRF model as a baseline for our experiments on a reduced-label dataset, which contains only a subset of all the available labels.

3.1.2 BiLSTM

An LSTM model is a recurrent neural network (RNN) that can find and maintain long-range dependencies in data [34, 35]. Due to the problem of vanishing gradients, regular RNNs are limited by how much important information can be stored in their *memory*. However, with LSTMs, thanks to a complex gating system, key information, no matter how far back in the sequence it is, can be maintained and accessed for the prediction task. A popular extension to the LSTM model is the bidirectional LSTM (BiLSTM), which has the ability to learn long-range dependencies in both directions—from left to right and right to left. The BiLSTM model has become a common foundation for several sequence tagging tasks [36], and it is considered a baseline in our analysis.

3.1.3 BERT

BERT is a pre-trained language model and stands for Bidirectional Encoder Representations from Transformers [8]. BERT's model architecture is based on the transformer

model proposed by Vaswani et al. [37], which is an attention mechanism that can be used to learn relationships between words and sub-words. Similar to the BiLSTM, the bidirectional nature of this model allows learning the context of a word based on the words both from left and right. A notable challenge in pre-training is defining a prediction task. As the main training strategy, BERT employs a masked language model objective that randomly masks a proportion of tokens in the input. The task is to predict the actual masked word based on its context. Additionally, BERT employs a next-sentence prediction task which requires the model to predict whether the second of two input sentences actually follows the first. In pre-training stage for these tasks, the BooksCorpus of 800 M words and English Wikipedia of 2500 M words are employed. BERT has achieved state-of-the-art performance on several NLP tasks. Accordingly, five BERT variants are considered in our experiments. These include a popular lighter BERT variant (DistillBERT), three models that are pre-trained on different medical corpora (BioBERT, BlueBERT, MS-BERT), and another model that is pre-trained on scientific text with its custom domain-specific vocabulary (SciBERT). Below, we briefly summarize each of these BERT variants.

- *DistilBERT*: This model is the *faster, cheaper, and lighter* version of the original BERT [38]. While full-size transformers-based models offer outstanding performance, their usage is limited by high computational complexity. DistilBERT is based on the same architecture as BERT but in a condensed and more efficient form. DistilBERT was shown to retain 97% of BERT's performance while being 60% faster [38].
- *BioBERT*: BioBERT is a BERT variant that is additionally pre-trained on large-scale biomedical corpora to excel in biomedical text-mining tasks. Specifically, it is pre-trained on PubMed abstracts that contain 4.5 B words and PubMed Central full-text articles with 13.5 B words.
- *BlueBERT*: Similar to BioBERT, BlueBERT (Biomedical Language Understanding Evaluation) was pre-trained on biomedical data [39]. This BERT variant trained on a very large corpus of more than 4 B words from PubMed abstracts and over 500 million words from MIMIC-III clinical notes.
- *MS-BERT*: This model is an extension of BlueBERT that is further pre-trained on over 35 million words extracted from multiple sclerosis clinical notes collected between 2015 and 2019 in Toronto (hence the *MS* in MS-BERT) [40].
- *SciBERT*: This model has two key differences from BERT_{base}: its pre-training corpus and vocabulary [41]. Similar to the other domain-specific variations, SciBERT is pre-trained on scientific literature. The corpus consisted of 1.14 million articles from Semantic Scholar both from the computer science and biomedical domains, resulting in 3.17 B words. Unique only to SciBERT, however, is its vocabulary. Generating its own vocabulary instead of reusing the one from BERT_{base} allowed SciBERT to capture the most frequently occurring words and sub-words from its specific domain.

3.2 Datasets

We use two distinct medical text datasets in our experiments: MeDAL [23] and UMN [42]. The MeDAL dataset consists of medical abstracts where certain long-form words have been manually swapped with their abbreviated form. A key feature of this dataset is the occurrence of multiple unique ABVs in one abstract. The UMN dataset, curated by the University of Minnesota's Digital Conservancy, is a collection of raw, anonymized clinical patient notes. There is only one target ABV with an associated label that can occur multiple times in this dataset. The MeDAL dataset offers the unique challenge of dealing with multiple unique ABVs in one instance, and hence, it provides a strong motivation for the use of sequence labeling methods. On the other hand, the UMN dataset potentially provides a more likely real-life application of AD, since it is composed of raw clinical notes. The following section describes each dataset in detail.

3.2.1 MeDAL Dataset

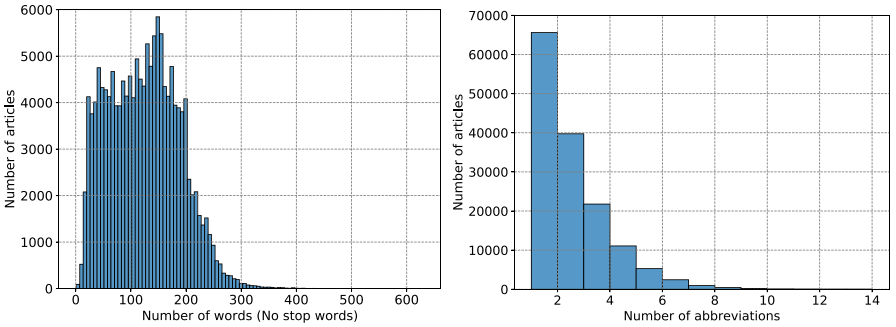
In our analysis, 2% of the full 14 million MeDAL abstract dataset was used. In this subset, there are 288,080 rows, and each row contains an abstract, the location of each ABV given by its index, and the corresponding long-form versions or senses. An example of a MeDAL dataset instance can be found in Table 3.

There are 557,248 unique words, 4866 unique ABVs, and 16,299 unique labels in the MeDAL dataset. However, to further reduce computational complexity and training time in our experiments, only the 300 most frequent ABVs and their 1005 most frequent labels were selected. This restriction reduced total rows to 147,728 and unique words to 320,168.

Figure 1 illustrates ABV and word count distribution across abstracts. The mean ABV count is 2.1 while the mean word count is 124. However, by removing the stop words like *the*, *at*, or *how*, the average word count drops to 77. In this subset, there are on average 3.35 unique senses associated with each abbreviation, with a minimum of 1 possible label (least ambiguous) to a maximum of 18 (most ambiguous). Figure 2

Table 3 A sample MeDAL data instance

Text	Locations	Labels
The kinetic disposition and beta-adrenergic blocking action in relation to the plasma level of a single oral dose proportional to the extent of the histamine release. It is concluded that the reduction in the in vitro amine uptake after anaphylactic and compound-induced histamine release is due to the fact that there are fewer intact granules capable of storing histamine and not primarily due to a damage to the mechanisms by which mast cells take up **BA** in vitro the observations further strengthen the view that anaphylactic and compound-induced **HR** are noncytolytic processes	(76, 90)	("**Biogenic Amines**," "**Histamine Release**")



(a) Word count distribution (b) Abbreviation count distribution

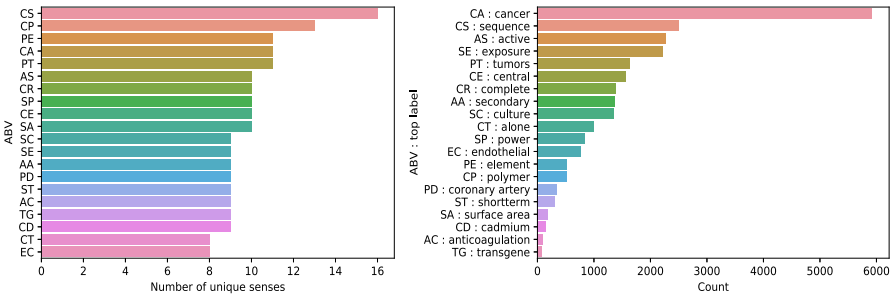
Fig. 1 Distribution of number of words and abbreviations for MeDAL dataset

shows the 20 ABVs with the greatest number of unique labels (see Fig. 2a), which are considered the most ambiguous, as well as the label from each ABV with the greatest number of occurrences in the dataset (see Fig. 2b).

These figures show that the dataset is imbalanced; specifically, the distribution of the number of examples across all MeDAL labels span from a minimum of 14 up to over 18,000. Figures 3 a and b display the most frequent bi- and tri-grams. The *n*-gram phrases and sample text in Table 3 show that most of the text is well structured and most terms revolve around the themes of study, research, and medical experimentation.

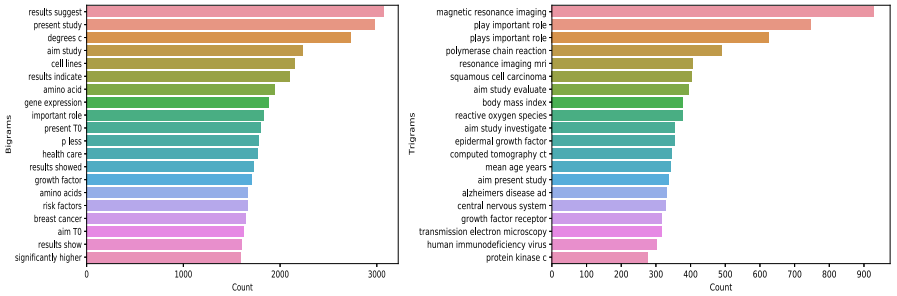
3.2.2 UMN Dataset

The UMN dataset consists of 36,996 rows, 74 unique ABVs, and 346 unique labels. There are a total of 39,110 unique words in the text column. To reduce the complexity of the dataset, any label with fewer than five examples is dropped from the dataset, resulting in a final total of 203 labels and 72 ABVs. Only including these labels in the dataset reduces the total rows to 35,518 and the total words to 37,283. As seen in Fig. 4, the word count distribution is centered around its mean of 39 words, or 59,



(a) Most ambiguous ABVs (Greatest number of unique labels) (b) Most frequent label of ambiguous ABVs

Fig. 2 Most ambiguous ABVs and their most frequently occurring labels in the MeDAL dataset



(a) Bi-grams distribution (b) Tri-grams distribution

Fig. 3 Top 20 most frequent bi- and tri-grams in the MeDAL dataset

including stop words. Table 4 presents a sample instance of the UMN dataset, which has a very similar format to the MeDAL dataset. However, in each row, there can only be one unique ABV present that can occur multiple times.

Looking deeper into the ABVs, there are on average 2.92 labels per ABV and the dispersion between the minimum and maximum number of senses is 1 and 8, respectively. Figure 5 displays the most ambiguous ABVs in UMN along with each of their top labels. Compared to the MeDAL, the dataset is less imbalanced with a minimum of 5 and a maximum of 1774 examples for a label, respectively. Finally, Figs. 6 a and b display the top bi- and tri-grams in the UMN dataset. We note that the text is more unstructured than the MeDAL dataset and, as expected, the distribution is centered around prescriptions and patient analysis.

3.2.3 Data Preprocessing

In our numerical analysis, to focus the models on more topic-specific terms and also to reduce training time, punctuation and stop words were removed from the text, any rows with ABVs beyond the 110th word index were dropped, and text columns were

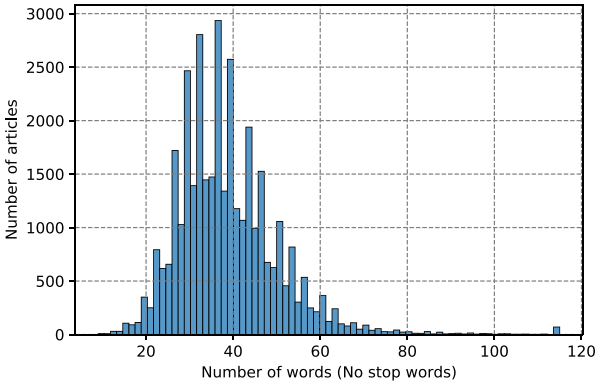


Fig. 4 Word count distribution of the abstracts in the UMN dataset

Table 4 A sample UMN data instance

Text	Locations	Labels
Her **PA** pressures were 44/26 with a wedge of 22 with a CVP of 10 and her heart rate of 120 to 139. Her cardiac index was 3. Nursing made aggressive attempts to bring her **PA** pressures down, and on the day 2 of admission, she was found to have her Nipride running at 7 mcg/kg/min. This was quickly weaned, and captopril was instituted with hydralazine as needed.	(1, 35)	("**Pulmonary Artery**," "**Pulmonary Artery**")

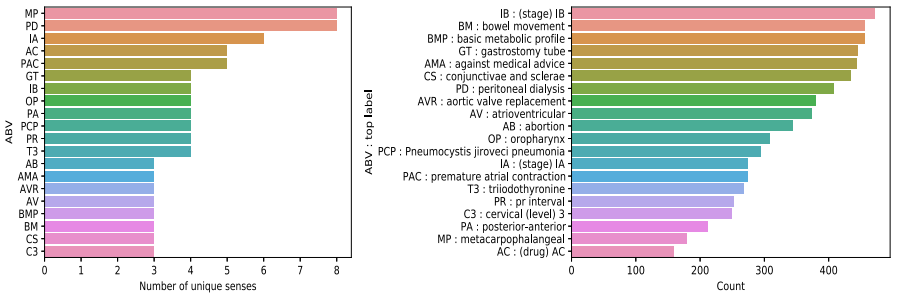
truncated to a maximum length of 115 words. Additionally, to satisfy sequence labeling model requirements, each token was mapped to its corresponding label; if the token was a regular word, it was assigned *NA_word*, and if it was an ABV, it was mapped to its corresponding sense.

The MeDAL dataset has an abundance of labels with an excessive amount of examples. While this does not necessarily hinder prediction performance, using all training examples can increase training time without significant performance improvements. Accordingly, for the MeDAL dataset, we dropped any instance where all labels in the row had at least 500 other rows to reference. These steps reduced the number of raws in the dataset by half from 147,728 to 73,196.

Both datasets are further subsampled to ensure complete training, reduce the training time, and avoid resource limitations in certain experiments. Therefore, only a subset of each dataset containing their respective top 12 most frequent ABVs and 40 most frequent labels was used. After applying the same preprocessing steps as described for MeDAL, 8472 rows remain in the UMN subset. We refer to the corresponding datasets as MeDAL-40 and UMN-40, respectively, in the rest of the paper.

3.3 Experimental Setup

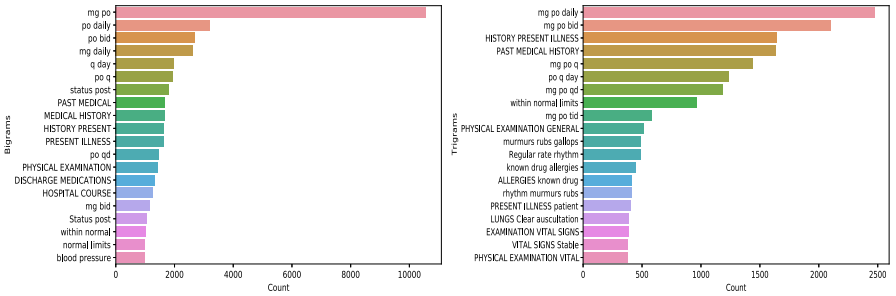
We evaluate our proposed sequence labeling method for AD using different models over MeDAL and UMN datasets. Five different BERT-based models including



(a) Most ambiguous ABVs (Greatest number of unique labels)

(b) Most frequent label of ambiguous ABVs

Fig. 5 Most ambiguous ABVs and their most frequently occurring label in the UMN dataset



(a) Bi-grams distribution (b) Tri-grams distribution

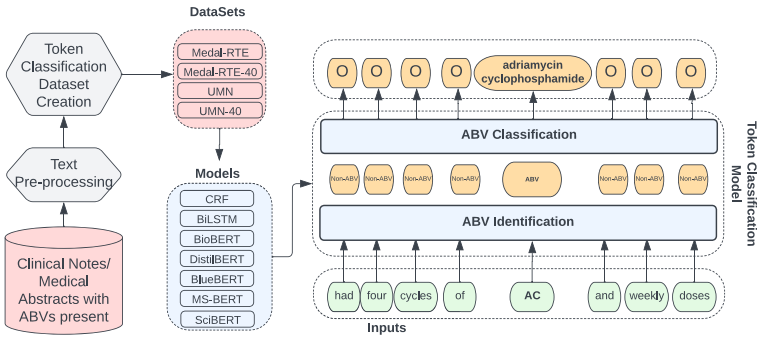
Fig. 6 Top 20 most frequent bi- and tri-grams in the UMN dataset (excluding any n-grams with digits)

DistilBERT, BioBERT, BlueBERT, MSBERT, and SciBERT are employed for the comparative analysis. We select BiLSTM and CRF as two baseline models to test our approach for the full-sized and limited 40-label versions of both datasets. Moreover, we compare the performance of our sequence labeling approach with popular text classification methods for AD.

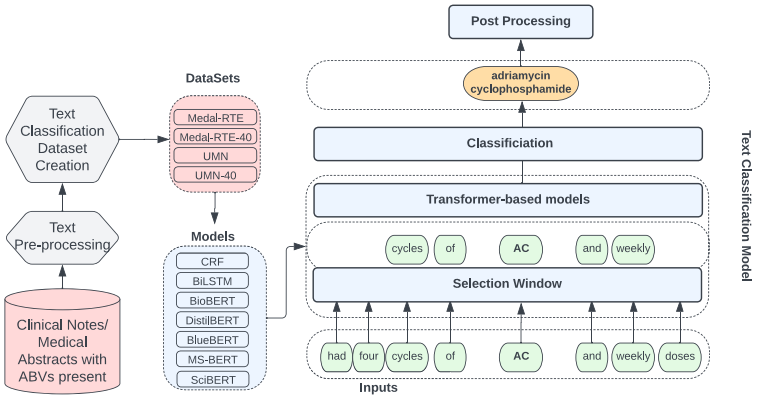
Figure 7a presents a visual representation of the proposed sequence labeling pipeline. According to this flow, each token of a sentence receives either the "other" (O) tag or the ABV sense tag as output labels. Within this context, our problem is more similar to a POS (part-of-speech) task than a NER. This is because multiple consecutive words can be associated with the same tag in NER, while in POS, each word is evaluated separately. Therefore, our model does not utilize the BIO format, which is a special structure developed for named entities. Rather, the proposed model simply produces either O or sense of ABV.

The standard text classification approach for AD is illustrated in Fig. 7b. Text classification models select a window of m words around the target ABV and pass the data to transformers-based models and the classifier. In this study, a window size of 40 is selected through extensive hyperparameter tuning experiments. Finally, a *post-processing* approach is implemented on the results of text classification models. In this approach, the predicted probability output is redistributed over the possible labels of the target ABV, and the label with maximum probability is selected as the output. Note that these possible labels are identified based on the corresponding labels of the ABV in the training data. However, this method is not implemented on sequence labeling raw outputs, since the methodology of the sequence labeling incorporates the specific ABVs for each sample, rendering the postprocessing redundant. Hyperparameter tuning experiments are employed for all the models and datasets, and the final hyperparameters are reported in Table 5.

The employed BiLSTM structure in our analysis is presented in Fig. 8. The model consists of an embedding layer, two dropout layers, two LSTM layers, one of which is Bidirectional, and a final TimeDistributed layer to output a label for every input token. BiLSTM uses the *Adam* optimizer with a learning rate of 0.005, and *sparse categorical cross-entropy* loss function to deal with the non-one-hot encoded labels. Additionally, to decrease the impact of the imbalanced distribution of labels with the high majority



(a) Sequence Labeling classification pipeline



(b) Text classification pipeline with window of size 2

Fig. 7 Flowchart of sequence labeling and text classification methods

Table 5 Selected model hyperparameters for each dataset

Model	Dataset	Hyperparameters
CRF	MeDAL	algorithm : *lbfgs*, c1 : 0.1, c2 : 0.1, max_iterations : 100
	UMN	algorithm : *lbfgs*, *c*1 : 0.1, c2 : 0.1, max_iterations : 100
BiLSTM	MeDAL	dropout_rate : 0.3, optimizer : *Adam*, learning_rate : 5e-3, activation : *softmax*, loss_function : *sparse_categorical_crossentropy*, num_epochs : 30, batch_size : 64
	UMN	dropout_rate : 0.3, optimizer : *Adam*, learning_rate : 5e-3, activation : *softmax*, loss_function : *sparse_categorical_crossentropy*, num_epochs : 30, batch_size : 64
DistilBERT BioBERT BlueBERT	MeDAL	max_sequence_length : 512, num_epochs : 5, learning_rate : 2e-5, weight_decay : 0.01, batch_size = 8
MS-BERT SciBERT	UMN	max_sequence_length : 512, num_epochs : 6, learning_rate : 2e-5, weight_decay : 0.01, batch_size = 8

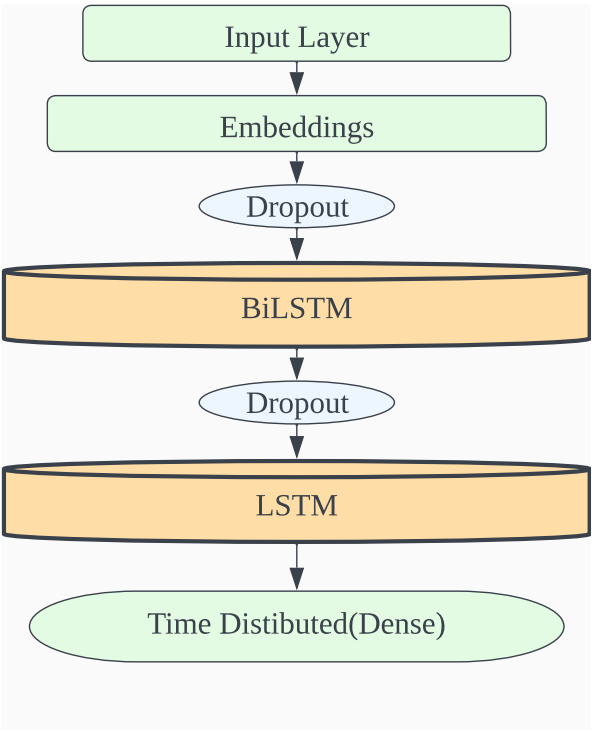


Fig. 8 BiLSTM model architecture

of tokens being assigned to non-ABV label (*NA_word*), we set the weight of all labels to 100 and keep *NA_word* at 1.

The training process for the pre-trained BERT models is as follows. We first tokenize the datasets using the respective BERT Fast Tokenizer and then create training batches using the *DataCollatorForTokenClassification* class which also dynamically pads the sequences in each batch to be the same length. We then fine-tune the model using the *AutoModelForTokenClassification* class from HuggingFace.[1]

The performance of the trained models is evaluated based on the 3-fold cross-validation method. Macro- and weighted-F1 scores are the main performance metrics reported in this study. The macro-averaged F1 score is computed by taking the arithmetic mean of all per-sense F1 scores. This method treats all senses (classes) equally, regardless of their support values. On the other hand, the weighted-average F1 score takes into account the support for each sense when calculating the mean of all per-sense F1 scores. The macro-F1 score calculation includes the *NA_word* label (the label that should be assigned to non-ABV tokens), whereas, for weighted-F1 score, it is excluded due to its abundantly large support and low importance. Macro- and weighted-precision and recall values are also included to provide a more complete performance analysis and compare the models on a deeper level.

[1] https://huggingface.co/

4 Numerical Results

In this section, we first compare the performance of different sequence labeling methods for the AD task. Next, we select the best-performing model from the first part and compare its performance with different text classification methods. Finally, we explore the performances of the BERT-based sequence labeling methods in comparison to the popular CRF approach over the limited-label MeDAL-40 and UMN-40 datasets.

4.1 Comparative Analysis of Sequence Labeling Models

In this experiment, we compare the performance of the BERT-based sequence labeling methods against the BiLSTM model. Experiments are conducted over MeDAL and UMN datasets with more than 1000 and 200 unique labels, respectively. We note that the large size of these datasets closely resembles the computational complexity of real-world scenarios. Table 6 presents the results of the experiments over both datasets, and Fig. 9 shows the box plot of macro-F1 scores as obtained over 3 folds.

We observe that all the BERT-based models outperform the baseline BiLSTM model by a large margin. For the MeDAL dataset, SciBERT is the best-performing model overall, with a macro-F1 score of 77.29%. On the other hand, the results show that all BERT models perform similarly for the UMN dataset. The significant difference between macro and weighted metrics highlights the imbalanced nature of the dataset. In particular, limited samples of specific labels deteriorate the macro metrics of the models. BioBERT and SciBERT for the MeDAL and BlueBERT for the UMN are found to be the best-performing models. The strong performance by BioBERT is expected as it was pre-trained on text data from a similar domain. BlueBERT was pre-trained on similar text as well, though on a more limited basis, and hence achieved much lower macro scores for the MeDAL dataset. BlueBERT's limitations on MeDAL dataset could be explained by the class imbalance issue that affects its performance particularly for macro average metrics.

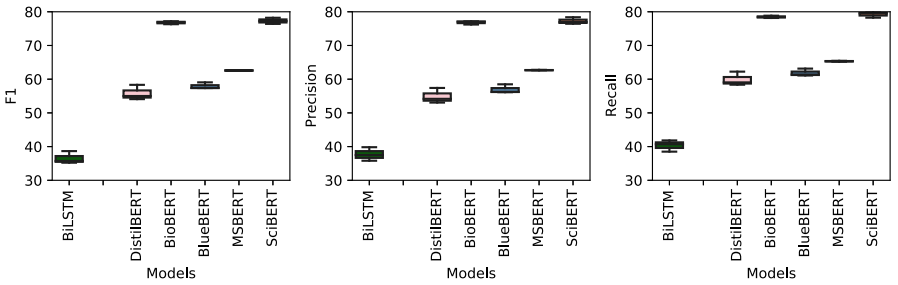
4.2 Comparison Against Text Classification Methods

Table 7 presents the postprocessed performances of different text classification models against the best-performing sequence labeling model from the previous section. In addition, the macro metrics are illustrated in Fig. 10. The impact of postprocessing on the results of the text classification models is thoroughly explored and reported in Sect. 1 of the Appendix.

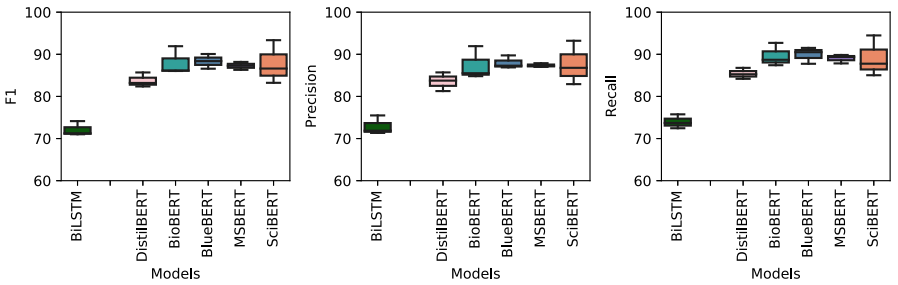
When we closely examine the performance of text classification models, we observe that the models produce similar results for both Medal and UMN data. However, we still see that the BioBERT model gives slightly better results in terms of average performance values. The DistilBERT model has shown a much weaker performance compared to other models, and there may be two reasons for this. Firstly, the DistilBERT model has fewer parameters than other models. Secondly, since it was trained on a general dataset rather than domain-specific data, its knowledge of the scientific field is insufficient compared to other models.

Table 6 Summary performance values for the sequence labeling models on full-label datasets. Results are averaged over 3 folds (highest scores on each metric are in bold)

Dataset	Model	Macro average			Weighted average		
		F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)
MeDAL	BiLSTM	36.57 ± 1.52	37.71 ± 1.65	40.34 ± 1.38	66.80 ± 1.96	67.05 ± 1.56	68.71 ± 2.01
	DistilBERT	55.80 ± 1.82	54.88 ± 1.84	59.88 ± 1.69	78.02 ± 1.69	77.43 ± 1.97	81.12 ± 1.24
	BioBERT	76.79 ± 0.37	76.81 ± 0.42	78.48 ± 0.28	90.29 ± 0.68	90.69 ± 0.89	91.19 ± 0.37
	BlueBERT	57.94 ± 0.79	56.95 ± 1.08	61.87 ± 0.90	79.37 ± 0.86	78.83 ± 1.59	82.29 ± 0.41
	MS-BERT	62.56 ± 0.08	62.69 ± 0.05	65.32 ± 0.11	81.11 ± 0.92	81.57 ± 1.29	83.13 ± 0.67
	SciBERT	**77.29 ± 0.74**	**77.25 ± 0.82**	**79.16 ± 0.65**	**90.53 ± 1.05**	**90.85 ± 1.37**	**91.58 ± 0.74**
UMN	BiLSTM	72.13 ± 1.42	72.90 ± 1.84	73.96 ± 1.36	88.40 ± 0.29	88.10 ± 0.23	89.60 ± 0.38
	DistilBERT	83.75 ± 1.41	83.57 ± 1.81	85.41 ± 1.07	92.53 ± 0.96	92.54 ± 1.46	93.58 ± 0.53
	BioBERT	88.03 ± 2.75	87.42 ± 3.20	89.61 ± 2.26	**95.05 ± 0.83**	94.79 ± 1.32	**95.97 ± 0.30**
	BlueBERT	**88.34 ± 1.42**	**87.98 ± 1.26**	**89.92 ± 1.59**	94.32 ± 0.68	94.30 ± 1.16	95.18 ± 0.46
	MS-BERT	87.26 ± 0.77	87.38 ± 0.30	89.01 ± 0.85	93.87 ± 0.54	94.47 ± 0.78	94.71 ± 0.54
	SciBERT	87.73 ± 4.21	87.64 ± 4.24	89.09 ± 3.97	94.69 ± 1.14	**94.92 ± 1.36**	95.29 ± 0.86



(a) Macro average metrics for MeDAL dataset



(b) Macro average metrics for UMN dataset

Fig. 9 Sequence labeling model performance comparison over full-label datasets

The results for the MeDAL dataset show that the SciBERT sequence labeling model outperforms all the text classification models over macro metrics. However, for the weighted metrics, SciBERT, BioBERT, and BlueBERT text classification models lead to comparable performances. Similarly, for the UMN dataset, the BlueBERT sequence labeling model outperforms all the other models in terms of macro metrics. On the other hand, BioBERT shows higher performance in terms of weighted metrics.

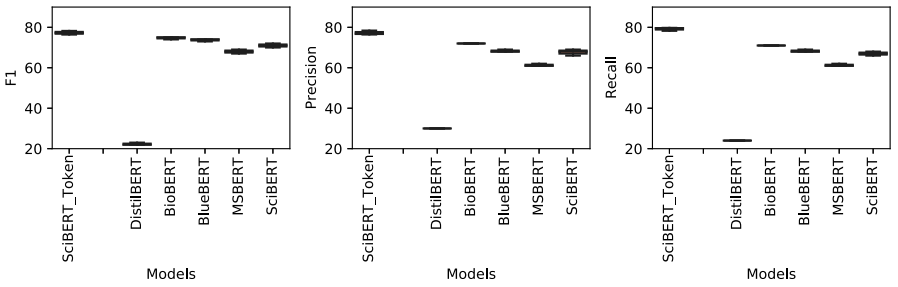
Overall, closely examining the relative performance of both datasets reveals that the best sequence labeling model was able to outperform the majority of the text classification models. That is, that sequence labeling can provide similar or better performance for medical AD tasks while also being able to manage text with multiple unique ABVs. Another important advantage of this model is that it can provide better results even for less-frequent abbreviations. We understand it from the difference between macro and weighted metrics, which shows the impact of the class imbalance for both MeDAL and UMN datasets. Lastly, the sequence labeling method does not require any post-processing, making it easier to implement and adopt in practice.

4.3 Comparison Against Baseline CRF Model

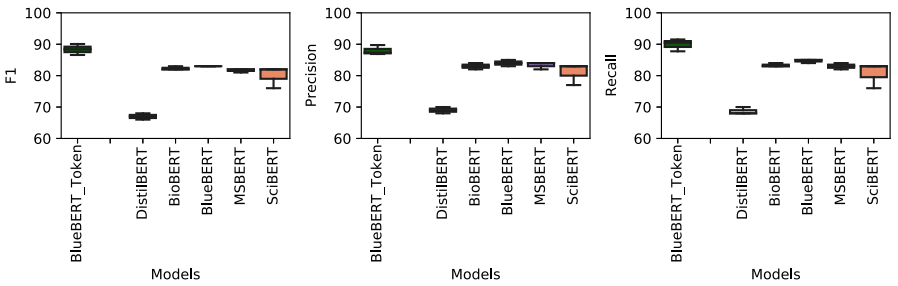
Table 8 presents the performances of different sequence labeling models against the CRF baseline model on a subsampled AD task with a total of 40 unique labels. Results of macro-averaged metrics are also illustrated in Fig. 11.

Table 7 Summary performance values for the text classification models on full-label datasets and comparison against sequence labeling baseline. Results are averaged over 3 folds (highest scores on each metric are in bold)

Dataset	Model	Macro average			Weighted average		
		F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)
				Text Classification Models			
MeDAL	DistilBERT	22.33 ± 0.47	30.0 ± 0.0	24.0 ± 0.0	59.33 ± 0.47	71.67 ± 0.47	63.67 ± 0.47
	BioBERT	74.67 ± 0.47	72.0 ± 0.0	71.0 ± 0.0	**90.67 ± 0.47**	**91.67 ± 0.47**	90.67 ± 0.47
	BlueBERT	73.67 ± 0.47	68.33 ± 0.47	68.33 ± 0.47	88.0 ± 0.0	89.0 ± 0.0	88.0 ± 0.0
	MS-BERT	68.0 ± 0.82	61.33 ± 0.47	61.33 ± 0.47	84.67 ± 0.47	85.67 ± 0.47	84.67 ± 0.47
	SciBERT	71.0 ± 0.82	67.67 ± 1.24	67.0 ± 0.82	89.33 ± 0.47	90.67 ± 0.47	89.67 ± 0.47
				Sequence Labeling Baseline			
	SciBERT	**77.29 ± 0.74**	**77.25 ± 0.82**	**79.16 ± 0.65**	90.53 ± 1.05	90.85 ± 1.37	**91.58 ± 0.74**
				Text Classification Models			
UMN	DistilBERT	67.0 ± 0.82	69.0 ± 0.82	68.67 ± 0.94	89.67 ± 0.47	90.0 ± 0.0	90.67 ± 0.47
	BioBERT	82.33 ± 0.47	83.0 ± 0.82	83.33 ± 0.47	**97.0 ± 0.0**	**97.0 ± 0.0**	**97.33 ± 0.47**
	BlueBERT	83.0 ± 0.0	84.0 ± 0.82	84.67 ± 0.47	96.0 ± 0.0	96.0 ± 0.0	96.0 ± 0.0
	MS-BERT	81.67 ± 0.47	83.33 ± 0.94	83.0 ± 0.82	95.33 ± 0.47	96.0 ± 0.0	96.0 ± 0.0
	SciBERT	80.0 ± 2.83	81.0 ± 2.83	80.67 ± 3.3	95.67 ± 1.89	95.67 ± 1.89	96.0 ± 1.41
				Sequence Labeling Baseline			
	BlueBERT	**88.34 ± 1.42**	**87.98 ± 1.26**	**89.92 ± 1.59**	94.32 ± 0.68	94.30 ± 1.16	95.18 ± 0.46



(a) Macro average metrics for MeDAL dataset



(b) Macro average metrics for UMN dataset

Fig. 10 Text classification model performance comparison over full-label datasets

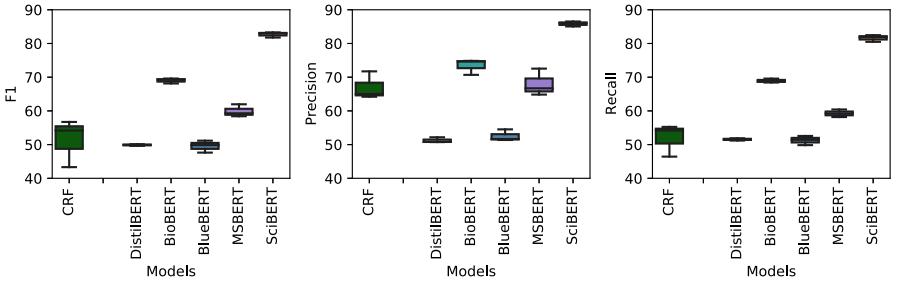
We observe that the majority of BERT-based models except DistilBERT outperform the CRF approach for both datasets. Moreover, SciBERT outperforms all the models for both datasets. Similar to full-size datasets, BioBERT and BlueBERT lead to a similar performance for the UMN dataset. Although the labels are sub-sampled and reduced to 40 most frequent labels, the visible difference between macro and weighted metrics shows that the label distributions in these datasets remain imbalanced. On the other hand, we note that reducing the number of labels in these datasets has led to significant reductions in the training times with the average training time of the models dropping from 40 to 6 min and 16 to 2 min for the MeDAL and UMN datasets, respectively.

5 Discussion and Conclusions

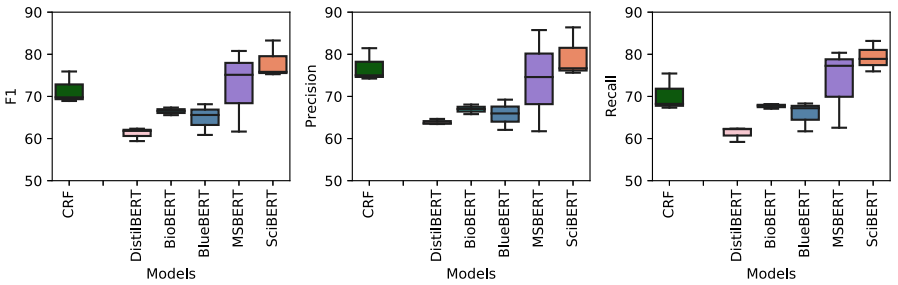
In this study, we investigated the ability of transformers-based text classification and sequence labeling models for the single-token medical AD task. While the majority of recent studies explore text classification methods for the medical AD task, we adopted a sequence labeling approach for this problem where each word (or ABV) is assigned a label.

Table 8 Summary performance values for the sequence labeling models on 40-label datasets. Results are averaged over 3 folds (highest scores on each metric are in bold)

Dataset	Model	Macro average			Weighted average		
		F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)
MeDAL-40	CRF	51.39 ± 5.81	66.98 ± 3.37	51.96 ± 3.93	66.02 ± 11.57	77.16 ± 7.36	67.01 ± 9.78
	DistilBERT	49.88 ± 0.20	51.24 ± 0.66	51.54 ± 0.27	71.84 ± 0.49	71.59 ± 0.56	74.66 ± 0.43
	BioBERT	68.99 ± 0.62	73.40 ± 1.92	68.93 ± 0.48	84.79 ± 0.16	84.63 ± 0.17	85.76 ± 0.20
	BlueBERT	49.54 ± 1.46	52.52 ± 1.42	51.27 ± 1.12	71.71 ± 1.40	71.56 ± 1.61	75.11 ± 1.14
	MS-BERT	59.88 ± 1.51	68.03 ± 3.28	59.27 ± 0.92	76.39 ± 1.05	75.88 ± 1.12	78.54 ± 0.96
	SciBERT	**82.68 ± 0.66**	**85.85 ± 0.61**	**81.61 ± 0.83**	**89.06 ± 0.30**	**88.73 ± 0.42**	**89.85 ± 0.14**
UMN-40	CRF	71.54 ± 3.12	76.90 ± 3.23	70.33 ± 3.63	91.97 ± 0.34	92.22 ± 0.26	92.94 ± 0.31
	DistilBERT	61.19 ± 1.28	63.91 ± 0.54	61.26 ± 1.46	92.02 ± 0.93	91.48 ± 0.84	93.48 ± 0.76
	BioBERT	66.50 ± 0.74	66.96 ± 0.93	67.69 ± 0.46	93.84 ± 0.64	93.24 ± 0.56	94.99 ± 0.60
	BlueBERT	64.87 ± 3.02	65.75 ± 2.93	65.75 ± 2.88	93.51 ± 0.44	92.90 ± 0.34	94.85 ± 0.35
	MS-BERT	72.53 ± 8.02	74.03 ± 9.81	73.40 ± 7.76	94.63 ± 0.50	94.08 ± 0.68	95.56 ± 0.29
	SciBERT	**78.13 ± 3.64**	**79.57 ± 4.84**	**79.35 ± 2.96**	**95.14 ± 0.53**	**95.02± 0.61**	**95.72 ± 0.42**



(a) Macro average metrics for MeDAL-40 dataset



(b) Macro average metrics for UMN-40 dataset

Fig. 11 Sequence labeling model performance comparison over limited-label datasets

The results showed all BERT-based models outperform the BiLSTM and CRF baseline models. Across our two main experiments, we found that BioBERT consistently performed well on both datasets and was notably resistant to data imbalance. Furthermore, its stronger relative performance on both MeDAL dataset variations could be attributed to the fact that it was pre-trained on 18 B words from medical abstracts. Similarly, MS-BERT was among the top performers on the UMN datasets which can also be credited to its relevant pre-training corpora. However, the most notable outcome from these experiments was found in SciBERT results. This model consistently performed the best or close to the best in both experiments and datasets while not being pre-trained on nearly as many medical abstract words as BioBERT, let alone any pre-training on clinical notes like MS-BERT. This success can be attributed to the use of a domain-specific vocabulary, instead of the one created by BERT_{base}, playing a key role in the performance of SciBERT, and should be a consideration in future works. Overall, through our detailed experiments, we were able to examine the efficacy of sequence labeling methods on the medical AD task and demonstrate its benefits over text classification-based approaches. Additionally, our results showed the effect of transfer learning, the significance of relevant pre-training corpora, and the importance of a model's resistance to label imbalance.

The datasets used in our study, while relevant, do not necessarily capture all possible environments where medical ABVs could occur. Accordingly, repeating our experiments with more datasets, ideally, clinical patient notes with multiple unique target ABVs could be helpful in medical AD research; however, we note that suitable public medical datasets are rarely made available. Regarding our chosen solution to this task, using sequence labeling methods on this problem only enhances the class imbalance issue due to the amount of non-ABV entities present in each text. Moreover, limited hyperparameter tuning can pose a threat to validity. Conducting more extensive hyperparameter tuning experiments could help further improve the performance for our medical AD tasks.

In our numerical analysis, we incorporated limited features for the CRF model, which may deteriorate the models' performance. Future work can benefit from a more extensive feature selection procedure. Furthermore, the postprocessing was found to have a significant impact for the text classification model performance. However, exploring and comparing alternative postprocessing methods could prove to be beneficial for this task. Finally, although the motivation of our study was to use sequence labeling methods for the medical AD task, based on the results from the text classification baseline, another possible route could be to combine text classification and sequence labeling methods into one comprehensive model, e.g., using a voting ensemble.

Author Contributions All the co-authors contributed to the conception, design, implementation, writing, and review of the paper. Author order is alphabetical.

Data Availability All the datasets used in our analysis are publicly available, and the links to these datasets are provided as follows: • MeDAL, https://www.kaggle.com/datasets/xhlulu/medal-emnlp; • UMN, https://conservancy.umn.edu/handle/11299/137703

Declarations

Ethical Approval Not applicable

Competing Interests The authors declare no competing interests.

Appendix. Text classification postprocessing results

In this section, we have reported the detailed results for the text classification experiments in Sect. 4.2. Table 9 presents the performance values before and after applying postprocessing. Overall, we observe that almost all the models benefit from postprocessing. In particular, DistilBERT, BlueBERT, and MS-BERT experience a significant performance improvement for the MeDAL dataset. On the other hand, BioBERT and SciBERT models' performances do not benefit from the postprocessing approach on the UMN dataset.

Table 9 Summary performance values for transformers-based text classification models for the full-label datasets with and without postprocessing. Results are averaged over 3 folds (highest scores on each metric are in bold)

Dataset	Model	Raw results						Post-processed results					
		Macro Average			Weighted average			Macro average			Weighted average		
		F1 (%)	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)
MeDAL	DistilBERT	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	22.3 ± 0.5	30.0 ± 0.0	24.0 ± 0.0	59.3 ± 0.5	71.7 ± 0.5	63.7 ± 0.5
	BioBERT	**74.7 ± 0.5**	**72.0 ± 0.0**	**71.0 ± 0.0**	**90.7 ± 0.5**	**91.7± 0.5**	**90.7 ± 0.5**	**74.7 ± 0.5**	**72.0 ± 0.0**	**71.0 ± 0.0**	**90.7 ± 0.5**	**91.7 ± 0.5**	**90.7 ± 0.5**
	BlueBERT	12.3 ± 0.5	11.0 ± 0.0	10.0 ± 0.0	24.7 ± 0.5	29.0 ± 0.0	24.0 ± 0.0	73.7 ± 0.5	68.3 ± 0.5	68.3 ± 0.5	88.0 ± 0.0	89.0 ± 0.0	88.0 ± 0.0
	MS-BERT	4.3 ± 0.5	5.0 ± 0.0	4.0 ± 0.0	11.0 ± 0.0	17.0 ± 0.0	12.0 ± 0.0	68.0 ± 0.8	61.3 ± 0.5	61.3 ± 0.5	84.7 ± 0.5	85.7 ± 0.5	84.7 ± 0.5
	SciBERT	57.0 ± 0.8	53.3 ± 1.3	53.0 ± 0.8	78.7 ± 0.5	80.7 ± 0.5	78.7 ± 0.5	71.0 ± 0.8	67.7 ± 1.2	67.0 ± 0.8	89.3 ± 0.5	90.7 ± 0.5	89.7 ± 0.5
UMN	DistilBERT	14.3 ± 0.5	14.7 ± 0.9	16.7 ± 0.5	32.0 ± 0.8	31.3 ± 1.3	38.0 ± 0.8	67.0 ± 0.8	69.0 ± 0.8	68.7 ± 0.94	89.7 ± 0.5	90.0 ± 0.0	90.7 ± 0.5
	BioBERT	**82.3 ± 0.5**	**83.0 ± 0.8**	**83.3 ± 0.5**	**97.0 ± 0.0**	**97.0 ± 0.0**	**97.3 ± 0.5**	82.3 ± 0.5	83.0 ± 0.8	83.3 ± 0.5	**97.0 ± 0.0**	**97.0 ± 0.0**	**97.3 ± 0.5**
	BlueBERT	44.0 ± 0.0	47.3 ± 1.7	45.0 ± 0.0	67.7 ± 0.5	68.0 ± 0.8	70.0 ± 0.0	**83.0 ± 0.0**	**84.0 ± 0.8**	**84.7 ± 0.5**	96.0 ± 0.0	96.0 ± 0.0	96.0 ± 0.0
	MS-BERT	37.3 ± 0.5	40.3 ± 0.5	38.3 ± 0.5	60.7 ± 0.5	61.0 ± 0.0	63.0 ± 0.0	81.7 ± 0.5	83.3 ± 0.9	83.0 ± 0.8	95.3 ± 0.5	96.0 ± 0.0	96.0 ± 0.0
	SciBERT	79.0 ± 2.8	80.3 ± 2.6	79.0 ± 2.8	94.3 ± 1.9	94.3 ± 1.9	94.7 ± 1.7	80.0 ± 2.8	81.0 ± 2.8	80.7 ± 3.3	95.7 ± 1.9	95.7 ± 1.9	96.0 ± 1.4

References

1. Navigli R (2009) Word sense disambiguation: a survey. ACM Comput Surv (CSUR) 41:1–69
2. Agirre E, Edmonds P (2007) Word sense disambiguation: algorithms and applications, vol. 33. Springer Science & Business Media
3. Abbreviation Definition & Meaning (2022). https://www.merriam-webster.com/dictionary/abbreviation?utm_campaign=sd&utm_medium=serp&utm_source=jsonld#note-2
4. Jaber A, Martínez P (2022) Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques. Methods Inf Med
5. Grossman LV, Mitchell EG, Hripcsak G, Weng C, Vawdrey DK (2018) A method for harmonization of clinical abbreviation and acronym sense inventories. J Biomed Inform 88:62–69
6. McInnes B, Pedersen T, Liu Y, Pakhomov S, Melton GB (2011) Using second-order vectors in a knowledge-based method for acronym disambiguation. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp 145–153
7. Xu H, Stetson PD, Friedman C (2012) Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In: AMIA Annual Symposium Proceedings, vol. 2012. American Medical Informatics Association, p 1004
8. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
9. Pakhomov S, Pedersen T, Chute CG (2005) Abbreviation and acronym disambiguation in clinical discourse. In AMIA Annual Symposium Proceedings, vol. 2005. American Medical Informatics Association, p 589
10. Joshi M, Pakhomov S, Pedersen T, Chute CG (2006) A comparative study of supervised learning as applied to acronym expansion in clinical reports. In AMIA Annual Symposium Proceedings, vol. 2006. American Medical Informatics Association, p 399
11. Moon S, Pakhomov S, Melton GB (2012) Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. In: AMIA Annual Symposium Proceedings, vol. 2012. American Medical Informatics Association, p 1310
12. Jaber A, Martínez P (2021) Disambiguating clinical abbreviations using pre-trained word embeddings. In: HEALTHINF, pp 501–508
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36:1234–1240
14. McCallum A (2012) Efficiently inducing features of conditional random fields. arXiv preprint. arXiv:1212.2504
15. Quinlan JR (2004) Data mining tools see5 and c5. 0. http://www.rulequest.com/see5-info.html
16. Wu Y, Xu J, Zhang Y, Xu H (2015) Clinical abbreviation disambiguation using neural word embeddings. In Proceedings of BioNLP 15, pp 171–176
17. Li I, Yasunaga M, Nuzumlalı MY, Caraballo C, Mahajan S, Krumholz H, Radev D (2019) A neural topic-attention model for medical term abbreviation disambiguation. arXiv preprint. arXiv:1910.14076
18. Johnson AE, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:1–9
19. Jin Q, Liu J, Lu X (2019) Deep contextualized biomedical abbreviation expansion. arXiv preprint. arXiv:1906.03360
20. Doğan RI, Leaman R, Lu Z (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 47:1–10
21. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinform 16:1–17
22. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D et al (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform 16:1–28
23. Wen Z, Lu XH, Reddy S (2020) MeDAL: medical abbreviation disambiguation dataset for natural language understanding pretraining. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, pp 130–135. https://aclanthology.org/2020.clinicalnlp-1.15, https://doi.org/10.18653/v1/2020.clinicalnlp-1.15
24. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). https://doi.org/10.18653/v1/n18-1202
25. Jin Q, Dhingra B, Cohen WW, Lu X (2019) Probing biomedical embeddings from language models. arXiv preprint. arXiv:1904.02181
26. D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, J. Fluck (2005) ProMiner: rule-based protein and gene entity recognition. BMC Bioinform 6:1–9
27. Quimbaya AP, Múnera AS, Rivera RAG, Rodríguez JCD, Velandia OMM, Peña AAG, Labbé C (2016) Named entity recognition over electronic health records through a combined dictionary-based approach. Proc Comput Sci 100:55–61
28. Zhang S, Elhadad N (2013) Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform 46:1088–1098
29. Settles B (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), pp 107–110
30. Yao L, Liu H, Liu Y, Li X, Anwar MW (2015) Biomedical named entity recognition based on deep neutral network. Int J Hybrid Inf Technol 8:279–288
31. Souza F, Nogueira R, Lotufo R (2019) Portuguese named entity recognition using BERT-CRF. arXiv preprint. arXiv:1909.10649
32. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, Wang P (2020) K-BERT: enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 2901–2908
33. Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data
34. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
35. Jozefowicz R, Zaremba W, Sutskever I (2015) An empirical exploration of recurrent network architectures. In: International Conference on Machine Learning. PMLR, pp 2342–2350
36. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint. arXiv:1508.01991
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
38. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint. arXiv:1910.01108
39. Peng Y, Yan S, Lu Z (2019) Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019), pp 58–65
40. MS - BERT (2020). https://huggingface.co/NLP4H/ms_bert
41. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 3615–3620. https://aclanthology.org/D19-1371, https://doi.org/10.18653/v1/D19-1371
42. Moon S, Pakhomov S, Melton G (2012) Clinical abbreviation sense inventory. https://conservancy.umn.edu/handle/11299/137703