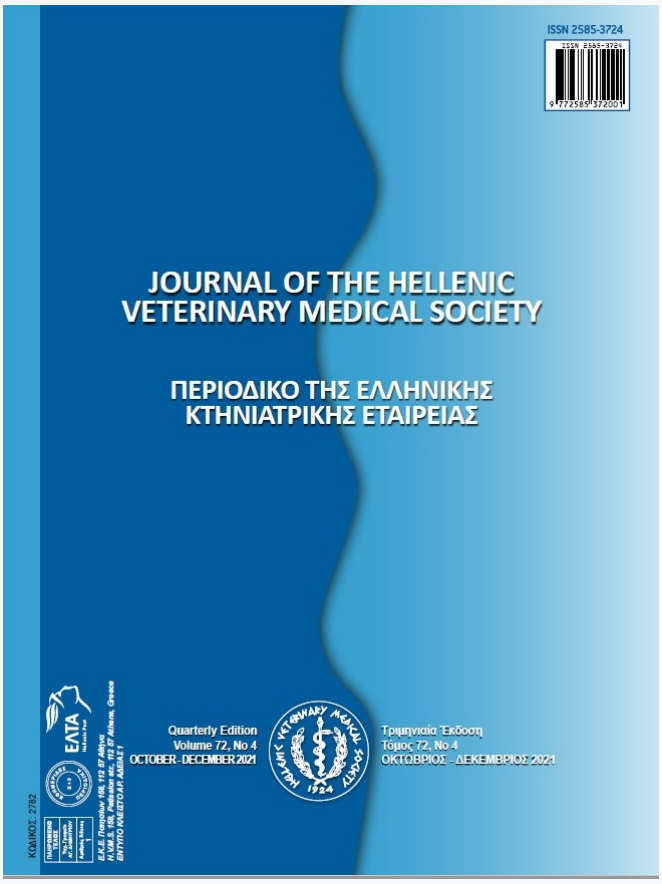
Journal of the Hellenic Veterinary Medical Society

Vol 72, No 4 (2021)



A 30 year topic analysis of Veterinary Medicine literature

I FYTILAKOS, V ALEXOPOULOS

doi: 10.12681/jhvms.29430

To cite this article:

FYTILAKOS, I., & ALEXOPOULOS, V. (2022). A 30 year topic analysis of Veterinary Medicine literature. *Journal of the Hellenic Veterinary Medical Society*, *72*(4), 3401–3414. https://doi.org/10.12681/jhvms.29430

J HELLENIC VET MED SOC 2021, 72(4): 3401-3414
ΠΕΚΕ 2021, 72(4): 3401-3414

Research article
Ερευνητικό άρθρο

A 30 year topic analysis of Veterinary Medicine literature

I. Fytilakos[1], V. Alexopoulos[2]

[1]*Department of Zoology-Marine Biology, National and Kapodistrian University of Athens*

[2]*School of Veterinary Medicine, University of Thessaly, Karditsa, Greece*

ABSTRACT: In the present study Latent Dirichlet allocation (LDA) was used as a generative probabilistic model to extract major topics in interdecadalresearch for the Veterinary Medicine scientific literature. A total of 22 topics were extracted during the 1991-2000 period, 23 topics during 2001-2010 and 60 topics during 2011-2020. Three different algorithms were used to validate the model: perplexity, silhouette clustering and gradient boosted trees. All three validation metrics showed that LDA performed well in extracting topics. Each decade was characterized by unique topics as well as common topics which existed throughout periods. The most frequent topics were identified and trends were quantified with the use of indexes. A list of the 30 most frequent and most associated with the term Veterinary Medicine words is provided. A shift in scientific thinking probably occurred during the 30-year-period in the process of incorporating the fields related to Veterinary students, antimicrobial resistance and animals' behavior.

***Keywords*:** Veterinary Medicine, latent Dirichlet allocation, topic modeling, model validation, literature trends

Corresponding Author:
Ioannis Fytilakos, Department of Zoology-Marine Biology, National and Kapodistrian University of Athens, Panepistimiopolis, 15701, Ilisia, Athens, Greece
E-mail address: ifitilakos@gmail.com

Date of initial submission: 28-09-2020
Date of acceptance: 12-11-2020

INTRODUCTION

The more classical approach of collecting information and exploratory analysis in scientific literature includes qualitative research methods, which offer high flexibility and focus well on understanding a problem. However, there are faster and less subjective methods to study the literature. Quantitative research methods offer solutions as they include easy data collection and analysis procedures and are not affected by the subjectivity of the researcher (Queirós et al., 2017).

The Latent Dirichlet allocation (LDA) is a generative probabilistic Bayesian model for collecting of discrete data such as text corpora (Blei et al., 2003). In LDA, text documents are used as a collection of words to identify underlying topics. Unsupervised machine learning techniques such as LDA, require little prior work from the researcher and are able to categorize big amount of data. However, LDA can also generate ambiguous topics which are hard to interpret and to classify as discussed previously in Web analysis research (Nanni, 2017). LDA has been successfully applied in the past in various fields such as in biology, biodiversity, climate change and animal communities (Zhang et al., 2019).

Studies in the literature of Veterinary Medicine with the use of LDA are scarce. However other advanced statistical techniques of machine learning have been applied on necropsy reports for detecting emerging diseases (Bollig et al., 2020) and in a literature review of urothelial cancer (Lin et al., 2020). Machine-learning-based literature mining may analyze large collections of documents, identify patterns in a dataset using statistical and computational methods or make predictions based on the discovered patterns (Lin et al., 2020). It is useful in summarizing key research themes and trends (Lin et al., 2020).

Previous works on text information extraction of literature mainly used text mining processesto study several subfields of Veterinary Medicine such as poor animal welfare (Contiero et al., 2019), epidemiology (Van der Waal et al., 2017), studies in livestock animals (Sahadevan et al., 2012), parasitology (Ellis et al., 2020), studies in antimicrobial prescribing practices (Welsh et al., 2017) or in geographic trends of science (Christopher andMarusic, 2013). Furthermore, another aspect of extracting information from text is the construction of automated electronic surveillance systems in order to predict emergency situations regarding disease outbreak (Lustgarten et al., 2020;Dórea et al., 2015), companion animal syndromes (Anholt et al., 2014), temporal and spatial features of diseases (Bollig et al., 2020) or in decision support frameworks (Jones-Diette et al., 2019).

Although many information collection techniques have been applied in Veterinary Medicine in the past for the purposes of surveillance systems, historical studies on an extended temporal scale has not been conducted. This is the first attempt that aims at clarifying the major scientific terms of literature during a 30 year period, from 1991to present. Topic extraction using LDA modeling is the main purpose of the present study to reveal the diachronically major topics in the field of Veterinary Medicine and to ascertain possible interdecadal differences in trends. Simultaneously, a lack of validation processes in related literature has been observed, thus a three-way validation approach was followed in this study to calculate the accuracy of the LDA model in predicting topics.

MATERIALS AND METHODS

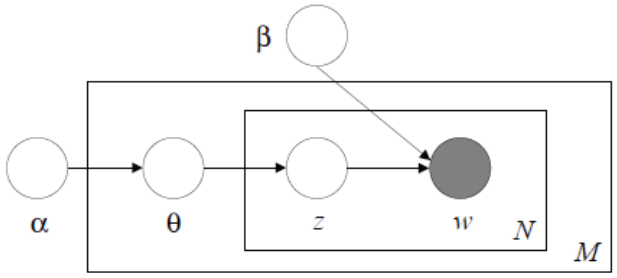
Abstracts of publications related to Veterinary Science were extracted from the Web of Science database. For this purpose the term “Veterinary Science” in quotes was entered in the search engine to extract only the Abstracts belonging to the category of Veterinary Sciences. Research articles, Reviews and Conference Proceeding papers were selected as they are the source of Abstracts and the study was planned at the level of decade; thus three decades 1991-2000, 2001-2010 and 2011-2020 were used as a filter to extract 857, 1.732 and 3.256abstracts respectively. Abstracts were stored in three separate files representing decades.

At first, a pre-processing stage transformed Abstract texts into words: special characters, symbols, numbers and articles were excluded from the analysis with the “stop words” procedure which was common for all three decades. KH coder (Higuchi, 2016) is able to analyze English data by grouping derivatives based on a built-in dictionary (it is called lemmatization and uses the Stanford POS Tagger toolkit) or by cutting the last letters and grouping words by their stem (it is called stemming and uses the Snowball Stemmer toolkit). For instance, the term “veterinary” during the stemming process becomes “veterinari”, terms “veterinarian/ veterinarians” become “veterinarian” and terms “tumor/ tumors/ tumorous/ tumoral” become “tumor”. Both toolkits were tested for their ability to group derivatives and to extract a large number of words. After the data preparation process, a pre-processing command was used to segment the word file

into words. This is a necessary internal process to organize the results in a SQL database form, to carry out searching and tabulating (Higuchi, 2016).

Quantitative analysis of text data followed. Words in the documents were counted to obtain the number of appearances. A word frequency table was constructed with the 30 most frequent words in each decade. A comparison of term frequencies was done to study differences between decades. For this purpose word frequencies were normalized as documents contained a different number of abstracts. A Word Association Table was constructed for each decade using the Jaccard coefficient to determine the associations between words and the term "Veterinary Science". The Jaccard coefficient emphasizes whether or not specific words co-occur, and is suitable for analyzing sparse data and is also calculated irrespectively of the term frequency (Higuchi, 2016). The values of Jaccard coefficient vary between 0 and 1. In KH Coder, words that appear frequently in the same sentence/paragraph are considered to be closely associated, and in that case the Jaccard coefficient reaches 1.

The Latent Dirichlet allocation (LDA) was used as a generative probabilistic model to extract topics from documents. LDA is based on the idea that documents represent a distribution of words which surround a topic. The model assumes that we predefine the number of topics into a document (k), one parameter for the distribution of topics into a document (α) and one parameter for the distribution of words into topics (β). These two parameters were set as α = 50/N and β = 0.1 in all runs of the algorithm. A plate notation from (Blei et al., 2003) of LDA variables is presented below, where:



M denotes the number of Abstracts
N is number of words in a given Abstract (Abstract i has N_i words)
α is the parameter of the Dirichlet prior on the per-Abstract topic distributions
β is the parameter of the Dirichlet prior on the per-topic word distribution
θ_i is the expected topic proportion of Abstract M, which is generated by a Dirichlet distribution parameterized by parameter α
z is the topic for the *n*th word in Abstract M and
w is the word in the *n*th position word of Abstract M.

Three measurements were used to evaluate the effectiveness of the LDA modeling those of perplexity, gradient boosted trees (GBT) and silhouette clustering. Validating LDA is a hard procedure as unsupervised machine learning uses data without pre-existing labels. For this purpose perplexity was used to define the ideal number of topics into each document. GBT were used to evaluate the performance of LDA in identifying topics and silhouette clustering with a manual labeling procedure was used in order to validate topics extracted with LDA.

Perplexity is a measurement of how well a probability model predicts a sample.In language modeling, perplexity decreases in the likelihood of the test data. A lower perplexity score indicates a better generalization performance (Blei et al., 2003). Due to the large volume of words collected during the 2011-2020 period, perplexity tends to get minimized at over 150 topics which makesit difficult to interpret such a large number of topics. Considering that word frequency of the endmost topics was very low only the 60 most important topics were finally presented concerning this decade in terms of word frequency.

$$perplexity\ (Document) = exp\left\{-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\right\}$$

M denotes the number of Abstracts
N is number of words in a given Abstract (Abstract i has N_i words)
p is the probability distribution of the model
w is aword-level variable.

Gradient boosted trees is an ensemble consisting of a set of alternative models using multiple learning algorithms to produce a more accurate classifier than that of the standard classifier (Opitz and Maclin, 1999). In this case the whole dataset (5.845 Abstracts) was divided in two random subsets for a total of10 random times to compare the model performance. These subsets were used in combinations of two to produce 45 different trees to evaluate the performance of LDA in identifying topics. Each subset was used to extract 40 topics, thus 80 topics were compared each time (40 topics from the training set and 40 topics from the prediction set) which were manually labeled. Manual labeling is a time-consuming process and rules have to be followed in order to obtain meaningful topics for a neutral individual. Three rules were applied to label topics: at least two words are necessary in order to label a topic meaningful; if there were more than two common words between topics, the label was given according to a third

word etc., and words in a similar context were considered as belonging to the same topic. For instance, the words tumor, tumour, malign, lymphoma, neoplasm, cancer were all identified under the topic of cancer and the words echocardiography, cardiac, pulmonary, valve, doppler, heart, pressure, arrhythmia, myocardium were all identified under the topic of heart-related problems. The weight of each word was used as an advisory index but not as a criterion in the identification of common topics between training and prediction sets. An example of labeling is given in Table 1.

Table 1. Example of the manual process followed to give labels to topics according to their meaning. Topics with the same context were given same labels (Topic 1 was the label for tumor related context) while topics with different context were given different labels (Topic 2 and 3)

Training set		**Prediction set**	
Word	Weight	Word	Weight
Topic1		Topic1	
tumor	394.0	tumor	284.0
cell	316.0	cell	256.0
histopatholog	150.0	tumour	153.0
tumour	144.0	lymphoma	104.0
histolog	115.0	cytolog	99.0
Topic2		Topic3	
effect	87.0	vaccin	198.0
dog	72.0	infect	182.0
agent	70.0	antibodi	101.0
chemotherapi	57.0	immun	81.0
treatment	56.0	virus	63.0

An indirect use of silhouette clustering was used to validate the results of LDA. Silhouette clustering is a method of interpretation and validation of consistency and cohesion within clusters of data. The average silhouette width can be used to evaluate cluster validity (Rousseeuw, 1987). As LDA extracted topics without any labels applied, the same manual procedure described above was used to give labels to each topic. An agglomerative hierarchical clustering of the two sets combined (each set consisted of 40 topics * five most frequent words of each topic = 200 words * two sets = 400 words) was used to identify groups of topics at the level of sets (training and prediction). Silhouette clustering was then applied to measure the consistency and the cohesion of 45 different combinations of clusters. A silhouette value (score) of one data point can be calculated with the formula:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \quad -1 \le s(i) \le 1$$

where a(i) is the dissimilarity 'within' a cluster and b(i) is the dissimilarity 'between' clusters.

The range of silhouette scores varies between +1 for objects that classified well in the predefined clusters (those of training and prediction sets) and -1 for objects that have been misclassified. In our case an accurate model would lead to small 'between' and 'within' distance dissimilarities between the training and the prediction sets, thus s(i) score would tend to zero. The overall average silhouette width for the entire plot (the average of the s(i) for all objects i belonging to the whole dataset)(Rousseeuw, 1987), was calculated with the Euclidean distance as a metric to indirectly calculate the accuracy of the model.

Three measures of topic diagnostic information were used to highlight the most frequent topics of each decade those of document entropy, document burstiness and corpus distance. Entropy is the degree of probability of a topic to occur in different documents (Abstracts in our case). The concept of entropy in information systems has been introduced by Shannon (1948) and has been widely used in topic extraction analysis. A topicof higher entropy is possible to have been extracted by a high number of several documents. Burstiness in natural language documents is the property of the most common words to represent a large number of topics (Boyd-Graber et al., 2014). This relationship of a few words representing a majority of topics and vice versa seems to follow a Zipfian distribution, an empirical law previously observed in social and physical sciences (Boyd-Graber et al., 2014). The combination of entropy and burstiness gives us a measure of the most common words and topics of literature. Finally the distance of topics from the corpus of each decade was measured with the Kullback-Leibler divergence distribution (Kullback andLeibler, 1951). A lesser distance from a corpus indicates that a topic is closely related to Veterinary medicine. Closely related topics consist of the most frequent words of the corpus. On the other hand, more distant topics are more distinct from the main corpus. A three-decade comparison was carried out to ascertain trends in corpus-related topics. A list of the most frequent words appearing before and after selected terms is presented in the Supplementary material, aiming not only to help but also to furtherly promote the interpretation of the topics and the comprehension of their position into the text.

Three open source free software were used for the procedures of document preprocessing, topic extraction and silhouette clustering: KH coder v. 3.Beta.01a (Higuchi, 2016), Orange v. 3.26.0 (Demsar et al., 2013) and RapidMinerv. 9.7.2+ (Mierswa et al., 2006).

RESULTS

Word stemming extracted almost the same number of words compared with the lemmatization procedure and was preferred for its better grouping ability (Table 2). From the total number of terms approximately half of them were excluded from the analysis (Table 2). The majority of words were common between decades with a different order of appearance (Table 3). Word association analysis revealed 14 strongly associated unique words in the between decade comparison (Table 3). Four words (dure, test, drug, system) were unique during 1991-2000, four words (provide, patient, student, medic) during 2001-2010 and six words (associ, dog, perform, compare, product, group) during 2011-2020. Normalized word frequency comparison revealed an increase in the use of common words such as dog, cat, pig, antibiotics and tumor and a decrease in words such as substance, vaccine, market, public and epidemiology during the decade 2011-2020. A detailed list of comparisons is given in Table 4.

Table 2. Comparison of two different algorithms for their ability to extract words from documents. The Tagger algorithm uses a built-in dictionary while the Stemmer algorithm groups words by their stem and then cuts their last letters

	1991-2000	2001-2010	2011-2020
	Stemmer		
Tokens in total	122627	314517	762825
Tokens in use	67094	175453	431440
	Tagger		
Tokens in total	126876	326773	797047
Tokens in use	67264	175970	433491

Table 3. List of the 30 most frequent and most associated words with the term Veterinary Medicine for each decade. Numbers correspond to frequencies and to Jaccard coefficient (JC).Words that appear frequently in the same abstract are closely associated thus JC reaches 1

Term Frequency						**Associated Words**					
1991-2000		**2001-2010**		**2011-2020**		**1991-2000**		**2001-2010**		**2011-2020**	
veterinari	1527	veterinari	3779	veterinari	6528	medicin	0.8672	medicin	0.8486	medicin	0.8609
medicin	1164	medicin	2538	medicin	4561	anim	0.2721	anim	0.3165	studi	0.4569
anim	813	anim	1833	anim	3809	clinic	0.2053	clinic	0.2841	anim	0.3992
dog	483	dog	1406	studi	3704	result	0.1787	studi	0.2726	result	0.3336
clinic	396	clinic	1192	dog	3521	studi	0.1759	human	0.2249	clinic	0.3129
diseas	378	studi	1121	clinic	2540	develop	0.1726	result	0.2195	human	0.2695
studi	358	case	962	treatment	2104	human	0.1725	univers	0.2047	increas	0.2440
result	309	human	948	result	2020	diseas	0.1713	diseas	0.1889	differ	0.2413
human	285	diseas	914	group	1989	import	0.1521	develop	0.1875	signific	0.2378
effect	281	student	723	diseas	1936	year	0.1448	year	0.1793	evalu	0.2332
develop	271	result	692	effect	1846	effect	0.1367	includ	0.1757	treatment	0.2312
treatment	269	develop	666	human	1846	increas	0.1287	increas	0.1681	associ	0.2293
test	268	health	648	signific	1760	method	0.1265	case	0.1632	effect	0.2200
drug	267	treatment	634	differ	1660	includ	0.1259	evalu	0.1570	includ	0.2169
method	246	resist	611	increas	1656	practic	0.1251	differ	0.1567	diseas	0.2118
case	243	report	600	case	1583	univers	0.1240	effect	0.1529	dog	0.1998
differ	226	increas	599	evalu	1583	treatment	0.1169	import	0.1501	perform	0.1967
year	219	effect	573	cell	1501	differ	0.1167	practic	0.1442	develop	0.1896
practic	206	year	569	student	1451	examin	0.1088	treatment	0.1440	report	0.1884
system	205	differ	564	perform	1377	review	0.1074	examin	0.1437	import	0.1860
cat	202	cat	552	report	1372	case	0.1066	provid	0.1330	year	0.1830
increas	202	practic	552	associ	1330	time	0.1044	patient	0.1324	compar	0.1829
import	201	evalu	547	cat	1321	dure	0.1028	signific	0.1312	case	0.1773
group	200	univers	533	resist	1307	test	0.1022	report	0.1279	product	0.1736
signific	200	examin	512	product	1216	signific	0.1005	method	0.1259	method	0.1725
veterinarian	193	includ	508	test	1189	determin	0.0995	review	0.1246	determin	0.1712
infect	188	veterinarian	508	compar	1182	evalu	0.0993	time	0.1239	time	0.1645
evalu	186	patient	491	includ	1177	articl	0.0978	student	0.1231	univers	0.1629
examin	186	hors	485	control	1146	drug	0.0968	articl	0.1215	examin	0.1623
includ	185	infect	481	sampl	1146	system	0.0955	medic	0.1177	group	0.1573

Table 4. Normalized word frequency comparisons of selected words presented in subcategories. Arrows indicate a frequency value over [⇧], below [⇩] or equal [▭] to the average of the 1991-2020 study period. Red gradient was used to highlight higher word frequency values of each subcategor

Animals	dog	14,61	21,20	30,77	cat	5,95	8,19	11,42	hors	5,03	7,17	6,41	pig	1,27	1,68	5,15	bird	1,18	0,55	2,81	sheep	1,08	0,87	1,56
	domest	0,99	1,36	1,53	mammari	0,62	0,90	0,68	cow	2,04	1,56	1,90	bovin	1,24	1,03	1,25	breed	2,50	2,72	3,30				
Diet	feed	1,58	0,97	6,37	milk	1,42	0,98	2,16	dairi	0,65	1,16	1,37	nutrit	1,48	0,94	1,14								
Other	infecti	1,14	1,26	1,24	care	1,21	2,72	4,00	chronic	0,87	1,32	2,12	protocol	0,81	0,81	1,69	routin	0,81	0,97	1,16	altern	1,33	1,09	2,01
Immune system	antibiot	3,73	2,55	5,41	antibodi	1,48	0,69	1,78	vaccin	2,47	3,31	2,27	antimicrobi	1,67	5,29	8,84	antigen	0,53	0,69	0,86	serum	2,72	2,86	4,96
	plasma	1,45	1,82	2,37	immun	1,24	1,47	2,54	virus	1,14	1,83	1,65	metabol	0,90	0,80	1,72	dna	0,56	0,58	1,03				
Professions	human	8,51	14,22	16,04	student	2,84	10,79	12,56	administ	0,93	1,18	1,95	econom	0,93	0,84	1,04	safeti	0,87	1,18	1,34	market	0,81	0,51	0,48
	owner	0,81	1,99	3,45	materi	0,62	1,64	2,55	public	2,62	5,70	2,69												
Pathology / Diseases	surgeri	2,87	2,58	3,60	gastrointestin	0,87	0,45	1,11	dose	2,32	2,69	3,81	tumor	1,91	3,69	4,03	disord	1,85	1,62	1,78	therapeut	1,76	2,02	2,87
	joint	1,73	0,60	1,34	pressur	1,54	1,29	1,83	pharmacokinet	1,48	1,03	1,31	salmonella	0,56	0,69	0,59	respiratori	1,33	1,41	1,41				
	renal	1,51	1,13	2,06	patholog	1,45	3,22	2,82	surgic	1,42	2,75	3,79	bone	1,36	2,35	3,18	pathogen	1,36	2,43	2,82	heart	0,68	1,09	1,81
	cardiac	0,65	1,04	1,41	diagnos	1,30	2,70	3,75	mastiti	0,59	0,43	0,65	biochem	0,68	0,60	1,43	diabet	0,68	0,78	0,32				
	epidemiolog	1,30	1,15	1,11	physiolog	1,27	1,18	1,14	strain	1,27	2,93	2,80	genet	1,24	1,21	1,56	mechan	0,96	1,61	2,27	oral	1,11	2,35	2,14
	arteri	1,05	0,92	1,60	enrofloxacin	1,02	0,80	0,63	pharmacolog	0,77	1,10	0,50	urinari	0,71	0,81	1,16	anaesthet	0,68	0,39	0,00				

LDA successfully extracted 22 topics during 1991-2000, 23 topics during 2001-2010 and 60 topics during 2011-2020 (Figure 1-3) as indicated by the perplexity score (Figure 4). Topics including the term "veterinari" differed between decades. Six topics with this term were extracted during 1991-2000, seven topics during 2001-2010 and six topics during 2011-2020 in a relevant different volume of literature Abstracts. During the first decade the terms study, animal-veterinarian, student-education, school, surgery-techniques and drugs-use-effects co-appeared with the term veterinary. During the second decade the terms practice, student-education-program, clinic-examination, human-treatment, study-group, faculty-school and animal-health co-appeared with the term veterinary. Finally the terms clinic-case, university-research, animal-human-health, disease-human-patient, clinic-studies and student co-appeared with the term veterinary during the third decade. Apart from the topics directly related to the overly generalized terms "Veterinari" and "medicine" which were common for all three decades, several other topics were extracted many of which were very specific. For instance, the topics "ultrasound-arteri-pregnanc-fetal-Doppler", "cell-tumor-tumour-dog-carcinoma" and "resist-antimicrobi-isol-strain-antibiot" refer to artery examination, cancer in dogs and antimicrobial resistance respectively.

GBT modeling indicated that LDA is able to extract topics with a root-mean-square error (RMSE) of 2.759 ±0.053. This means that proportionally with the 40 topics extracted from each subset during the validation process, a mean of 2.759 topics were not predicted correctly. Cluster analysis of the 45 different combinations of sets grouped the 80 topics into two clusters at the level of training and prediction sets indicating the degree of correctly predicted (or respectively incorrectly predicted) topics. Silhouette scores of each cluster analysis indicated the number of topics from the training set that was predicted incorrectly (Figure 5).An overall average of s(i) = 0.079864 (±2σ of 0.02429) means that there are small 'between' and 'within' dissimilarities between training and prediction sets, thus the model is accurate.

Each decade consisted of different topics of high entropy and high burstiness (Figure 1-3). Topics "dog-cat-case-clinic-examin", "studi-data-age-veterinari-sampl" and "veterinari-anim-medicin-veterinarian-develop" were most frequent during the first decade. Topics "veterinari-student-medicin-educ-pro-

gram", "veterinari-medicin-use-practic-inform", "clinic-examin-case-veterinari-medicin" and "veterinari-human-medicin-treatment-diseas" were most frequent during the second decade. Finally topics "student-veterinari-medicin-studi-survey", "anim-veterinari-medicin-health-human", "diet-group-feed-supplement-fed","veterinari-medicin-diseas-human-patient" and "case-dog-veterinari-clinic-report" were most frequent during the third decade. A distance from the corpus revealed that "antimicrobial resistance" and "veterinary students" were more distinct topics from the corpus of Veterinary medicine during 1991-2000 while "animal behavior problems" and "surgeries" were more distinct during 2011-2020 (Figure 6).

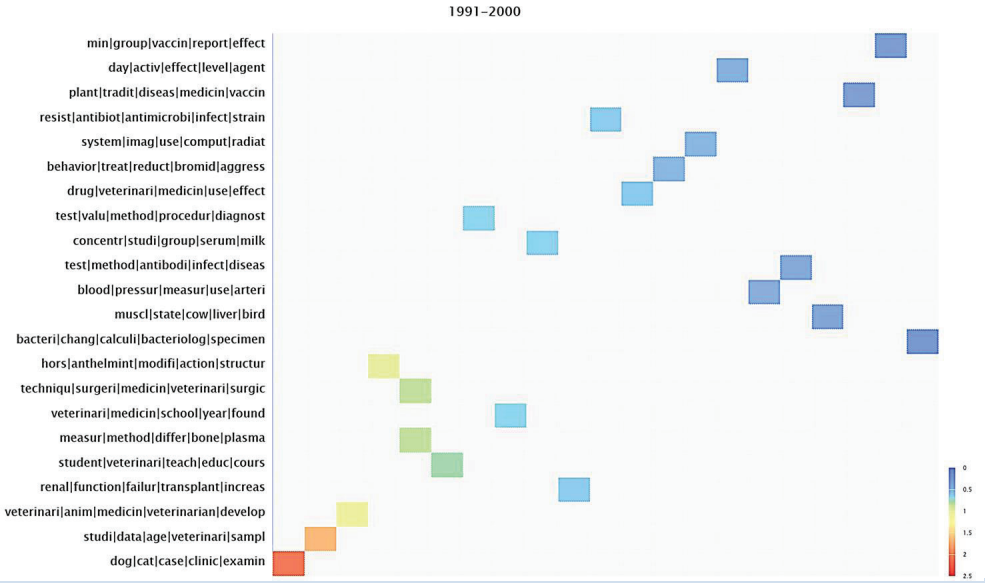


Figure 1. Topics extracted from the period of 1991-2000 with the LDA analysis. The combination of entropy and burstiness measures was used as an index to quantify the most frequent topics of literature (red color)

Figure 2. Topics extracted from the period of 2001-2010 with the LDA analysis. The combination of entropy and burstiness measures was used as an index to quantify the most frequent topics of literature (red color)

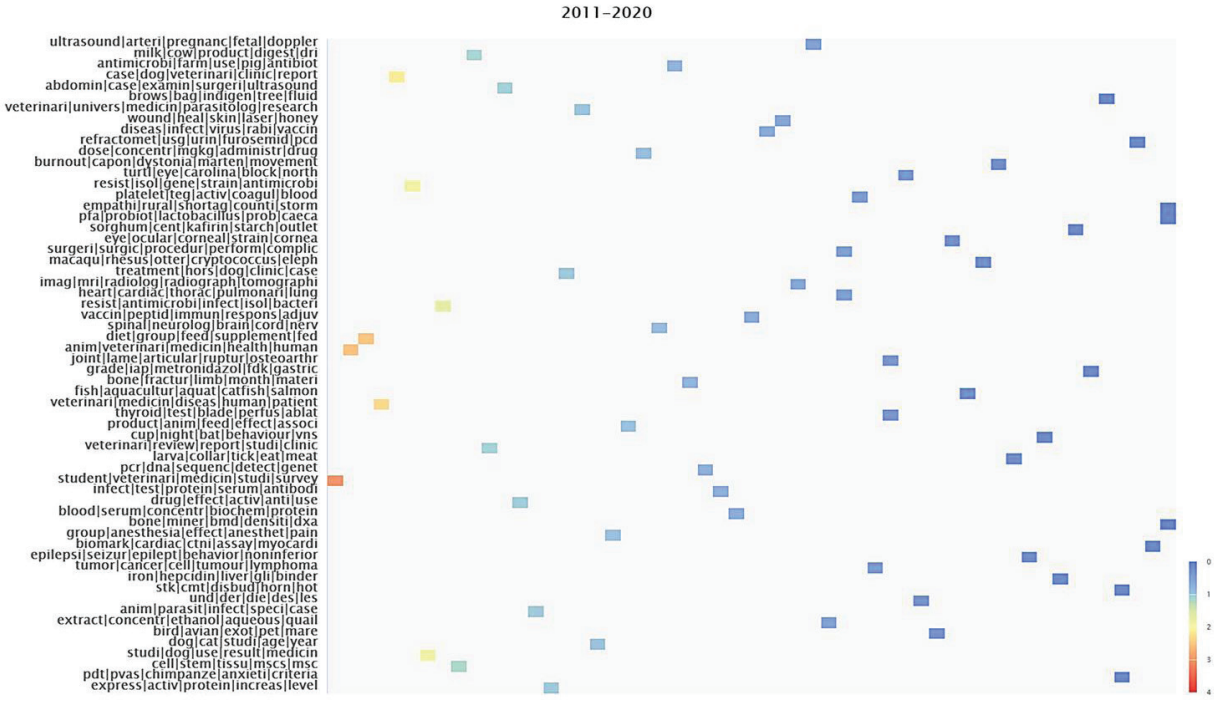


Figure 3. Topics extracted from the period of 2011-2020 with the LDA analysis. The combination of entropy and burstiness measures was used as an index to quantify the most frequent topics of literature (red color)

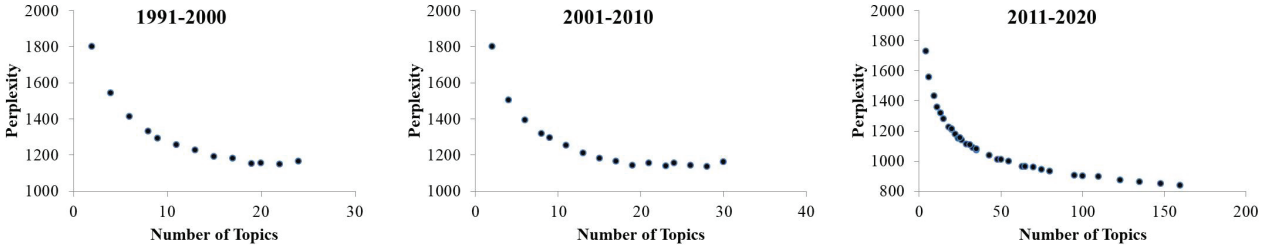


Figure 4. Perplexity is a measure of how well the LDA model predicted a sample. A lower perplexity score indicates better generalization performance of the model

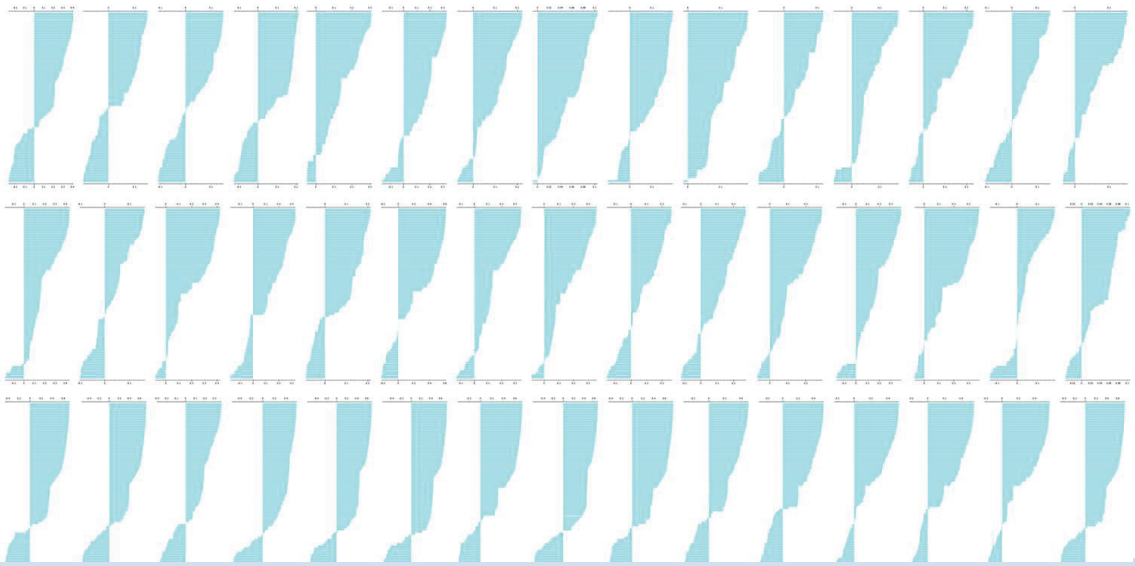


Figure 5. Silhouette graphs showing the consistency and the cohesion of 45 different combinations of training and prediction sets. Small 'between' and 'within' dissimilarities between training and prediction sets, give a small silhouette score [s(i) close to zero] and would imply that the model is accurate. The overall average width for each plot is calculated with the Euclidean distance to give the average s(i)

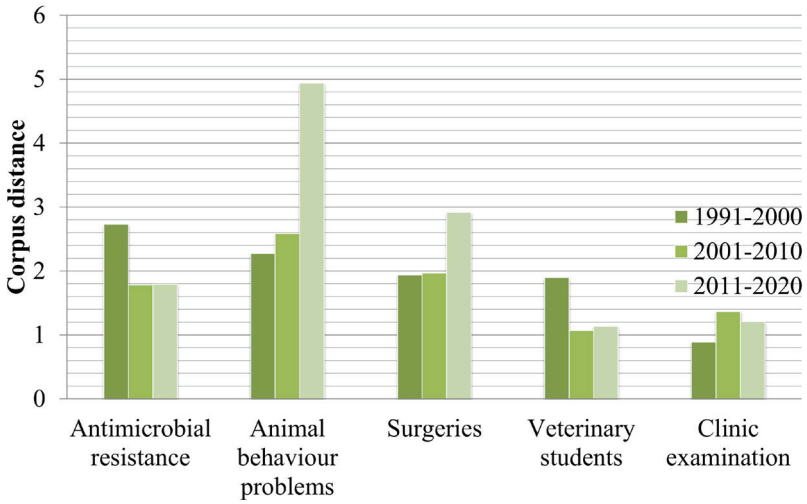


Figure 6. Comparison of selected topics described in terms of their context. Distance of topics from the corpus ofVeterinary medicine-was measured with the Kullback-Leibler divergence distribution. Lower distance from a corpus indicates that a topic is closely related with Veterinary medicine

DISCUSSION

The interpretation of topics is not an easy mission due to the generality of some topics or because many of their words do not match in meaning. The extraction of ambiguous topics has been reported in the past (Nanni, 2017). However, the LDA model managed to uncover a satisfying number of topics, many of which are well interpretable. The term "veterinari" was included in a number of topics during each decade, showing that these topics were not able to acquire a special meaning. Apart from these strongly linked and sometimes overgeneralized topics, other more specific and more meaningful topics existed. The basic analysis in extracting the meaning of a document by giving us a number of topics is the LDA modeling. The interpretation process is subjective and shall focus on the more meaningful topics. To avoid this subjectivity the indexes of topic diagnostic information, document entropy, document burstiness and corpus distance, will constitute a more quantitative approach which helps us objectively interpret results in contrast with the direct subjective human interpretation. They can provide a useful automated summary of topic quality (Boyd-Graber et al., 2014).

LDA analysis constitutes the main technique applied in the present study to extract the major scientific topics of Veterinary literature. At first, the three measures of topic diagnostic information are discussed combined to analyze the relation between persistently present topics during the three decades and the main corpus of literature. Secondly, an effort is made to identify major shifts occurred in the overall topic rank. For this purpose, a simplification is used to shortly describe one topic (Supplementary material). For instance the topics "student|veterinari|teach|educ|cours" and "veterinari|student|medicin|educ|program" are both described as "students" and the topics "resist|antibiot|antimicrobi|infect|strain" and "resist|antimicrobi|isol|strain|antibiot" are both described as "antibiotic resistance" because they refer to students and antibiotics resistance respectively. Thirdly, trends of selected words representing topics are discussed with the use of word frequencies (collocation statistics - Supplementary material) to further facilitate the interpretation of topics under a more analytical aspect.

The use of document entropy and document burstiness helped to identify and rank the most frequent topics. A focus on two animals (dog and cat) during the first time period, a focus on students' educational programs during the second period and a focus on students surveys during the third period indicate different hot topics of each period. Simultaneously, student-related topics were less distinct from the corpus during the last two periods. This means that veterinary student topics became more frequent and were progressively incorporated in the main body of scientific literature. It is possible that a shift in scientific thinking had occurred during the 30-year-period. Students' attitude, learning, motivation, competence in science, learning in practice are some of the aspects of the use of this term in literature (Mich et al., 2010; Jones et al., 2019). A relevant incorporation of a topic into the main corpus of Veterinary Medicine occurs with the topic of antibiotic resistance. The latter is less frequent during the first period and becomes a hot topic less distanced from the corpus during the past two pe-

riods. It is possible that scientists working on antibiotics take time to find novel solutions to this problem or they are aware of microorganisms' resistance to drugs (Toutain et al., 2016) thus the reporting of problems is progressively accumulated in literature.

The topic of animal behavior is extracted in all the three periods from the LDA model but it does not belong to the most frequent ones. This probably shows that veterinarians have already incorporated a perception that behavior problems are equally important with others topics such as clinical examination even before the nineties. The large distance from the corpus during 2011-2020 indicates that during the past few years this topic is linked toseveral words other than the most common ones. These different words probably reflect a variety of new subjects introduced in larger quantities into scientific community which were not strongly present in the past. Indeed, abstracts collected during the present study, contained the term behavior in the context of behavior of aggressive dogs (Csoltova et al., 2017), changes of behavior under a specific therapy (Packer et al., 2016), behavior studies, behavior alterations or behavior abnormalities (Tynes and Sinn, 2014).

A diachronically increasing interest of scientific literature for the topics of students and antimicrobial resistance and a decreasing interest for the topics of surgeries and animal behavior have been observed (Fig. 1-3 and Supplementary Material). On the contrary, the topic of clinical examination constantly seems to be of high interest among publications as it shows up in the five most frequent topics during the 30-years-period. During the 1991 - 2000 period the most frequent topics included clinical examination, were followed by topics including surgeries, students, antimicrobial resistance and animal behavior (Supplementary material). During the 2001 - 2010 period a shift was noticed in topic frequencies rank: the most frequent topic included students while clinical examination, antimicrobial resistance, surgeries and animal behavior followed. Lastly, during the 2011 - 2020 period the most frequent topics included students followed by topics including clinical examination, antimicrobial resistance, surgeries and behavior. The frequencies of these topics show their relative position in the overall ranking of each decade and probably reflect similar shifts in the interests of each decade. Furthermore, during the last period topics including clinical examination, antimicrobial resistance and surgeries were extracted in two different versions. It is possible that a broadening occurred regarding their discussion in the scientific literature. Other topics emerge during specific time periods such as tumors during 2001 - 2020, vaccinations during 2001 - 2010, dosage (dose, mg/kg, concentration) during 2001 - 2020, dairy (milk, cow) during 1991 - 2000 and 2011 - 2020 and blood pressure during 1991 - 2010. It is difficult to tell whether there is a specific incident that provoked these shifts in topic ranking (for instance a pandemic) or if advances in technology have promoted the interest of scientists (for instance new tumor confrontation techniques). It is possible that a reporting increase in the national veterinary registration systems of each country acts as a signal to activate further scientific research of specific diseases.

Collocation statistics retrieved with KH coder (Supplementary material) contributed to the objective interpretation of selected words representing topics. Four of these words, those of "behavior", "resistance", "student" and "surgery", were selected as representatives of the topics that were firmly extracted during the three-decade-period and displayed changes intothe topic ranking of each decade (animal behavior, antimicrobial resistance, students and surgeries respectively). Words appearing before or after these selected terms are directly depended on them (they display greater weight). It is more probable to see the word "veterinary" and "female" one word left from "student" and word "educ" one word right from "student" during 1991 - 2000. All these words combined thus indicate that they are connected under this point of view: veterinary student's education or female student's education. During 2001 - 2010 other words co-appear with the word student such as medicine, learning, graduate, interest, evaluation, training, participation or experience. All these words combined thus indicate that they are connected under the following point of view: veterinary student's learning, veterinary student's training etc. During 2011 - 2020 new words appear close to the term student: performance, perception, studies, assessment, training, attitude, experience or improvement. All these words combined indicate that they are connected under this point of view: veterinary student's performance or veterinary student's perception etc. It is possible that a shift in the scientific thinking occurred during 1991 - 2020 from the simple perspective of veterinary student's education (Heath et al., 1996) to a more competitive aspect including performance, assessment of their experience etc (Zenner et al., 2005).

It is more probable to encounter the words "veterinary", "animal", "invas", "cardiovascular" one word left from "surgeri" and the words "procedur" and "depart" one word right from "surgeri" during 1991 - 2000. All these words combined thus suggest that they are connected under this point of view:veterinary surgeries or invasive surgeries or surgeries procedure. During 2001 - 2010 new words appear close to the word surgeries such as "dure", "clinic", "laparoscop", "abdomin", "convent", "hors", "colic" etc. All these words combined suggest that they are connected under this point of view:veterinary surgeries or laparoscopic surgeries or conventional surgeries etc. During 2011 - 2020 new words appear close to the word surgeries such as "dog", "perform", "course", "spinal", "open". All these words combined suggest that they are connected under this point of view:surgeries duration or veterinary surgeries or surgeries performance or dog surgeries etc. It is possible that the scientific field of surgeries is constantly evolving and dealing with ever-changing topics.

It is more probable to see the words "problem"," animal", "therapy", "pet", "pharmacotherapy" close to the word "behavior" during 1991 - 2000. New words appear close to "behavior" during 2001 - 2010 those of "medicin", "veterinary" and "cours". During 2011 - 2020 words that appear close to "behavior" include "problem", "change", "relat", "dog" and "intervent". All these words combined suggest that they are connected under this point of view: animal behavior problem during 1991 - 2020, but with some differentiations between decades from behavior therapy/pharmacotherapy to behavior course and then to behavior change. It is possible that scientists make efforts to intervene into the animal behavior problems through therapies. However during 2001 - 2020 an effort is made to attribute biological interpretations into the aggressive behavior of animals. Some types of agitated behavior indeed have a strong genetic basis (Grandin and Deesing, 2014).

It is more probable to see the words "antibiot", "bacteri" and "pathogen" during 1991 - 2000, the words "antimicrobi", "isol" and "methicilin" during 2001 - 2010 and the words "antimicrobi", "antibiot" and "multidrug" during 2011 - 2020close to the term "resist". All these words combined suggest that they are connected under this point of view: antibiotic resistance or antimicrobial resistance or isolates of bacteria to study the antibiotic resistance etc. Many of the terms closely present with "resist" are common between the three decades. However it is possible that each decade is characterized by different priorities regarding research on antibiotic resistance as new terms stand out during the 2001 - 2020 period, those of multidrug, *Staphylococcus aureus* and gene nevertheless this is not so evident and cannot be distinctly supported by the specific results.

Previous text mining works have revealed a common number of terms also described in the present study. The subject of public awareness regarding the way in which farm animals are kept, the use of antimicrobials to increase animal performance and animal welfare have been reported before (Contiero et al., 2019). The subject of infectious diseases transmitted from animals to humans (or the opposite) and antimicrobial resistance due to prescriptions and results in human health have also been reported in the past (Lustgarten et al., 2020).

The validation of the LDA model carried out with several ways to verify that the accuracy in extracting the same topics was good at a satisfactory level. Each time the LDA model is applied a number of less frequent and hard to interpret topics is extracted (Nanni, 2017). We have to take into consideration that a part of the topics that LDA failed to correctly extract is these nonessential topics. On the other hand the LDA model is suitable in correctly identifying the most important of them over a large period of time. More specialized queries in the future may contribute to revealing possible trendsof less studied topics of the scientific literature.

CONCLUSIONS

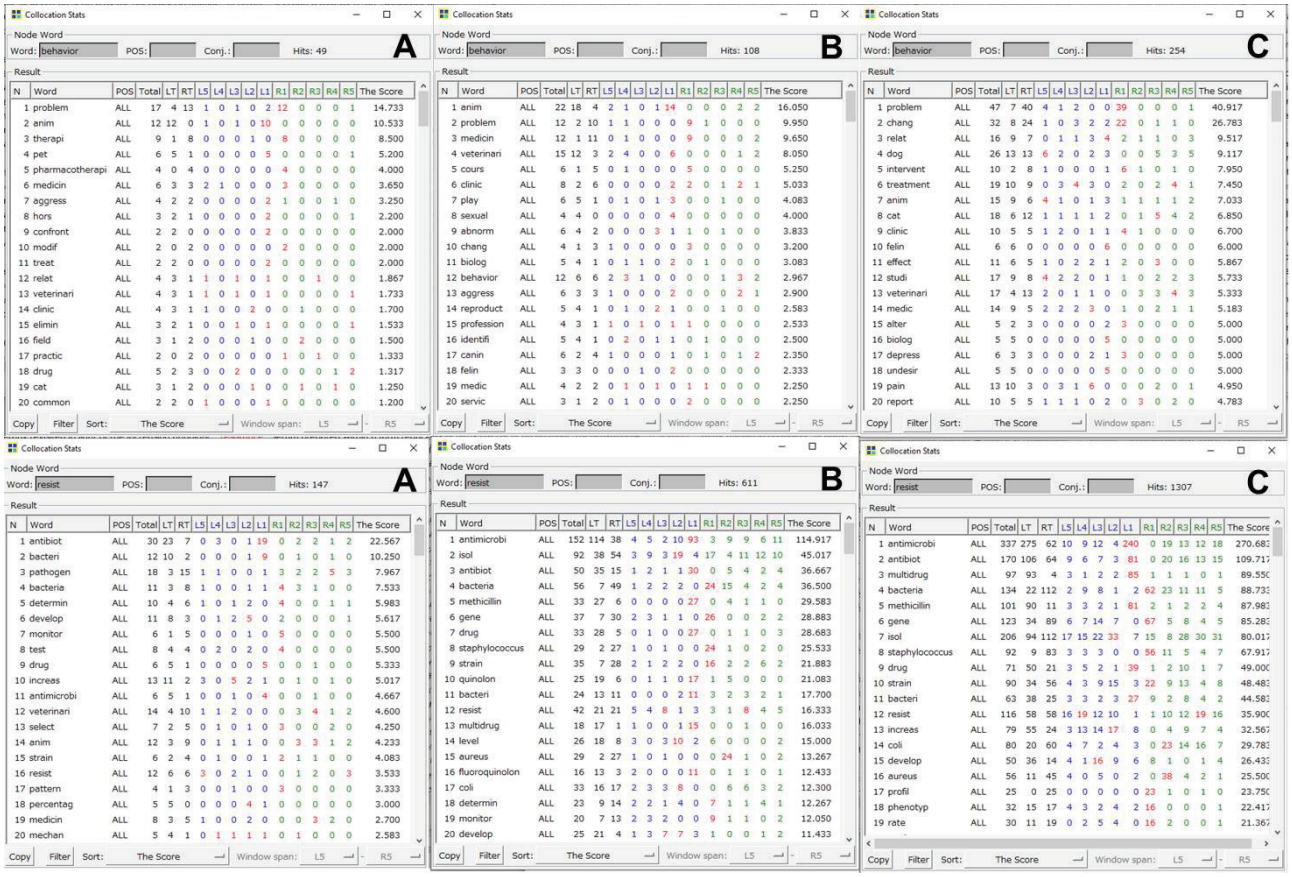
LDA managed to reveal the most frequent topics of three continuous decades. The number of topics extracted during each period increases proportional to the volume of scientific literature. Differences throughout decades occur and may reflect perceptions of researchers. Topics related with veterinary students and antibiotic resistance are probably incorporated into the main corpus of literature during the 2001 - 2020 period while topics related with animal behavior were probably enriched with a variety of new subfields not recorded in the past. Quantitative literature research is an appropriate tool in identifying trends in topics.

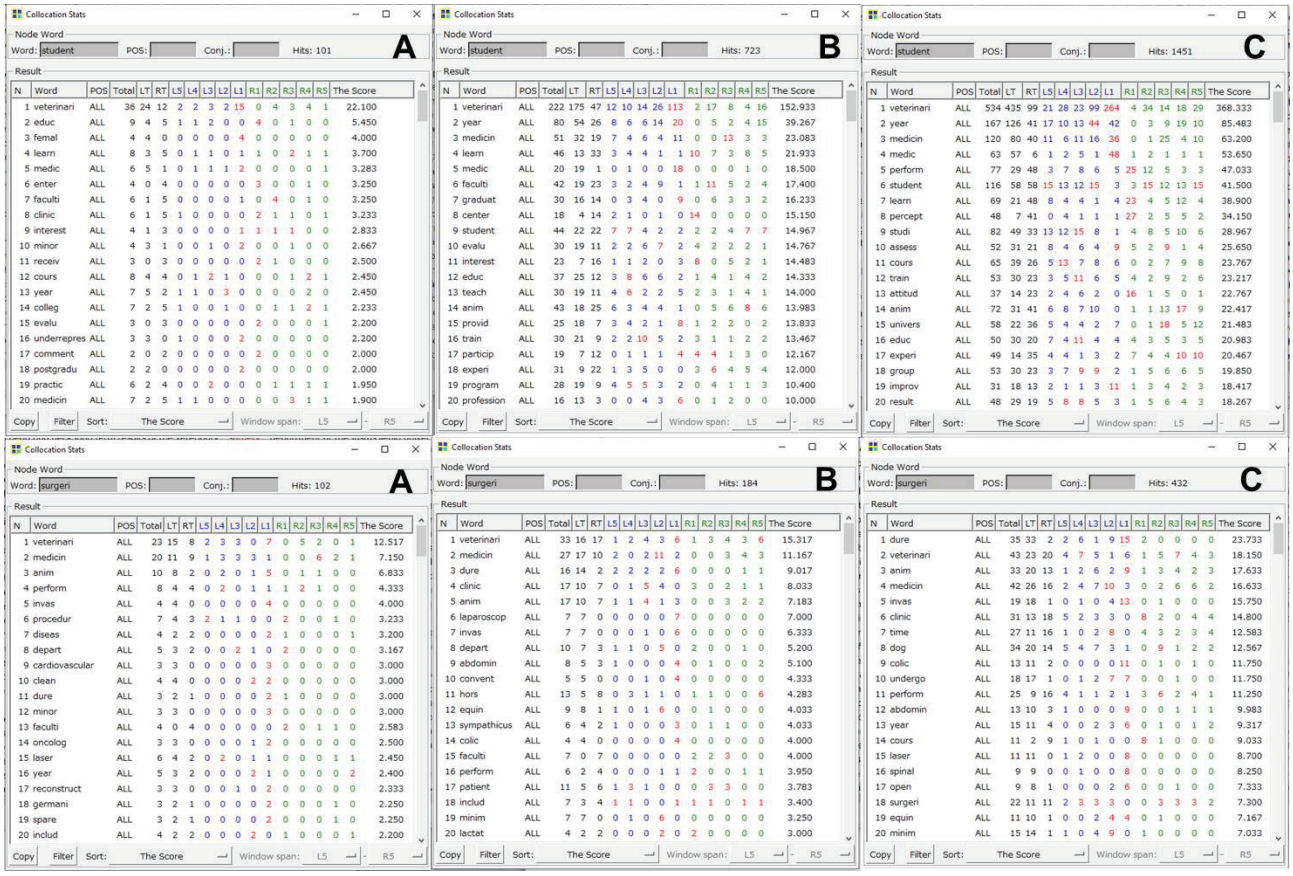
ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers for their useful comments.

SUPPLEMENTARY MATERIAL

Collocation statistics retrieved with KH coder. Word frequencies appearing before (L) and after (R) a node word are presented. The list shows the words that directly depend on the node word. Words that appear closer to the node words display greater weight, thus a higher Score. Four examples of strongly connected words with "Veterinary", persistently present in all the 1991 - 2020 period are shown here those of "behavior", "resistance", "student" and "surgery".[A = 1991 - 2000, B = 2001 - 2010, C = 2011 - 2020]





List of the topics extracted during 1991 - 2020 with the use of the LDA model. Topics are presented in descending order with the use of entropy and burstiness as indexes to measure the most common words and topics of literature. Abbreviations of selected topics were used to help track their order during each time period.

	1991 - 2000		2001 - 2010		2011 - 2020			
1	dog|cat|case|clinic|examin	CE	veterinari|student|medicin|educ|program	ST	student|veterinari|medicin|studi|survey	ST	imag|mri|radiolog|radiograph|tomographi	
2	studi|data|age|veterinari|sampl		clinic|examin|case|veterinari|medicin	CE	anim|veterinari|medicin|health|human		ultrasound|arteri|pregnanc|fetal|doppler	
3	veterinari|anim|medicin|veterinarian|develop		veterinari|medicin|use|practic|inform		diet|group|feed|supplement|fed		extract|concentr|ethanol|aqueous|quail	
4	hors|anthelmint|modifi|action|structur		veterinari|human|medicin|treatment|diseas		veterinari|medicin|diseas|human|patient		surgeri|surgic|procedur|perform|complic	SU
5	techniqu|surgeri|medicin|veterinari|surgic	SU	resist|antimicrobi|isol|strain|antibiot	AR	case|dog|veterinari|clinic|report	CE	heart|cardiac|thorac|pulmonari|lung	
6	measur|method|differ|bone|plasma		anim|health|veterinari|diseas|develop		resist|isol|gene|strain|antimicrobi	AR	platelet|teg|activ|coagul|blood	
7	student|veterinari|teach|educ|cours	ST	infect|vaccin|diseas|pcr|mrsa	VA	studi|dog|use|result|medicin		tumor|cancer|cell|tumour|lymphoma	TU
8	test|valu|method|procedur|diagnost		cell|tumor|tumour|dog|carcinoma	TU	resist|antimicrobi|infect|isol|bacteri	AR	joint|lame|articular|ruptur|osteoarthr	
9	veterinari|medicin|school|year|found		hors|surgic|case|fractur|surgeri	SU	cell|stem|tissu|mscs|msc		thyroid|test|blade|perfus|ablat	
10	concentr|studi|group|serum|milk	DA	dose|administr|concentr|oral|mgkg	DO	milk|cow|product|digest|dri	DA	turtl|eye|carolina|block|north	
11	renal|function|failur|transplant|increas		dog|blood|method|samp|concentr		veterinari|review|report|studi|clinic	CE	und|der|die|des|les	
12	resist|antibiot|antimicrobi|infect|strain	AR	cell|immun|respons|express|activ		abdomin|case|examin|surgeri|ultrasound	SU	bird|avian|exot|pet|mare	
13	drug|veterinari|medicin|use|effect		imag|use|techniqu|digit|ultrasound		drug|effect|activ|anti|use		eye|ocular|corneal|strain|cornea	
14	behavior|treat|reduct|bromid|aggress	AB	studi|dog|group|veterinari|result		anim|parasit|infect|speci|case		fish|aquacultur|aquat|catfish|salmon	
15	system|imag|use|comput|radiat		blood|arteri|oxygen|pressur|use	BP	express|activ|protein|increas|level		macaqu|rhesus|otter|cryptococcus|eleph	
16	day|activ|effect|level|agent		dog|heart|breed|system|wave		treatment|hors|dog|clinic|case		burnout|capon|dystonia|marten|movement	
17	blood|pressur|measur|use|arteri	BP	veterinari|medicin|faculti|school|cpcr		veterinari|univers|medicin|parasitolog|research		larva|collar|tick|eat|meat	
18	test|method|antibodi|infect|diseas		product|feed|goat|meat|extract		dog|cat|studi|age|year		epilepsi|seizur|epilept|behavior|noninferior	AB
19	muscl|state|cow|liver|bird		renal|creatinin|kidney|serum|control		group|anesthesia|effect|anesthet|pain		cup|night|bat|behaviour|vns	
20	plant|tradit|diseas|medicin|vaccin	VA	case|rabi|report|number|anim		product|anim|feed|effect|associ		iron|hepcidin|liver|gli|binder	
21	min|group|vaccin|report|effect	VA	pet|anim|behavior|problem|dog	AB	dose|concentr|mgkg|administr|drug	DO	sorghum|cent|kafirin|starch|outlet	
22	bacteri|chang|calculi|bacteriolog|specimen		cow|interview|milk|dairi|herd	DA	spinal|neurolog|brain|cord|nerv		grade|iap|metronidazol|fdk|gastric	
			diabet|glucos|allergen|mellitus|colt		antimicrobi|farm|use|pig|antibiot		brows|bag|indigen|tree|fluid	
					bone|fractur|limb|month|materi		pdt|pvas|chimpanze|anxieti|criteria	
					pcr|dna|sequenc|detect|genet		stk|cmt|disbud|horn|hot	
					infect|test|protein|serum|antibodi		refractomet|usg|urin|furosemid|pcd	
					blood|serum|concentr|biochem|protein		biomark|cardiac|ctni|assay|myocardi	
					vaccin|peptid|immun|respons|adjuv	VA	empathi|rural|shortag|counti|storm	
					diseas|infect|virus|rabi|vaccin	VA	pfa|probiot|lactobacillus|prob|caeca	
					wound|heal|skin|laser|honey		bone|miner|bmd|densiti|dxa	

Abbreviations

Clinical Examination	**CE**	**Tumor**	**TU**
Surgeries	**SU**	**Vaccinations**	**VA**
Students	**ST**	**Dosage**	**DO**
Antimicrobial Resistance	**AR**	**Dairy**	**DA**
Animal Behavior	**AB**	**Blood Pressure**	**BP**

REFERENCES

Anholt RM, Berezowski J, Jamal I, Ribble C Stephen C (2014) Mining free-text medical records for companion animal enteric syndrome surveillance. Prev Vet Med 113(4): 417-422.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan): 993-1022.

Bollig N, Clarke L, Elsmo E, Craven M (2020). Machine learning for syndromic surveillance using veterinary necropsy reports.PloS one 15(2): p.e0228105.

Boyd-Graber J, Mimno D, Newman D (2014) Care and feeding of topic models: Problems, diagnostics, and improvements. In: Handbook of mixed membership models and their applications Chapman and Hall/CRC: pp 225-255.

Christopher MM, Marusic A (2013) Geographic trends in research output and citations in veterinary medicine: insight into global research capacity, species specialization, and interdisciplinary relationships. BMC Vet Res 9(1): 115.

Contiero B, Cozzi G, Karpf L, Gottardo F (2019) Pain in Pig Production: Text Mining Analysis of the Scientific Literature. J Agric Environ Ethics 32: 401-412.

Csoltova E, Martineau M, Boissy A, Gilbert C (2017) Behavioral and physiological reactions in dogs to a veterinary examination: Owner-dog interactions improve canine well-being. Physiol Behav 177: 270-281.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, J Mach Learn Res 14(Aug): 2349−2353.

Dórea FC, Lindberg A, Elvander M (2015) Veterinary syndromic surveillance in practice: costs and benefits for governmental organizations. Infect Ecol Epidemiol 5(1): 29973.

Ellis J, Ellis B, Velez-Estevez A, Reichel M, Cobo M (2020) 30 years of parasitology research analysed by text mining. Parasitology 1: 1-58.

Grandin T, Deesing MJ (2014) Genetics and behavior during handling, restraint, and herding. In: Genetics and the behavior of domestic animals 2nd ed, Academic Press, New York:pp 115-158.

Heath TJ, Lanyon A, Lynch-Blosse M (1996) A longitudinal study of veterinary students and recent graduates: 3. Perceptions of veterinary education. Aust Vet J 74(4): 301-304.

Higuchi K (2016) KH Coder 3 reference manual. Kioto (Japan): Ritsumeikan University.

Jones BD, Byrnes MK, Jones MW (2019) Validation of the MUSIC Model of Academic Motivation Inventory: Evidence for use with veterinary medicine students. Front Vet Sci 6: 11.

Jones-Diette JS, Dean RS, Cobb M, Brennan ML (2019) Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK. Prev Veterinary Med 167: 61-67.

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1): 79-86.

Lin H, Sheu PC, Tsai JJP, Charles CN, Wang, CYC (2020) Text mining in a literature review of urothelial cancer using topic model. BMC Cancer 20: 462.

Lustgarten JL, Zehnder A, Shipman W, Gancher E, Webb TL (2020) Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives-a joint paper by the Association of Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA). JAMIA Open3(2): 306-317.

Mich PM, Hellyer PW, Kogan L, Schoenfeld-Tacher R (2010) Effects of a pilot training program on veterinary students' pain knowledge, attitude, and assessment skills. J Vet Med Educ 37(4): 358-368.

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006). Yale: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining: pp. 935-940.

Nanni F (2017) The Web as a Historical Corpus: Collecting, Analysing and Selecting Sources on the Recent Past of Academic Institutions. (Doctoral dissertation, alma).Universita di Bologna 242p.

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. J Artif Intell Res 11: 169-198.

Packer RM, Law TH, Davies E, Zanghi B, Pan Y, Volk HA (2016) Effects of a ketogenic diet on ADHD-like behavior in dogs with idiopathic epilepsy. Epilepsy Behav 55: 62-68.

Queirós A, Faria D, Almeida F (2017) Strengths and limitations of qualitative and quantitative research methods. Eur J Educ Stud 9: 369-387.

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J ComputAppl Math 20: 53-65.

Sahadevan S, Hofmann-Apitius M, Schellander K, Tesfaye D, Fluck J, Friedrich CM (2012) Text mining in livestock animal science: introducing the potential of text mining to animal sciences. J Anim Sci 90(10): 3666-3676.

Shannon CE (1948) A mathematical theory of communication. The Bell system technical journal 27(3): 379-423.

Toutain PL, Ferran AA, Bousquet-Melou A, Pelligand L, Lees P (2016) Veterinary medicine needs new green antimicrobial drugs. Front Microbiol 7: 1196.

Tynes VV, Sinn L (2014) Abnormal repetitive behaviors in dogs and cats: a guide for practitioners. Vet Clin North Am Small Anim Pract 44(3): 543-564.

Van der Waal K, Morrison RB, Neuhauser C, Vilalta C, Perez AM (2017) Translating big data into smart data for veterinary epidemiology. Front Vet Sci 4: 110.

Welsh CE, Parkin TDH, Marshall JF (2017) Use of large-scale veterinary data for the investigation of antimicrobial prescribing practices in equine medicine. Equine Vet J 49(4): 425-432.

Zenner D, Burns GA, Ruby KL, De Bowes RM, Stoll SK (2005) Veterinary students as elite performers: preliminary insights. J Vet Med Educ 32(2), 242-248.

Zhang Y, Tao J, Wang J, Ding L, Ding C, Li Y, Qichao Zhou, Dunhai Li, Zhang H (2019) Trends in diatom research since 1991 based on topic modeling. Microorganisms 7(8): 213.