Author/s:
Hur, B;Baldwin, T;Verspoor, K;Hardefeldt, L;Gilkerson, J

Title:
Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes

Date:
2020-01-01

Citation:
Hur, B., Baldwin, T., Verspoor, K., Hardefeldt, L. & Gilkerson, J. (2020). Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp.156-166. ASSOC COMPUTATIONAL LINGUISTICS-ACL. https://doi.org/10.18653/v1/2020.bionlp-1.17.

Persistent Link:
https://hdl.handle.net/11343/274182

# Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes

**Brian Hur**[1,2] **Timothy Baldwin**[2] **Karin Verspoor**[2,3]
**Laura Hardefeldt**[1] **James Gilkerson**[1]
[1]Asia-Pacific Centre for Animal Health
[2]School of Computing and Information Systems
[3]Centre for the Digital Transformation of Health
The University of Melbourne, Australia
{b.hur,tbaldwin,karin.verspoor,laura.hardefeldt,jrgilk}@unimelb.edu.au

## Abstract

Identifying the reasons for antibiotic administration in veterinary records is a critical component of understanding antimicrobial usage patterns. This informs antimicrobial stewardship programs designed to fight antimicrobial resistance, a major health crisis affecting both humans and animals in which veterinarians have an important role to play. We propose a document classification approach to determine the reason for administration of a given drug, with particular focus on domain adaptation from one drug to another, and instance selection to minimize annotation effort.

## 1 Introduction

Microorganisms — such as bacteria, fungi, and viruses — were a major cause of death until the discovery of antibiotics (Demain and Sanchez, 2009). However, antimicrobial resistance ("AMR") to these drugs has been detected since their introduction to clinical practice (Rollo et al., 1952), and risen dramatically over the last decade to be considered an emergent global phenomenon and major public health problem (Roca et al., 2015). Companion animals are capable of acquiring and exchanging multidrug-resistant pathogens with humans, and may serve as a reservoir of AMR (Lloyd, 2007; Guardabassi et al., 2004; Allen et al., 2010; Graveland et al., 2010). In addition, AMR is associated with worse animal health and welfare outcomes in veterinary medicine (Duff et al.; Johnston and Lumsden). "Antimicrobial Stewardship" is broadly used to refer to the implementation of a program for responsible antimicrobial usage, and has been demonstrated to be an effective means of reducing AMR in hospital settings (Arda et al., 2007; Pulcini et al., 2014; Baur et al., 2017; Cisneros et al., 2014). A key part of antimicrobial stewardship is having the ability to monitor antimicrobial usage patterns, including which antibiotic is given and the reason — or "indication" — for its use. This data is generally captured within free text clinical records created at the time of consult. The primary objective of this paper is to develop text categorization methods to automatically label clinical records with the indication for antibiotic use.

> History: Examination: Still extremely pruritic. There is no frank blood visible. And does not appear to be overt inflammation of skin inside EAC. Laboratory: Assessment: Much improved but still concnered there might be some residual pain/infection. This may be exac by persistent oilinesss from PMP over the last week. Treatment: Cefovecin 1mg/kg sc Owner will also use advocate; Advised needs to lose weight. To be 7kg Plan: Owner may return to recheck in ten days at completion of cefo duration.

Figure 1: Sample clinical note, in which the indication of use for *cefovecin* would be EAR DISORDER

We perform this research over the VetCompass Australia corpus, a large dataset of veterinary clinical records from over 180 of the 3,222 clinical practices in Australia which contains over 15 million clinical records and 1.3 billion tokens (McGreevy et al., 2017). An example of a typical clinical note is shown in Figure 1. We aim to map the indication for an antimicrobial into a standardized format such as Veterinary Nomenclature (VeNom) (Brodbelt, 2019), and in doing so, facilitate population-scale quantification of antimicrobial usage patterns.

As illustrated in Figure 1, the data is domain specific, and expert annotators are required to label the training data. This motivates the use of



*Proceedings of the BioNLP 2020 workshop*, pages 156–166

approaches to minimize the amount of annotation effort required, with specific interest in adapting models developed for one drug to a novel drug.

Previous analysis of this dataset has focused on labeling the antibiotic associated with each clinical note (Hur et al., 2020). In that study, it was found that *cefovecin* along with *amoxycillin clavulanate* and *cephalexin* were the top 3 antibiotics used. As *cefovecin* was the most commonly used antimicrobial with the most critical significance for the development of AMR, it was targeted for additional studies to understand the specific indications of use. The indication of use was manually labeled in 5,008 records. However, there were still over 79,000 clinical records with instances of *cefovecin* administration that did not have labels, in addition to over 1.1 million other clinical records involving other antimicrobial drug administrations missing labels.

Having only a single type of antimicrobial agent labeled causes challenges for training a model to classify the indication of use for other antimicrobials, as antimicrobials vary in how and why they are used, with the form of drug administration (oral, injected, etc.) and different indications of use creating distinct contexts that can be seen as sub-domains. Therefore, models that allow for the transfer of knowledge between the sub-domains of the various antimicrobials are required to effectively label the indication of use.

To explore the interaction between learning methods and the resource constraints on labeling, we develop models using the complete set of labels we had available, but also models derived using only labels that can be created within two hours, following the paradigm of Garrette and Baldridge (2013).

Specifically, our work explores methods to improve the performance of classifying the indication for an antibiotic administration in veterinary records of dogs and cats. In addition to classifying the indication of use, we explore how data selection can be used to improve the transfer of knowledge derived from labeled data of a single antimicrobial agent to the context of other agents. We also release our code, and select pre-trained models used in this study at: https://github.com/havocy28/VetBERT.

## 2 Related Work

Clinical coding of medical documents has been previously done with a variety of methods (Kiritchenko and Cherry, 2011; Goldstein et al., 2007; Li et al., 2018a). Additionally, classifying diseases and medications in clinical text has been addressed in shared tasks for human texts (Uzuner et al., 2010). Previous methods have also been explored for extracting the antimicrobials used, out of veterinary prescription labels, associated with the clinical records (Hur et al., 2019), and labeling of diseases in veterinary clinical records (Zhang et al., 2019; Nie et al., 2018) as well exploring methods for negation of diseases for addressing false positives (Cheng et al., 2017; Kennedy et al., 2019). Our work expands on this work by linking the indication of use to an antimicrobial being administered for that diagnosis.

Contextualized language models have recently gained much popularity due to their ability to greatly improve the representation of texts with fewer training instances, thereby transferring more efficiently between domains (Devlin et al., 2018; Howard and Ruder, 2018). Pre-training these language models on large amounts of text data specific to a given domain, such as clinical records or biomedical literature, has also been shown to further improve the performance in biomedical domains with unique vocabularies (Alsentzer et al., 2019; Lee et al., 2019). These models can also accomplish many tasks in an unsupervised manner. For example, Radford et al. (2019) showed that free text questions could be fed through a language model and generate the correct answer in many cases. In our experiments, we demonstrate the usefulness of contextualized language models by pre-training BERT on a large set of veterinary clinical records, and further explore its usefulness for domain adaptation through instance selection.

Domain adaptation is a task which has a long history in NLP (Blitzer et al., 2006; Jiang and Zhai, 2007; Agirre and De Lacalle, 2008; Daumé III, 2007). There has been further work demonstrating the usefulness of reducing the covariance between domains through adversarial learning (Li et al., 2018b). More recently, it has been shown that domain adversarial training can be effectively done using contextualized models, such as BERT, through using a two-step domain-discriminative data selection (Ma et al., 2019). We adapt these methods to our task to create a more generalizable

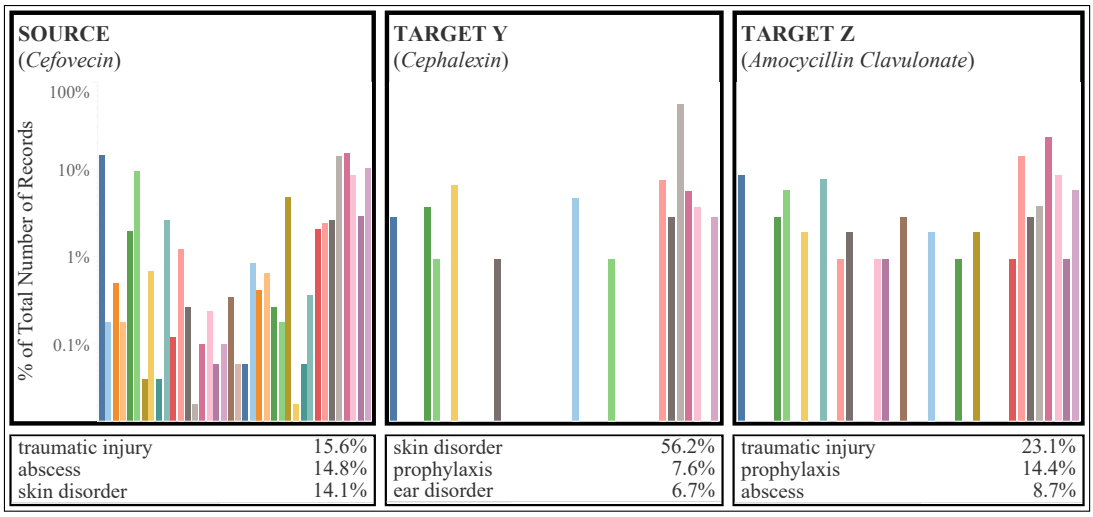


Figure 2: Distribution of labels from the SOURCE and TARGET domains (log scale). The Top-3 labels are noted below each chart.

model that can adapt between domains more effectively.

Previous experiments have used active learning to improve clinical concept extraction with weak supervision (Kholghi et al., 2016). Our work expands on this work through combining approaches to domain adaptation and the effective use of a small number of labels through the development of additional instance selection methods.

## 3 Dataset

### 3.1 Creating a set of terms

Standardized terminologies such as VeNom and SNOMED (NIH, 2019) have been created for medical diagnosis codes. While SNOMED has a veterinary extension, VeNom was created specifically for veterinary clinical text and can be mapped back to SNOMED, and is also part of the Unified Medical Language System (UMLS) (Bodenreider, 2004) used widely within human medicine. Therefore, VeNom codes are used here to create labels for the indication of drug administration (Brodbelt, 2019).

The VeNom codes we adopt are not fully comprehensive; they are a subset of the codes used by (O'Neill et al., 2019) which map specific VeNom codes to more generalized codes. These codes were provided by the Royal College of Veterinary Medicine for this study. In this subset of terms, specific labels such as EXTRACTION OF UPPER LEFT PREMOLAR 4 are simply mapped to DENTAL DISORDER. There were a total of 52 of these terms, of which 38 actually occur in our target dataset.

### 3.2 Data sub-domains

We consider the individual antibiotic agents in our dataset to be sub-domains, as they are administered differently (e.g. orally vs. injectable), and in response to different indications. In our experiments, we target *cefovecin* (injectable), *amoxycillin clavulanate* (oral or injectable), and *cephalexin* (oral). In addition, *cefovecin* and *amoxycillin clavulanate* are used broadly for many indications, while *cephalexin* is primarily used for skin infections.

### 3.3 Extracting and labeling the data

A corpus of 5,008 clinical records, where patients had been given *cefovecin*, were sourced from VetCompass Australia using methods previously described in Hur et al. (2019). The indication of use for *cefovecin* was then labeled by a veterinarian.

A subset of 100 of these annotations were labeled by another veterinarian and used to calculate agreement, which was measured as Cohen's Kappa = 0.78, with raw agreement of 0.80. An additional 105 and 104 records were randomly selected for each of *cephalexin* and *amoxycillin clavulanate*, respectively, and annotated by the same two veterinarians.

The variance between the distribution of indications for *cefovecin*, *cephalexin*, and *amoxycillin clavulanate* is presented in Figure 2.

An additional set of 3000 unannotated clinical notes was sampled, comprising 1000 clinical notes

for each of the three antibiotics of interest. We use these to train a domain classification filter (to identify which antimicrobial is administered), and for data selection. Any notes with fewer than 5 tokens were removed from the corpus.

### 3.4 Training and development sets

The training of the indication-of-use classifier was performed using the dataset pertaining to *cefovecin*, based on a 90:10 split of train and development data. In evaluation, we will refer to the development set as "SOURCE".

The labeled datasets for *amoxycillin clavulanate* and *cephalexin* are used to test cross-domain accuracy, and are referred to as "TARGET Y" for *cephalexin* and "TARGET Z" for *amoxycillin clavulanate*. The test data used for "TARGET Y" and "TARGET Z" were fixed in all tests and strictly disjoint from any training.

The estimated number of records that could be annotated within two hours was 250, based on the annotation of the three datasets. To assess the setting of having only two hours of annotation time, a subset of 250 records was sampled and annotated for for training and taken only from the "SOURCE" data according to one of the various instance selection methods described in the Approach section.

## 4 Approach

In this section we detail our approach, as illustrated in Figure 3.

**Pretraining**

In order to fine-tune our model to veterinary clinical notes, we took ClinicalBERT (Alsentzer et al., 2019) and repeated the pretraining steps as described by Devlin et al. (2018) using the entire corpus of 15 million clinical notes from VetCompass Australia. We refer to this model as "VetBERT".

**Training classifiers**

A baseline classifier for indication of antibiotic administration was trained using an LSTM ("LSTM": Gers et al. (1999)) with a 100 dimension embedding layer with 0.3 dropout, implemented in keras (Chollet et al., 2015). We also use a baseline BERT model using BERT-Base ("BERT"), in addition to a model based on VetBERT. Both the BERT and VetBERT classifiers were trained using an Adam optimizer, maximum of 512 word pieces, batch size of 32, softmax loss, and Learning Rate of 2e-5. Models trained on the full training set were

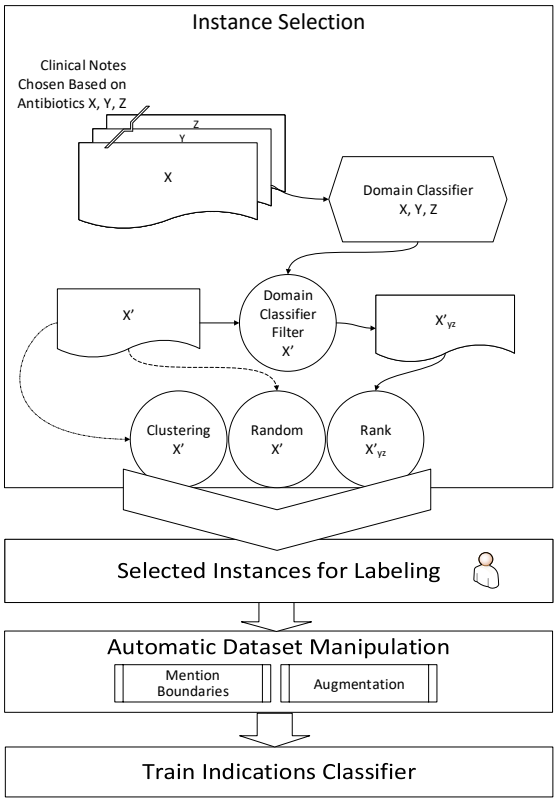


Figure 3: Outline of the proposed approach.

trained for 3 epochs, while models based on limited training data (see Section 4.3) were trained for 60 epochs. All models were tested with 5 different random seeds, and results averaged across them.

Table 1 shows the performance of VetBERT and the two baseline methods both in-domain ("SOURCE"), and for the two out-of-domain antimicrobials using the training data from SOURCE.

While the performance of VetBERT exceeded the interannotator agreement of 78% in-domain, the out-of-domain performance over TARGET Z in particular was substantially less, at 65.4% accuracy. To improve cross-domain performance, we add instance selection and dataset manipulation methods, as described below.

### 4.1 Instance selection

We hypothesize that filtering out training data that is dissimilar to the target domain will improve performance, despite the lower volume of training data. To this end, we experiment with domain-based instance selection.

We model domain similarity using a domain classification model, trained on the domain (i.e. administered antimicrobial) associated with a given medical record. Note that this is directly avail-

| | SOURCE | TARGET Y | TARGET Z |
|---|---|---|---|
| LSTM | 51.5±3.5 | 47.4±3.4 | 29.2±5.0 |
| BERT | 73.4±1.1 | 71.0±1.3 | 58.1±1.4 |
| VetBERT | 80.1±0.7 | 81.5±2.1 | 65.4±1.5 |
| VetBERT+A | 80.9±0.6 | **83.4**±1.5 | 68.1±2.1 |
| VetBERT+M | 80.5±0.6 | 80.2±1.3 | 65.8±2.6 |
| VetBERT+M+A | **81.2**±0.6 | 82.1±1.8 | 66.5±1.7 |
| VetBERT+D | 78.3±0.7 | 79.1±2.6 | 66.7±1.5 |
| VetBERT+D+A | 80.5±0.5 | 83.1±1.4 | **68.5**±2.2 |
| VetBERT+D+M | 78.5±0.8 | 78.3±3.5 | 66.7±0.9 |
| VetBERT+D+M+A | 80.3±0.4 | 82.7±2.1 | 67.5±2.2 |

Table 1: Predictive accuracy (%) of reason for antimicrobial administration in the SOURCE and TARGET domains, trained on all available source-domain training data. Notation: +D = domain-based instance selection, +M = mention boundary tagging, +A = data augmentation

able as an artefact of the dataset construction, and doesn't require any manual annotation. Specifically, we identify instances of source domain X (*cefovecin*) for which we have labeled data, which are most similar to instances from target domains Y and Z, i.e., records in which *cephalexin* or *amoxycillin clavulanate*, respectively, were administered. Determination of similarity is based on the probabilistic output of a domain classifier over the three domains. In Figure 3, we label this subset of the training data "$X'_{YZ}$", reflecting the fact that it is a subset of X similar to Y and Z. This subset of X is then used to train a second classifier focused on the primary task, namely the reason for administering an antibiotic.

To build the domain classification model, we follow the procedure of Ma et al. (2019), first training a domain classifier for 1 epoch, based on the datasets of 1000 instances each of the three domains. We used the same model architecture as the VetBERT model, with a softmax classification layer. This model was then applied to the 5,008 training instances for *cefovecin*, which were sorted in increasing score over domain X (i.e. in decreasing order of similarity to the target domains), the Top-1000, 2000, 3000, or 4000 records were selected, and the VetBERT model was trained over that subset of the training data. The best results were found to occur for 3000 samples. Models with domain-based instance selection are indicated with "+D" in Table 1.

The domain classifier filtering method results in an improvement for TARGET Z (66.7%), but drop in accuracy for TARGET Y (79.1%).

### 4.2 Automated dataset manipulation

We also explore the use of dataset manipulation, in two forms: (1) mention boundary tagging; and (2) data augmentation.

#### 4.2.1 Mention Boundary Tagging

To sensitize the model to the specific drug of interest, we add special learnable embedding vectors to the start and end of each antibiotic mention, based on the findings of Logeswaran et al. (2019) and Wu et al. (2019). Similar to Wu et al. (2019), we used special tokens to mark the boundaries of the tokens that contained a partial string match for the antibiotic of interest. This allows for the model to attend to these tokens at every layer of the network while training the classifier, and ideally better generalize across antimicrobials. The partial string matches were created by identifying strings that contained the prefixes *clav* or *amoxyclav* for *amoxycillin clavulanate*, *ceph*, *rilex* or *kflex* for *cephalexin*, and *conv* or *cefov* for *cefovecin*. These prefixes were sourced from a previous study exploring mention detection of antimicrobials (Hur et al., 2019). We signal the use of mention boundary embeddings with "+M" in the results tables.

#### 4.2.2 Data augmentation

Synonym-based data augmentation has been successfully applied to contexts including word sense disambiguation (Leacock and Chodorow, 1998), sentiment analysis (Li et al., 2017), text classification (Wei and Zou, 2019), and argument analysis (Joshi et al., 2018).

We perform data augmentation on clinical notes by replacing synonyms using WordNet (Fellbaum,

| | SOURCE | TARGET Y | TARGET Z |
|---|---|---|---|
| VetBERT+rank[linear] | 74.3±0.2 | 76.6±3.0 | **66.9**±2.2 |
| VetBERT+rank[linear]+A | 75.8±1.3 | **81.0**±2.6 | 63.7±1.4 |
| VetBERT+rank[linear]+M | 73.4±0.9 | 77.1±1.9 | 65.9±2.4 |
| VetBERT+rank[linear]+M+A | 75.7±0.8 | **81.0**±2.8 | 63.8±3.5 |
| VetBERT+rank[exp] | 68.3±2.1 | 66.5±2.1 | 58.1±1.5 |
| VetBERT+rank[exp]+A | 76.6±0.3 | 76.7±2.4 | 65.4±1.0 |
| VetBERT+rank[exp]+M | 68.9±2.0 | 66.7±1.5 | 57.9±2.1 |
| VetBERT+rank[exp]+M+A | **76.9**±0.2 | 77.3±2.3 | 64.4±1.5 |
| VetBERT+rank[rand] | 73.5±1.8 | 75.4±2.3 | 61.9±2.8 |
| VetBERT+rank[rand]+A | 74.8±1.3 | 78.9±3.1 | 64.2±2.5 |
| VetBERT+rank[rand]+M | 73.9±1.2 | 76.2±2.8 | 62.1±1.1 |
| VetBERT+rank[rand]+M+A | 74.9±0.4 | 80.6±1.3 | 63.1±2.6 |

Table 2: Predictive accuracy (%) of reason for antimicrobial administration over the SOURCE and TARGET domains, trained on 2-hours' worth of labeled data with the three domain similarity selection methods over the top-3000 from $X'_{YZ}$ of random sampling ("+rank[rand]"), modified exponential sampling ("+rank[exp]"), and linear step-wise sampling ("+rank[linear]").

| | SOURCE | TARGET Y | TARGET Z |
|---|---|---|---|
| VetBERT+rand | 70.9±1.5 | 76.2±1.6 | 58.0±2.0 |
| VetBERT+rand+A | 69.7±0.4 | 75.8±1.1 | 59.6±0.0 |
| VetBERT+rand+M | 70.5±0.1 | 77.4±0.6 | 57.4±2.4 |
| VetBERT+rand+M+A | 69.9±0.9 | 77.4±0.6 | 59.6±1.7 |
| VetBERT+rank[linear] | 74.3±0.2 | 76.6±3.0 | 66.9±2.2 |
| VetBERT+rank[linear]+A | **75.8**±1.3 | **81.0**±2.6 | 63.7±1.4 |
| VetBERT+rank[linear]+M | 73.4±0.9 | 77.1±1.9 | 65.9±2.4 |
| VetBERT+rank[linear]+M+A | 75.7±0.8 | **81.0**±2.8 | 63.8±3.5 |
| VetBERT+cluster | 73.4±1.1 | 68.6±1.3 | 63.0±2.1 |
| VetBERT+cluster+A | 73.9±0.1 | 75.2±2.7 | **67.8**±0.7 |
| VetBERT+cluster+M | 73.3±0.5 | 66.2±0.7 | 62.5±1.4 |
| VetBERT+cluster+M+A | 72.8±0.6 | 75.2±0.0 | 63.5±5.4 |

Table 3: Predictive accuracy (%) of reason for antimicrobial administration in the SOURCE and TARGET domains, trained on 2-hours' worth of labeled data, with random selection ("+rand"), linear step-wise sampling ("+rank[linear]"; results duplicated from Table 2), and clustering ("+cluster").

2012), based on random sampling. In this way, we create up to two additional training instances[1] in addition to the original instance, potentially tripling the amount of training data. We signal the use of data augmentation with "+A" in the results tables.

#### 4.2.3 Results for dataset augmentation methods

Mention boundary tagging and data augmentation generally led to improvements in results both in- and out-of-domain, as seen in Table 1. The highest accuracy over the source domain 81.2% was obtained with both mention boundary tagging and data augmentation (without instance selection), while the best out-of-domain results were obtained with data augmentation (with or without instance selection).

### 4.3 Instance selection under two annotation-hour constraint

All results to date have been based on the generous supervision setting of 3000 instances, or ap-

[1] In the instance of there being no synonym substitutes for any words in the original clinical note, no additional training instances are generated.

proximately 24 hours' annotation time. One natural question, inspired by the work of Garrette and Baldridge (2013) in the context of part-of-speech tagging in low-resource languages, is whether similar results can be achieved with a more realistic budget of expert annotation time. Specifically, we assume access to only 2 hours of expert annotator time, which translates to the annotation of 250 clinical notes. We propose three approaches to instance selection under this constraint: (1) domain similarity selection; and (2) clustering. We contrast these with a random selection baseline ("+rand" in our results tables).

#### 4.3.1 Domain similarity selection

Our first approach is based on the instance selection method from Section 4.1, except that we now select only 250 instances from SOURCE for annotation, based on their similarity with the target domain (as distinct from the top-3000 instances in Table 1). That is, we take the top-3000 instances from $X'_{YZ}$ and perform additional sub-sampling, in the form of: (a) random sampling ("+rank[rand]");[2] (b) modified exponential sampling ("+rank[exp]"); or (c) linear step-wise sampling ("+rank[linear]").

Modified exponential sampling is implemented by mapping 3000 onto an exponential scale of 250 steps over the 3000 results, rounding to the nearest integer, and additionally rounding up in the case that there is a collision with a value earlier in the series. That is, instead of the (rounded) series being $0, 0, 0, ..., 2879, 2938, 2999$ it becomes $0, 1, 2, ..., 2879, 2938, 2999$.

Linear step-wise sampling involves separating the domain space evenly, and taking every $n$th sample where $n = \lfloor \text{len}(N)/x \rfloor$ where $x$ is the number of labeled instances (= 250) and $N$ is the total number of samples (= 3000).

Results for the different instance selection methods are presented in Table 2. The best-performing method is step-wise sampling, achieving out-of-domain accuracy which is competitive with the results from Table 1 over 12 times the amount of training data.

#### 4.3.2 Clustering-based instance selection

Our second approach is based on the intuition that the diversity in the training data will optimize performance. We achieve this by clustering the source domain instances, and selecting prototypical instances from each cluster.

First, we generate a representation of each source-domain clinical note using the pretrained `VetBERT` model, based on the `[CLS]` token in the second-last layer of the model. Next, we cluster the instances into 250 clusters using $k$-means++ (Arthur and Vassilvitskii, 2006), and select the instance closest to the centroid for each cluster. This method is labeled "+cluster" in Table 3.

Clustering results in the highest accuracy for TARGET Z of 67.8%, but weaker results for TARGET Y.

## 5 Discussion

### 5.1 Pretraining Improvements

Pretraining `BERT` to the veterinary domain using the VetCompass Australia corpus showed the most dramatic improvement in our experiments. This was demonstrated by marked improvement over other baselines, without any additional steps (Table 1: `VetBERT` vs. `BERT` and `LSTM`). However, even with the pretraining used to create `VetBERT`, there was significant degradation in performance across the domains where there were fewer training instances (`VetBERT` in Table 1 vs. `VetBERT+rand` in Table 3).

### 5.2 Sub-domain transfer performance

The relative performance over TARGET Z as compared to TARGET Y when transferring from SOURCE was generally poor (Tables 1 and 3). This could be due to TARGET Y sharing more similarities with SOURCE, along with the more skewed class distribution in TARGET Y (Figure 2), potentially making it an easier classification task. More analysis is needed to understand this effect.

### 5.3 Optimizing for two hours of annotation time

When optimizing for two hours of annotation time, there were consistent improvements with the instance selection methods, compared to random selectin (Table 3: `VetBERT+rand` vs. others).

### 5.4 Dataset manipulation methods

The results for data augmentation and the addition of mention boundary embeddings were not as clear, in that they sometimes resulted in improvements and sometimes did not (Table 2 and 3: `+A` and `+M` vs. others). The clustering

[2] Note that this differs from +rand in that it is over the top-3000 instances, whereas +rand is over all 5008 annotated instances.

method generally performed better with data augmentation and worst with mention boundary embeddings (Table 3: `VetBERT+cluster+A` vs. `VetBERT+cluster+M+A` and `VetBERT+cluster+M`).

### 5.5 Limitations

The primary limiting factor was also the motivation of this study, namely the difficulty in obtaining sufficient high-quality annotations to perform accurate analysis of the model performance. We were also limited in that the instance selection was performed retrospectively over the 5008 annotated instances, and we were limited to the instances provided for the SOURCE domain, rather than a larger sample that could be obtained from VetCompass. There are also additional domains of data within this corpus that should be evaluated, such as records from specialty practices vs. records from general practices. This was shown to result in significant degradation of performance by Nie et al. (2018), and is a potential area for future research.

## 6 Conclusions and future work

In conclusion, we proposed a range of methods to transfer knowledge derived from labeled data for one antimicrobial agent to other agents, considering the additional constraint of a limited annotation resource time of two hours. While the in-domain accuracy of 83% exceeds the raw inter-annotator agreement of 80% (Cohen's Kappa = 0.78) on the source domain, transfer to other classes is still substantially lower with an average of 76% between the two classes. This shows that while the accuracy on classifying diseases is on par with human classifications for a single disease, there is still room for improvement on transferability to new data subdomains.

The primary question is whether the labels created are good enough to report the reason for antibiotic administration in epidemiological reporting and antimicrobial stewardship guidelines. While the labels for why *cefovecin* was administered were better than the current standard of using expert annotations, our results indicate that accuracy varies substantially depending on the antibiotic being administered, and testing of the accuracy for each individual antibiotic should be evaluated prior to reporting the results based on labels generated by any model.

In future research, these methods could be improved through utilization of available resources such as UMLS or Drugbank to identify clinical use guidelines for antibiotics, to allow for training or adapting a model with few or no annotations. Additionally, further work is required to apply these models into a data pipeline to create labels for VetCompass data to enable analysis of the key reasons for antimicrobial administration in veterinary hospitals across Australia.

## Acknowledgments

We thank Simon Sŭster, Afshin Rahimi, and the anonymous reviewers for their insightful comments and valuable suggestions.

This research was undertaken with the assistance of information and other resources from the VetCompass Australia consortium under the project "VetCompass Australia: Big Data and Real-time Surveillance for Veterinary Science", which is supported by the Australian Government through the Australian Research Council LIEF scheme (LE160100026).

## References

Eneko Agirre and Oier Lopez De Lacalle. 2008. On robustness and domain adaptation using SVD for word sense disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 17–24.

Heather K. Allen, Justin Donato, Helena Huimi Wang, Karen A. Cloud-Hansen, Julian Davies, and Jo Handelsman. 2010. Call of the wild: antibiotic resistance genes in natural environments. *Nature Reviews Microbiology*, 8(4):251–259.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings.

Bilgin Arda, Oguz Resat Sipahi, Tansu Yamazhan, Meltem Tasbakan, Husnu Pullukcu, Alper Tunger, Cagri Buke, and Sercan Ulusoy. 2007. Short-term effect of antibiotic control policy on the usage patterns and cost of antimicrobials, mortality, nosocomial infection rates and antibacterial resistance. *Journal of Infection*, 55(1):41–48.

David Arthur and Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

David Baur, Beryl Primrose Gladstone, Francesco Burkert, Elena Carrara, Federico Foschi, Stefanie Döbele, and Evelina Tacconelli. 2017. Effect of antibiotic stewardship on the incidence of infection and

colonisation with antibiotic-resistant bacteria and Clostridium difficile infection: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, 17(9):990–1001.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

David Brodbelt. 2019. VeNom Coding – VeNom Coding Group. http://venomcoding.org/.

Katherine Cheng, Timothy Baldwin, and Karin Verspoor. 2017. Automatic Negation and Speculation Detection in Veterinary Clinical Text. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 70–78, Brisbane, Australia.

François Chollet et al. 2015. Keras. https://keras.io.

J. M. Cisneros, O. Neth, M. V. Gil-Navarro, J. A. Lepe, F. Jiménez-Parrilla, E. Cordero, M. J. Rodríguez-Hernández, R. Amaya-Villar, J. Cano, A. Gutiérrez-Pizarraya, E. García-Cabrera, and J. Molina. 2014. Global impact of an educational antimicrobial stewardship programme on prescribing practice in a tertiary hospital centre. *Clinical Microbiology and Infection*, 20(1):82–88.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Arnold L. Demain and Sergio Sanchez. 2009. Microbial drug discovery: 80 years of progress. *The Journal of Antibiotics*, 62(1):5–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

A Duff, SK Keane, and Laura Y. Hardefeldt. Descriptive study of antimicrobial susceptibilty patterns from equine septic synovial structures. In *Proceedings of the 39th Bain Fallon Memorial Lectures*, volume 2017:11. Equine Veterinarians Australia.

Christiane Fellbaum. 2012. WordNet. *The Encyclopedia of Applied Linguistics*.

Dan Garrette and Jason Baldridge. 2013. Learning a Part-of-Speech Tagger from Two Hours of Annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. In *Proceedings of the 9th International Conference on Artificial Neural Networks: ICANN '99*, pages 850–855.

Ira Goldstein, Anna Arzumtsyan, and Özlem Uzuner. 2007. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2007, page 279. American Medical Informatics Association.

Haitske Graveland, Jaap A. Wagenaar, Hans Heesterbeek, Dik Mevius, Engeline van Duijkeren, and Dick Heederik. 2010. Methicillin Resistant Staphylococcus aureus ST398 in Veal Calf Farming: Human MRSA Carriage Related with Animal Antimicrobial Usage and Farm Hygiene. *PLOS ONE*, 5(6):e10990.

Luca Guardabassi, Stefan Schwarz, and David H. Lloyd. 2004. Pet animals as reservoirs of antimicrobial-resistant bacteria. *Journal of Antimicrobial Chemotherapy*, 54(2):321–332.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia.

B. Hur, L. Y. Hardefeldt, K. Verspoor, T. Baldwin, and J. R. Gilkerson. 2019. Using natural language processing and VetCompass to understand antimicrobial usage patterns in Australia. *Australian Veterinary Journal*, 97(8):298–300.

Brian A. Hur, Laura Y. Hardefeldt, Karin M. Verspoor, Timothy Baldwin, and James R. Gilkerson. 2020. Describing the antimicrobial usage patterns of companion animal veterinary practices; free text analysis of more than 4.4 million consultation records. *PLOS ONE*, 15(3):1–15.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271.

GCA Johnston and JM Lumsden. Antimicrobial susceptiblity of bacterial isolates from 27 thoroughbreds with arytenoid chondropathy. In *Proceedings of the 39th Bain Fallon Memorial Lectures*, volume 2017:11. Equine Veterinarians Australia.

Anirudh Joshi, Timothy Baldwin, Richard O Sinnott, and Cecile Paris. 2018. UniMelb at SemEval-2018 task 12: Generative implication using LSTMs, Siamese networks and semantic representations with synonym fuzzing. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1124–1128.

Noel Kennedy, Dave C Brodbelt, David B Church, and Dan G O'Neill. 2019. Detecting false-positive disease references in veterinary clinical notes without manual annotations. *NPJ Digital Medicine*, 2(1):1–7.

Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2016. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, 23(2):289–296.

Svetlana Kiritchenko and Colin Cherry. 2011. Lexically-triggered hidden markov models for clinical document coding. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 742–751.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv:1901.08746 [cs]*. ArXiv: 1901.08746.

Min Li, Zhihui Fei, Min Zeng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang. 2018a. Automated icd-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4):1193–1202.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018b. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the EACL (EACL 2017)*, pages 21–27, Valencia, Spain.

David H. Lloyd. 2007. Reservoirs of antimicrobial resistance in pet animals. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 45 Suppl 2:S148–152.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. *arXiv:1906.07348 [cs]*. ArXiv: 1906.07348.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China.

Paul McGreevy, Peter Thomson, Navneet K Dhand, David Raubenheimer, Sophie Masters, Caroline S Mansfield, Timothy Baldwin, Ricardo J Soares Magalhaes, Jacquie Rand, Peter Hill, Anne Peaston, James Gilkerson, Martin Combs, Shane Raidal, Peter Irwin, Peter Irons, Richard Squires, David Brodbelt, and Jeremy Hammond. 2017. VetCompass Australia: A National Big Data Collection System for Veterinary Science. page 15.

Allen Nie, Ashley Zehnder, Rodney L. Page, Arturo L. Pineda, Manuel A. Rivas, Carlos D. Bustamante, and James Zou. 2018. DeepTag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain. *arXiv:1806.10722 [cs]*. ArXiv: 1806.10722.

NIH. 2019. SNOMED CT.

Dan G. O'Neill, Alison M. Skipper, Jade Kadhim, David B. Church, Dave C. Brodbelt, and Rowena M. A. Packer. 2019. Disorders of Bulldogs under primary veterinary care in the UK in 2013. *PLOS ONE*, 14(6):e0217928.

C. Pulcini, E. Botelho-Nevers, O. J. Dyar, and S. Harbarth. 2014. The impact of infectious disease specialists on antibiotic prescribing in hospitals. *Clinical Microbiology and Infection*, 20(10):963–972.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

I. Roca, M. Akova, F. Baquero, J. Carlet, M. Cavaleri, S. Coenen, J. Cohen, D. Findlay, I. Gyssens, O. E. Heure, G. Kahlmeter, H. Kruse, R. Laxminarayan, E. Liébana, L. López-Cerero, A. MacGowan, M. Martins, J. Rodríguez-Baño, J. M. Rolain, C. Segovia, B. Sigauque, E. Tacconelli, E. Wellington, and J. Vila. 2015. The global threat of antimicrobial resistance: science for intervention. *New Microbes and New Infections*, 6:22–29.

I. M. Rollo, J. Williamson, and R. L. Plackett. 1952. Acquired Resistance To Penicillin And To Neoarsphenamine In Spirochaeta Recurrentis. *British Journal of Pharmacology and Chemotherapy*, 7(1):33–41.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388, Hong Kong, China.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Zero-shot Entity Linking with Dense Entity Retrieval. *arXiv:1911.03814 [cs]*. ArXiv: 1911.03814.

Yuhui Zhang, Allen Nie, Ashley Zehnder, Rodney L. Page, and James Zou. 2019. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. *NPJ Digital Medicine*, 2(1).