frontiers | Frontiers in Veterinary Science

TYPE Perspective
PUBLISHED 22 August 2024
DOI 10.3389/fvets.2024.1352726

OPEN ACCESS

EDITED BY
Moh A. Alkhamis,
Kuwait University, Kuwait

REVIEWED BY
Heinzpeter Schwermer,
Federal Food Safety and Veterinary Office (FSVO), Switzerland
Severino Segato,
University of Padua, Italy

*CORRESPONDENCE
P.-J. M. Noble
✉ rtnorle@liverpool.ac.uk

†PRESENT ADDRESSES
Heather Davies,
Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, ON, Canada
Ghada Alfattni,
Department of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia
Mercedes Arguello Casteleiro,
Electronics and Computer Science, University of Southampton, Southampton,
United Kingdom

RECEIVED 08 December 2023
ACCEPTED 17 July 2024
PUBLISHED 22 August 2024

CITATION
Davies H, Nenadic G, Alfattni G, Arguello Casteleiro M, Al Moubayed N, Farrell S, Radford AD and Noble P-JM (2024) Text mining for disease surveillance in veterinary clinical data: part two, training computers to identify features in clinical text.
Front. Vet. Sci. 11:1352726.
doi: 10.3389/fvets.2024.1352726

COPYRIGHT
© 2024 Davies, Nenadic, Alfattni, Arguello Casteleiro, Al Moubayed, Farrell, Radford and Noble. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Text mining for disease surveillance in veterinary clinical data: part two, training computers to identify features in clinical text

Heather Davies[1†], Goran Nenadic[2], Ghada Alfattni[2†], Mercedes Arguello Casteleiro[2†], Noura Al Moubayed[3], Sean Farrell[3], Alan D. Radford[1] and P.-J. M. Noble[1*]

[1]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, United Kingdom, [2]Department of Computer Science, Manchester University, Manchester, United Kingdom, [3]Department of Computer Science, Durham University, Durham, United Kingdom

In part two of this mini-series, we evaluate the range of machine-learning tools now available for application to veterinary clinical text-mining. These tools will be vital to automate extraction of information from large datasets of veterinary clinical narratives curated by projects such as the Small Animal Veterinary Surveillance Network (SAVSNET) and VetCompass, where volumes of millions of records preclude reading records and the complexities of clinical notes limit usefulness of more "traditional" text-mining approaches. We discuss the application of various machine learning techniques ranging from simple models for identifying words and phrases with similar meanings to expand lexicons for keyword searching, to the use of more complex language models. Specifically, we describe the use of language models for record annotation, unsupervised approaches for identifying topics within large datasets, and discuss more recent developments in the area of generative models (such as ChatGPT). As these models become increasingly complex it is pertinent that researchers and clinicians work together to ensure that the outputs of these models are explainable in order to instill confidence in any conclusions drawn from them.

KEYWORDS

big data, machine learning, neural language modeling, clinical records, companion animals

1 Introduction

Natural Language Processing (NLP) is a rapidly growing branch of machine learning aimed at understanding unstructured text using computational methodologies. NLP provides frameworks for computer systems to understand, interpret, and generate human language, which has important implications for applications such as machine translation, text classification, named entity recognition and summarization (1). NLP involves the development of algorithms and computational models to help achieve these challenging goals. However, language does not always follow defined rules but is complex, ambiguous and context-dependent with complications that include regional and dialect differences, emergence of new words, phrases, abbreviations, and colloquialisms.

Within healthcare, NLP is increasingly used to analyze and extract useful information from large volumes of unstructured clinical text data such as electronic health records (EHRs) and clinical notes in an automated process, enhancing the speed and efficiency of clinical decision-making, supporting the detection of disease outbreaks and the ongoing monitoring of disease incidence and prevalence. One study used EHRs from nine hospitals to support an understanding of the outbreak's transmission routes (2). Another area in which NLP has been applied is pharmacovigilance, to capture adverse drug events being reported within clinical narratives to gauge the prevalence and severity and can support the understanding of multi-drug interactions (3, 4).

With the availability of large volumes of veterinary EHRs through frameworks such as SAVSNET and VetCompass, there is a growing need to develop tools and methodologies to fully utilize the rich source of disease information that lies within them (5, 6). As we will describe, machine learning models have the potential to reveal disease syndromic signals within complex textual inputs and have become increasingly accessible even to researchers on modest budgets. This democratization of access to machine learning methods with the attendant potential to screen clinical records at scale has the potential to enhance our understanding of disease patterns in veterinary medicine profoundly.

In the second part of this mini-series, we will discuss the applications and potentials of machine learning methodology to extract valuable insights from unstructured clinical records. We explore how such tools are the building blocks for improving the capabilities of downstream applications such as disease epidemiology and outbreak surveillance. We examine the role of language models, such as bidirectional encoder representations through transformers (BERT) (7) and generative pretrained transformers such as chatGPT (8), Llama (9) (and there are now many more of these to choose from), to extract word meanings to understand the nuances of language and spelling variations within the corpora to better adapt fixed rule-based systems before evaluating them as independent classification tools. By providing an overview of the field's current state, we aim to highlight the pivotal role of machine learning-based text mining in enhancing companion animal care and disease surveillance in veterinary medicine.

2 Text-mining veterinary clinical notes using machine learning

2.1 Machine-learning for word meaning

Machine-learning (ML) is becoming increasingly important for text analysis. In many cases ML relies on neural networks. These are computational representations or software simulations of biological neural networks wherein virtual neurons (or nodes) in multiple layers, are interconnected. Each node aggregates the value of connections from nodes in the layer above, the mathematical weighting of these connections adjusts how much any given connection contributes to the activation of a node. In the simplest neural network these weightings are adjusted by evaluating training data over many iterations and at each iteration adjusting these weightings until the input to the network leads to the "correct" output. The number of weightings (connections) is sometimes referred to as the number of parameters (10).

In part one of this mini-series we discussed the utility of keyword searches for identifying features of EHRs. Veterinary free-text invariably includes both technical scientific and colloquial language, including non-standard abbreviations and misspellings. As it is difficult to curate a complete list of the different ways in which veterinary professionals will describe the same observation, dictionary development can be time consuming and the use of standardized ontologies may result in a loss of recall. Furthermore, even the most complete dictionaries will require updating due to the ever changing nature of nature language. However, there are some simple machine learning approaches which can be implemented in order to augment these dictionaries. An example of such approaches is word embeddings.

Word embedding involves creating vector representations of words (coordinates for a word in a multidimensional word-space) which encapsulate word meaning, therefore allowing mathematical analysis of text. An efficient method for creating embeddings, word2vec, was developed by Mikolov et al., training a neural network to predict words in sentences based on the words surrounding them using a large corpus (hundreds of thousands to millions) of documents (11). The resultant neural network weightings are effectively a vector (usually 200–300 numbers) representing the words embedding. Words, spellings and abbreviations with similar meanings can be identified due to the mathematical similarity of the vectors representing them. A similar procedure can be used to "embed" sentences and passages of text giving these numerical representations of their overall meaning (12).

This approach has been used to expand a dictionary of dietary supplements as described in clinical notes (13), with the model identifying between one and 12 variants for each supplement, resulting in retrieval of 8.39% more clinical notes on average. Similarly, word2vec was used to identify misspellings of pharmaceutical words in clinical notes (14) and resulted in identification of 150 new terms which were used to create an extended lexicon.

The specific advantage of this approach is that the model is trained on data taken from the target corpus, allowing development of embeddings more representative of the language used within that corpus. However, word2vec models do not capture remote relationships between words and attribute a single vector for words that might have multiple meanings. This limitation is avoided in later approaches such as ELMO (15) and transformer architecture (as described below).

2.2 Language models as tools for record annotation

Language models can accurately capture semantic and syntactic structures, which is critical to leverage the rich sources of information that unstructured clinical narratives have within them. Such language models permit flexibility in understanding by capturing patterns and relationships across large volumes of data rather than pre-defined rules; the dynamic nature of language requires an equally malleable system to capture the many ways clinicians can articulate their notes. Rule-based systems are limited to defined inputs where subtle variants in language can add significant complexities to their design and, therefore, can only practically be used for searching for one item at a time such as for single disease scope studies. Rule-based systems also rely heavily on the developers' domain-specific knowledge and manual readings of the records to produce. Neural network-based architectures present in language models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), were a significant innovation away from statistical and rule-based frameworks and introduced the concept of word embeddings, a movement toward capturing rich contextual word representations (16).

Recently, the transformer architecture, which capitalizes on the concept of attention (the relationship between words in phrases/sentences sometimes separated by some distance), has enabled new state-of-the-art performances across many NLP tasks (17). Transformers are the core architecture behind Bidirectional Encoder Representation for Transformers (BERT) (7), Generative Pretraining (GPT) (18), and Language Models for Dialog Applications (LaMDA) (19). A key difference to word2vec embeddings is that these models allow for context-specific representations of words to allow different meanings to be coded differently. For example the word "discharge" may mean fluid leaking from a wound or releasing a patient from hospital and would have the same embedding in word2vec models but is ultimately treated as different entities in BERT models depending on context (7).

Disease coding frameworks, such as the International Classification of Disease (ICD), provide a robust methodology for understanding mortality and morbidity information for research and epidemiology (20). However, disease coding of unstructured clinical notes is challenging and is inherently time-consuming, expensive, and prone to errors (21–24). For these reasons, the concept of automating such a process with language models has been well-explored, with previous research exploring the application of RNNs, CNNS and, more recently, the incorporation of attention mechanisms and the transformer architectures (25–29). The desire for automated disease annotation frameworks to exist within veterinary medicine is no different. Previous works have aimed to apply SNOMED-CT diagnosis labels using bidirectional long-short-term memory networks (BLSTMs), a variant of RNNs, using 112,558 expert annotations from a tertiary referral centre showing promising results (30). Further works capitalized from this integrating a transformer architecture allowing for a hierarchical organization of automated disease codings (31). Language models have also been used to understand veterinarians reasoning behind an antimicrobial administration; here, a BERT model was additionally trained on 15 million clinical notes from the VetCompass Australia corpus (32).

In summary, language models present promise for automating annotations within unstructured EHRs. Their capacity to analyze extensive datasets and discern intricate linguistic relationships empowers these models to enhance annotation precision and speed, facilitating more comprehensive data analysis in disease epidemiology. An exemplification of this potential lies in the development of PetBERT, a substantial language model trained on a corpus exceeding 500 million tokens sourced from the SAVSNET dataset (33). This dataset comprises clinical narratives from diverse veterinary practices across the UK. Through fine-tuning, PetBERT was transformed into a multi-label classifier proficient in automatically coding veterinary clinical EHRs using the International Classification of Diseases 11 framework. Impressively, it achieved F1 scores surpassing 83% across 20 disease codings with minimal annotation requirements. Moreover, they serve as foundational structures for bolstering disease outbreak detection capabilities. Employing this syndromic labeling system, we identified a documented disease outbreak. Comparative analysis between PetBERT's automated identification and the previously employed clinician-assigned point-of-care labeling strategy revealed PetBERT's capability to identify the outbreak up to three weeks earlier. The demonstrated proficiency of PetBERT in automating coding processes within veterinary clinical narratives underscores the transformative potential of language models in augmenting disease surveillance and timely outbreak detection within veterinary medicine.

2.3 Unsupervised machine learning

The machine-learning approaches described above often rely on identifying groups of records for study and in the case of using neural language models will often entail training models using gold-standard annotations made by experts. These are often referred to as 'supervised' systems where the researcher/developer directs what the system learns. Unsupervised systems involve presenting a dataset to a model without stipulating an underlying structure to be detected and to identify the underlying structure within the data in the hope that this will expose relevant clinical features. One such approach is topic modeling. A key approach to this was published by Blei et al. (34) using a statistical approach to text (latent dirichlet allocation) which modeled the assumption that documents contain a distribution of topics and that topics are made up of a specific distribution of words. Reverse engineering this allowed identification of the word distributions present in a set of topics which were inferred from the text itself rather than a prior assumption of what topics were present in the data. This method is made readily available through accessible programming interfaces (35). This approach has been applied to veterinary clinical data allowing clinically relevant topics to be identified in veterinary text to the extent that a topic identified through this method displayed a clear temporal pattern matching a known national outbreak of gastroenteric disease in dogs (36). Topic modeling has been recognized as valuable tool for bioinformatic research (37) and as a tool for clinical research (38, 39).A key feature of topic modeling is that topics discovered in documents are easily interpreted due to each topic having a list of words (and probabilities for those words) that make them up, such that in the example above the outbreak identified above could be clearly seen to be gastroenteric disease given that it comprised words like "diarrhoea," "vomit," and "food." Furthermore, topic-modeling methods allow for evaluation of the weighting of words that comprise the topic with time (or across other categories such as breed, age, and date) highlighting evolution of themes with time that, in the case of disease phenotype, might illustrate emergence of new syndromes (36, 40). More recently, transformer-based models that create whole document embeddings i.e., representation of clinical notes as 768 dimensional arrays based on the meaning of words/tokens in the documents provide another route to topic modeling. Following a reduction of dimensionality in these arrays, a clustering algorithm can be used to cluster documents. An analysis of term-frequency-inverse document frequency (tf-idf) identifies key-words reflected in the these clusters. These words then indicate themes or topics in those documents. This approach is encapsulated in the BERTopic

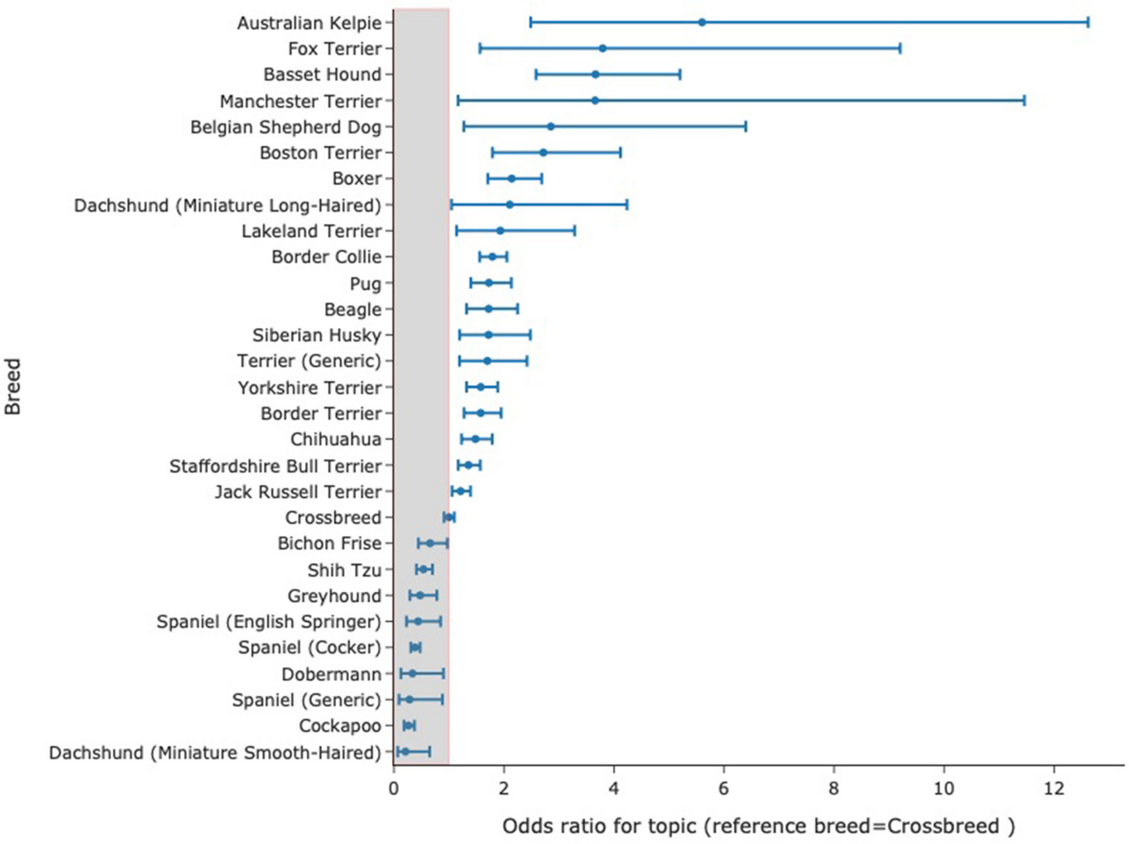
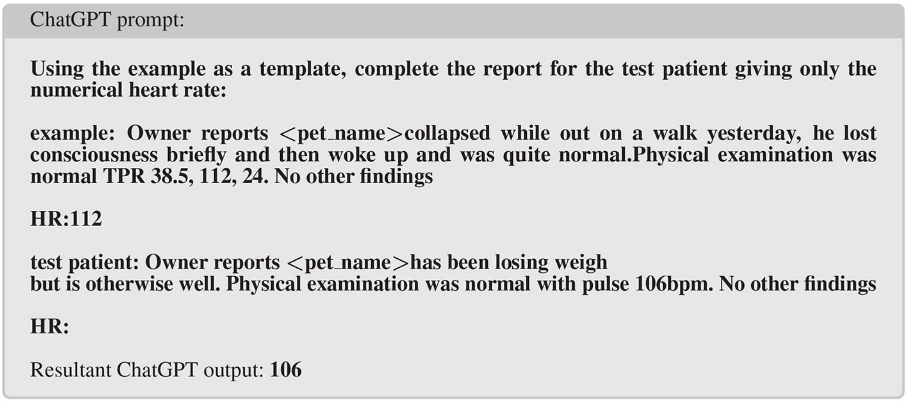


FIGURE 1
Distribution of odds ratios with 95% confidence interval for a given topic to occur in records from specific breeds. The model comprised 200 topics, generated entirely by the computer without external "hints," many having clear clinical relevance. Here, odds ratio for records with topic 23 as their most probable topic are calculated against the odds of records from crossbreeds having that topic as the most probable one. Data for topic 23 with key-words: seizure, seizures, bloods, diazepam, epiphen, phenobarb, epilepsy, pexion, lasted, episode.


ChatGPT prompt:

Using the example as a template, complete the report for the test patient giving only the numerical heart rate:

example: Owner reports <pet_name>collapsed while out on a walk yesterday, he lost consciousness briefly and then woke up and was quite normal.Physical examination was normal TPR 38.5, 112, 24. No other findings

HR:112

**test patient: Owner reports <pet_name>has been losing weigh
but is otherwise well. Physical examination was normal with pulse 106bpm. No other findings**

HR:

Resultant ChatGPT output: **106**

FIGURE 2
Example of prompt engineering to extract selected data (in this case heart rate) from clinical narratives.

Package (41) and when applied to 1,000,000 SAVSNET records, we produced a topic model comprising 200 different topics each with keywords produced by the model without external prompting or intervention. Each topic is characterized by the keywords present and these usually have a clear clinical correlate for instance words describing ear disease ("ear," "canal," "left ear," "wax," "right ear," "otitis," "discharge," "left," "drops," "both," "osurnia") and words describing wound management ("collar," "wound," "buster collar," "poc," "keep," "healing," "looks," "wound looks," "healed," "post," "off," "licking," and "well"). The probability distribution for a given topic could be calculated across breeds. An example of how a topic relating to seizures (with keywords such as seizure, bloods, diazepam epiphen, and epilepsy) created using BERTopic in this way is distributed in records from dogs of different breeds is shown in Figure 1. This data was very similar to published breed-related data on seizuring (42). While topics may not be precise, they can allow rapid and comprehensive screening of large volumes of records for a huge variety of disease phenotypes in a single study.

2.4 Generative models

The neural network models described so far have ranged from tens of thousands of connections (parameters) through to hundreds of millions of parameters (in BERT Language models). More recently models have been developed that have billions to trillions of parameters. As a comparison, the human neocortex is estimated to have in the order of 200 trillion synapses (43). These generative models are often trained using tasks such as text completion (particularly next word prediction) and question/answer using massive training sets from diverse sources. Examples include GPT-3, ChatGPT, created by OpenAI (44) and OPT and Llama from Meta (45). The complexity of the internal wiring of these models combined with the extensive training data set leads to behavior that is uncannily human. The main form of interaction is to provide a prompt to which the model generates an answer (this is after all the task the model is trained on). Thus prompting ChatGPT (8) with the text *"tell me in 30 words, why dogs vomit"* returns the text *"Dogs can vomit due to various reasons such as eating too fast, consuming something toxic, having an underlying medical condition, or experiencing motion sickness."* On the face of it, this conversational dialog appears challenging to extract structured data from but the model will respond with more useful text if prompted in a more structured manner. So called "prompt engineering" allows the prompter to coerce the model to return structured outputs and can be used to classify text from clinical records. For instance when trying to extract an important clinical index of health such as heart rate data from free-text narratives, a suitably engineered prompt can do this (Figure 2). This technique would allow evaluation of a population at scale where previously extensive manual reading or arcane rule-based classification of records might be needed covering only a small subset of available records.

The ability to run prompts across numerous texts is made available through a programming interface which, in theory, could allow for screening of extremely large numbers of records but there are several drawbacks: Firstly, for large volumes, record screening can have a cost per record which can be come substantial for very large datasets (millions of records) with multiple prompts; Secondly, some large language models capture the prompt text for further model training which can lead to some of the potentially sensitive material appearing in outputted text for other users. In order to run such large language models on a local machine (i.e., a copy which will not train on the text or be prompted by external users) requires substantial computing resource. Additionally, when using more complex prompts looking for complicated responses, further issues arise: given the very varied nature of the training data, these models can output opinion that can be misleading, completely fictitious (often referred to as "hallucination") and even prejudiced/unwholesome. In the case of ChatGPT, beyond the prompt-response training, further fine tuning has been implemented using reinforcement learning from human feedback among other tools to try to forestall misleading or offensive content (44). In our preliminary experiments utilizing ChatGPT to identify overweight animals and body condition scores based on notes written by clinicians during consultations, ChatGPT's performance compared very favorably with a rule-based classifier used for the same task (46). Manual reading of records remains the gold standard against which such approaches are validated.

3 Conclusion

Machine learning and artificial intelligence are revolutionizing our ability to automate the generation of signals relating to disease phenotypes in veterinary patients using EHRs. The underlying machinery of the tools is becoming less and less explainable as models with massive numbers of parameters are trained on vast datasets. The impact of large language models will become very significant in the coming years and it will be important that users understand the provenance of data from these models and that researchers work to ensure that the outputs from these models are explainable. While the methods discussed in this second part of the review series have clear benefits in screening huge volumes of data sometimes without requiring any stipulation of the nature of signals to detect, there will remain a role for manual review of records identified by these tools to maintain confidence that valid conclusions are drawn from them.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HD: Writing - review & editing, Writing - original draft. GN: Writing - review & editing. GA: Writing - review & editing. MA: Writing - review & editing. NA: Writing - review & editing, Writing - original draft. SF: Writing - review & editing, Writing - original draft. AR: Writing - review & editing. P-JN: Writing - review & editing, Writing - original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Work for this review carried out by GA, GN, and MA at Manchester University was funded by CWG grant from Dogs Trust (SAVSNET Agile) and Healtex: UK Healthcare Text Analytics Research Network (EP/N027280/1, EPSRC). HD was funded by University of Liverpool. P-JN was funded by CWG grant from Dogs Trust (SAVSNET Agile). SF was supervised by NA on a BBSRC funded Ph. D. studentship. This work was partly supported by Dogs Trust through the SAVSNET Agile award.

Acknowledgments

This work would not have been possible without the contribution of practicing vets across the UK contributing to the SAVSNET project and without the help and support of the SAVSNET core team including Bethaney Brant, Steve Smyth, and Gina Pinchbeck.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Cambria E, White B. Jumping NLP curves: a review of natural language processing research. *IEEE Comput Intell Mag.* (2014) 9:48–57. doi: 10.1109/mci.2014.2307227

2. Sundermann AJ, Miller JK, Marsh JW, Saul MI, Shutt KA, Pacey M, et al. Automated Data Mining of the electronic health record for investigation of healthcare-associated outbreaks. *Infect Contr Hospit Epidemiol.* (2019) 40:314–9. doi: 10.1017/ice.2018.343

3. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf.* (2017) 40:1075–89. doi: 10.1007/s40264-017-0558-6

4. Liu F, Jagannatha A, Yu H. Towards Drug Safety Surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from Electronic Health Records. *Drug Saf.* (2019) 42:95–7. doi: 10.1007/s40264-018-0766-8

5. Radford A, Tierney, Coyne KP, Gaskell RM, Noble PJ, Dawson S, et al. Developing a network for small animal disease surveillance. *Vet Rec.* (2010) 167:472–4. doi: 10.1136/vr.c5180

6. McGreevy P, Thomson P, Dhand NK, Raubenheimer D, Masters S, Mansfield CS, et al. VetCompass Australia: a national big data collection system for veterinary science. *Animals.* (2017) 7:74. doi: 10.3390/ani7100074

7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference* (2018). p. 4171–86. Available online at: http://arxiv.org/abs/1810.04805 (accessed November 23, 2023).

8. OpenAI. *ChatGPT (Version 3.5)* (2023). Available online at: https://www.openai.com (accessed November 23, 2023).

9. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. ArXiv [preprint] arXiv: abs/2302.13971. (2023). doi: 10.48550/arXiv.2302.13971

10. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci.* (2021) 2:1. doi: 10.1007/s42979-021-00815-1

11. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations, Workshop Track Proceedings*. Scottsdale, AZ: ICLR 2013 (2013).

12. Wu L, Yen IEH, Xu K, Xu F, Balakrishnan A, Chen PY, et al. Word mover's embedding: from Word2Vec to document embedding. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics (2018). p. 4524–34.

13. Fan Y, Pakhomov S, McEwan R, Zhao W, Lindemann E, Zhang R. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open.* (2019) 2:246–53. doi: 10.1093/jamiaopen/ooz007

14. Workman ET, Divita G, Shao Y, Zeng-Treitler Q. A proficient spelling analysis method applied to a pharmacovigilance task. *Stud Health Technol Informat.* (2019) 264:452–6. doi: 10.3233/SHTI190262

15. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, LA: Association for Computational Linguistics (2018). p. 2227–37. Available online at: https://aclanthology.org/N18-1202 (accessed November 23, 2023).

16. Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning. ICML '08*. New York, NY: Association for Computing Machinery (2008). p. 160–7.

17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv [preprint] arXiv: abs/1706.03762. (2017). doi: 10.48550/arXiv.1706.03762

18. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv [preprint] arXiv:abs/2005.14165. (2020). doi: 10.48550/arXiv.2005.14165

19. Thoppilan R, Freitas DD, Hall J, Shazeer N, Kulshreshtha A, Cheng H, et al. LaMDA: language models for dialog applications. arXiv [preprint] arXiv: abs/2201.08239. (2022). doi: 10.48550/arXiv.2201.08239

20. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. *BMC Med Informat Decision Mak.* (2021) 21:6. doi: 10.1186/s12911-021-01534-6

21. Lloyd SS, Rissing JP. Physician and coding errors in patient records. *J Am Med Assoc.* (1985) 254:1330–6.

22. Hasan M, Meara RJ, Bhowhick BK. The quality of diagnostic coding in cerebrovascular disease. *Int J Qual Health Care.* (1995) 7:407–10.

23. Farzandipour M, Sheikhtaheri A, Sadoughi F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). *Int J Inform Manag.* (2010) 30:78–84. doi: 10.1016/j.ijinfomgt.2009.07.002

24. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res.* (2005) 40:1620–39. doi: 10.1111/j.1475-6773.2005.00444.x

25. Shi H, Xie P, Hu Z, Zhang M, Xing EP. Towards automated ICD coding using deep learning. arXiv [preprint] arXiv: abs/1711.04075. (2017). doi: 10.48550/arXiv.1711.04075

26. Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. arXiv [preprint] arXiv: abs/1912.00862. (2019). doi: 10.48550/arXiv.1912.00862

27. Cao P, Chen Y, Liu K, Zhao J, Liu S, Chong W. HyperCore: hyperbolic and co-graph representation for automatic ICD coding. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics (2020). p. 3105–14. Available online at: https://aclanthology.org/2020.acl-main.282 (accessed November 23, 2023).

28. Zhang Z, Liu J, Razavian N. BERT-XML: large scale automated ICD coding using BERT pretraining. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics (2020). p. 24–34. Available online at: https://aclanthology.org/2020.clinicalnlp-1.3 (accessed November 23, 2023).

29. Pascual D, Luck S, Wattenhofer R. Towards BERT-based automatic ICD coding: limitations and opportunities. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics (2021). p. 54–63. Available online at: https://aclanthology.org/2021.bionlp-1.6 (accessed November 23, 2023).

30. Nie A, Zehnder A, Page RL, Zhang Y, Pineda AL, Rivas MA, et al. DeepTag: inferring diagnoses from veterinary clinical notes. *NPJ Digit Med*. (2018) 1:8. doi: 10.1038/s41746-018-0067-8

31. Zhang Y, Nie A, Zehnder A, Page RL, Zou J. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. *NPJ Digit Med*. (2019) 2:1. doi: 10.1038/s41746-019-0113-1

32. Hur B, Baldwin T, Verspoor K, Hardefeldt L, Gilkerson J. Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics (2020). p. 156–66. Available online at: https://aclanthology.org/2020.bionlp-1.17 (accessed November 23, 2023).

33. Farrell S, Appleton C, Noble PJM, Al Moubayed N. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Sci Rep*. (2023) 13:18015. doi: 10.1038/s41598-023-45155-7

34. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Machine Learn Res*. (2003) 3:993–1022.

35. Rehurek R, Sojka P. *Gensim–Python Framework for Vector Space Modelling*. Brno: NLP Centre, Faculty of Informatics, Masaryk University (2011).

36. Noble PJM, Appleton C, Radford AD, Nenadic G. Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. *PLoS ONE*. (2021) 16:e0260402. doi: 10.1371/JOURNAL.PONE.0260402

37. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus*. (2016) 5:1608. doi: 10.1186/s40064-016-3252-8

38. Pérez J, Pérez A, Casillas A, Gojenola K. Cardiology record multi-label classification using latent Dirichlet allocation. *Comput Methods Progr Biomed*. (2018) 164:111–9. doi: 10.1016/j.cmpb.2018.07.002

39. Ghosh S, Chakraborty P, Nsoesie EO, Cohn E, Mekaru SR, Brownstein JS, et al. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Sci Rep*. (2017) 7:40841. doi: 10.1038/srep40841

40. Blei DM, Lafferty JD. Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine Learning. ICML '06*. New York, NY: Association for Computing Machinery (2006). p. 113–20.

41. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv [preprint] arXiv:220305794. (2022). doi: 10.48550/arXiv.2203.05794

42. Erlen A, Potschka H, Volk HA, Sauter-Louis C, O'Neill DG. Seizure occurrence in dogs under primary veterinary care in the UK: prevalence and risk factors. *J Vet Intern Med*. (2018) 32:1665–76. doi: 10.1111/jvim.15290

43. Nguyen T. Total number of synapses in the adult human neocortex. *Undergrad J Math Model*. (2010) 3:26. doi: 10.5038/2326-3652.3.1.26

44. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. *arXiv preprint*. (2022). doi: 10.48550/arXiv.2203.02155

45. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, et al. OPT: open pre-trained transformer language models. *arXiv preprint*. (2022). doi: 10.48550/arXiv.2205.01068

46. Fins IS, Davies H, Farrell S, Torres JR, Pinchbeck G, Radford AD, et al. Evaluating ChatGPT text-mining of clinical records for obesity monitoring. *Vet Record*. (2023) 2023:e3669. doi: 10.1002/vetr.3669