Text mining for identification of biological entities related to antibiotic resistant organisms

Kelle Fortunato Costa[1], Fabrício Almeida Araújo[2,3], Jefferson Morais[4], Carlos Renato Lisboa Frances[1] and Rommel T. J. Ramos[5]

[1] Programa de pós-graduação em Engenharia Elétrica, Universidade Federal do Pará, Belém, Pará, Brazil
[2] Biological Science Institute, Universidade Federal do Pará, Belém, Pará, Brazil
[3] Universidade Federal Rural da Amazônia, Belém, Pará, Brazil
[4] Universidade Federal do Pará, Belém, Pará, Brazil
[5] Biological Science Institute, Universidade Federal do Para, Belém, Pará, Brazil

ABSTRACT

Antimicrobial resistance is a significant public health problem worldwide. In recent years, the scientific community has been intensifying efforts to combat this problem; many experiments have been developed, and many articles are published in this area. However, the growing volume of biological literature increases the difficulty of the biocuration process due to the cost and time required. Modern text mining tools with the adoption of artificial intelligence technology are helpful to assist in the evolution of research. In this article, we propose a text mining model capable of identifying and ranking prioritizing scientific articles in the context of antimicrobial resistance. We retrieved scientific articles from the PubMed database, adopted machine learning techniques to generate the vector representation of the retrieved scientific articles, and identified their similarity with the context. As a result of this process, we obtained a dataset labeled "Relevant" and "Irrelevant" and used this dataset to implement one supervised learning algorithm to classify new records. The model's overall performance reached 90% accuracy and the f-measure (harmonic mean between the metrics) reached 82% accuracy for positive class and 93% for negative class, showing quality in the identification of scientific articles relevant to the context. The dataset, scripts and models are available at https://github.com/engbiopct/TextMiningAMR.

Subjects Bioinformatics, Microbiology, Computational Science, Data Mining and Machine Learning
Keywords Antimicrobial resistance, Biological literature, Text mining, Machine learning

INTRODUCTION

Antibiotics are the most successful drugs of the last 100 years, responsible for saving countless lives and enabling modern medical procedures that would otherwise be unthinkable. However, all antibiotics derived from secondary or fully synthetic microbial metabolism products are subject to resistance (*Wright, 2011*).

Antimicrobial resistance (AMR) has been increasingly recognized as an important public health problem worldwide, considering that infections caused by

multidrug-resistant organisms (MDR) result in a significant increase in mortality and cause a tremendous economic burden (*Tran, Munita & Arias, 2015*). In addition to the costs of rising hospital admission rates, it is estimated that by 2,050 there will be an economic loss of $100 trillion in global antibiotic production (*Review on Antimicrobial Resistance, 2016*).

In recent years, the scientific community has intensified efforts to combat this problem by making available a wide range of public databases specific to AMR, such as: National Database of Antibiotic Resistant Organisms (NDARO) (*Annual Reports for NLM Program and Services, 2016*), Comprehensive Antibiotic Resistance Database (CARD) (*Alcock et al., 2020*), Resfinder (*Zankari et al., 2012*), ResfinderFG (*Munk et al., 2018*), Resfams (*Gibson, Forsberg & Dantas, 2015*), Antibiotic Resistance Genes Database (ARDB) (*Liu & Pop, 2009*), MEGARes (*Lakin et al., 2017*), Antibiotic Resistance Gene Annotation (ARG-ANNOT) (*Gupta et al., 2014*), Mustard (*Ruppe et al., 2019*), Functional Antibiotic Resistance Metagenomic Element (FARME database) (*Wallace et al., 2017*), SARG (v2) (*Yin et al., 2018*), Lahey list of β-lactamases (*Bush & Jacoby, 2010*), β-Lactamase Database (BLDB) (*Naas et al., 2017*), Lactamase Engineering Database (LacED) (*Thai, Bos & Pleiss, 2009*; *Thai & Pleiss, 2010*), Comprehensive β-Lactamase Molecular Annotation Resource (CBMAR) (*Srivastava et al., 2014*), among others, which are frequently used as reference databases, with gene sequences related to antimicrobial resistance and metadata that enrich the characterization of sequences. Due to the increase in the volume of literature related to biological and health sciences, the curation process has become challenging for researchers and biocurators who use these databases as a source of research, mainly due to the time required to locate relevant information about biological entities related to antibiotic-resistant organisms. Even queries in specialized databases in biomedical literature such as PubMed (scientific and medical abstracts/citations), PubMed Central (full-text journal articles), NLM Catalog (index of NLM collections), Books (books and reports), and MeSH (ontology used for PubMed indexing) (*Sayers et al., 2019*), tends to make document selection difficult due to a large amount of retrieved items.

In this context, the adoption of text mining (TM) techniques are viable alternatives (*Wei, Kao & Lu, 2013*), as they can help in different stages of the standard biocuration workflow. According to (*Hirschman et al., 2012*), the steps are:

- Selection: search for articles relevant to the curation.
- Identification and standardization of bioentities: detection of mentions of bioentities relevant to the curatorship; for example, genes, proteins, or small molecules, linked to unique identifiers from databases such as UniProt, EntrezGene, or ChEBI.
- Detection of annotation events: identification and encoding of events, such as descriptions of protein-protein interactions, characterizations of gene products in terms of cellular location, molecular function, involvement in the biological process and phenotypic effect.
- Evidence qualifier association: association of experimental evidence that supports the annotation event performed due to biocuration efforts.

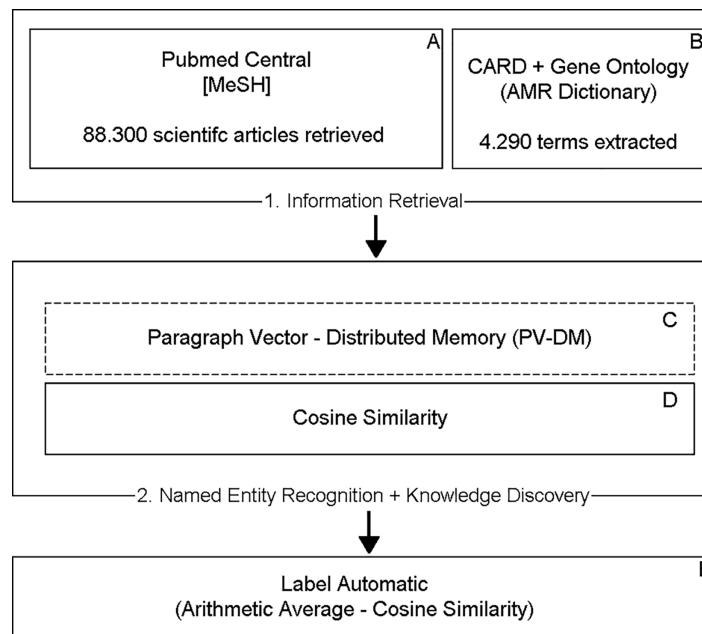- Completion and verification of the database record.

TM technologies combine knowledge resources such as controlled vocabularies, taxonomies, and ontologies with linguistic analysis and machine learning to deal with language variations and extract not only terms from the text but also relationships between terms (*Chaix et al., 2018*).

In the last decade, several applications in the biomedical area were developed (*Fleuren et al., 2011*; *Fontelo, Liu & Ackerman, 2005*; *Perez-Iratxeta, Bork & Andrade, 2001*; *Lewis et al., 2006*; *Fontaine et al., 2009*; *States et al., 2009*; *Huang et al., 2013*; *Hokamp & Wolfe, 2004*; *Plikus, Zhang & Chuong, 2006*; *Becker et al., 2003*; *Douglas, Montelione & Gerstein, 2005*; *Brancotte et al., 2011*; *De et al., 2010*; *Smalheiser, Zhou & Torvik, 2008*; *Chen & Sharp, 2004*; *Li et al., 2013*; *Glynn, Kerin & Sweeney, 2010*; *Xuan et al., 2007*; *Giglia, 2011*; *Tsuruoka et al., 2011*; *Fernandez, Hoffmann & Valencia, 2007*; *Raja, Subramani & Natarajan, 2013*; *Pafilis et al., 2009*; *Rebholz-Schuhmann et al., 2008*; *Plake et al., 2006*; *Soldatos et al., 2010*; *Franceschini et al., 2013*) using one or more of the following TM steps: (i) retrieving textual resources relevant to a particular subject of interest, a process known as information retrieval (IR), (ii) detect the occurrence of specific keywords of interest and the relationships between these keywords and (iii) infer new relationships based on known facts, and this step is called knowledge discovery (KD) (*Fleuren & Alkema, 2015*).

One of the most used machine learning techniques in knowledge discovery, especially in document screening, is text classification. However, supervised classification requires the prior labeling of a training set, a non-trivial task for human curators, as it requires a lot of time and effort.

In this sense, *Suomela & Andrade (2005)* propose a methodology for automatic (binary) classification of large volumes of data, adopting the word counting technique, known as the bag of words (BOW) (*Manning, Raghavan & Schütze, 2008*), a textual representation that composes the vector space model (*Salton, Wong & Yang, 1975*), where documents are converted into vectors of words. A weighting scheme is applied to each word, which can be a simple word count or a metric such as Term Frequency-Inverse Document Frequency (TF-IDF) (*Manning, Raghavan & Schütze, 2008*; *Paik, 2013*; *Chen, 2017*) and based on the arithmetic mean of the weights of these words, text summaries are classified as relevant or irrelevant.

This methodology was implemented from specific abstracts of an area of interest, extracted from Pubmed, and inspired the development of the MedlineRanker application (*Fontaine et al., 2009*) and served as a baseline for this study, which instead of the bag of words, adopts a representation approach based on neural networks (*Bengio et al., 2006*), called Paragraph Vector-Distributed Memory (PV-DM) (*Le & Mikolov, 2014*), capable of revealing semantic characteristics between documents, a property that makes this approach useful for many natural language processing (NLP) tasks and justifies its wide use in works involving natural language understanding (*Collobert & Weston, 2008*; *Zhila et al., 2013*), machine translation (*Mikolov, Le & Sutskever, 2013*; *Zou et al., 2013*), image comprehension (*Frome et al., 2013*) and relational extraction (*Socher et al., 2013a*).

Figure 1 Proposed TM model. Steps (A) and (B) include retrieving the information. Steps (C) and (D) include the recognition of entities and the discovery of knowledge, resulting in a metric (cosine similarity) responsible for determining the binary classification performed in step (E).

Full-size ⬚ DOI: 10.7717/peerj.13351/fig-1

The study of antimicrobial resistance genes is essential to public health. However, there is challenging to handle and extract the amount of available data of scientific and medical manuscripts from public databases without computational methods. Thus, this work proposes an unsupervised learning-based TM approach for ranking the relevance of articles on AMR context to generate a set of training, accurate enough to generalize new data, maximizing the efficiency of the supervised classifiers.

MATERIALS AND METHODS

Labeling pipeline

Figure 1 shows the TM steps implemented in this work in order to label the data.

Information retrieval

An Application Programming Interface (API) was implemented to retrieve a collection of relevant articles in the Drug Resistance, and Microbial domain through the Pubmed Central (PMC) database, which contains free full-text files of the Library's of Medicine and the US National Institutes of Health (NIH/NLM) biomedical literature are available. In the API, the E-Search and E-Fetch tools from the E-utilities package were used, which provide a structured interface for accessing the Entrez system, the NCBI database system, which currently includes 38 databases, covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and biomedical literature (NCIBI Homepage, https://www.ncbi.nlm.nih.gov/books/NBK3827/). Table 1 presents the set of parameters incorporated into the search (E-Search).

Table 1 Parameters *E-Search* PubMed central.

Parameters	Value
URL	https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pmc&term=%22drug%20resistance,%20microbial%22[MeSH%20Terms]%20OR%20(%22drug%22[All%20Fields]%20AND%20%22resistance%22[All%20Fields]%20AND%20%22microbial%22[All%20Fields])%20OR%20%22microbial%20drug%20resistance%22[All%20Fields]%20OR%20(%22drug%22[All%20Fields]%20AND%20%22resistance%22[All%20Fields]%20AND%20%22microbial%22[All%20Fields])%20OR%20%22drug%20resistance,%20microbial%22[All%20Fields])%20AND%20%22open%20access%22[filter]&retmax=10000
db	PMC (full text articles)
Term	("drug resistance, microbial"[MeSH Terms] OR ("drug"[All Fields] AND "resistance"[All Fields] AND "microbial"[All Fields]) OR "microbial drug resistance"[All Fields] OR ("drug"[All Fields] AND "resistance"[All Fields] AND "microbial"[All Fields]) OR "drug resistance, microbial"[All Fields])
Free text articles	Open access

Table 2 Parameters Doc2Vec algorithm.

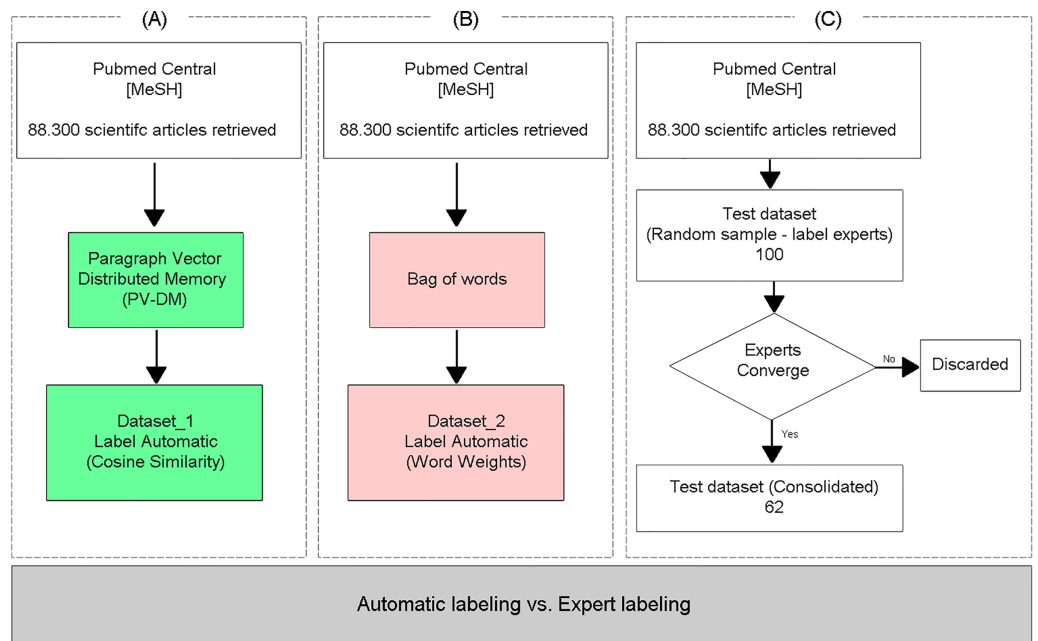| Parameters	Value	Description
vector_size *(int, optional)*	300	Dimensionality of the feature vectors
alpha *(float, optional)*	0.025	The initial learning rate
min_alpha *(float, optional)*	0.00025	Learning rate will linearly drop to *min_alpha* as training progresses
workers *(int, optional)*	18	Use these many worker threads to train the model (= faster training with multicore machines)
min_count *(int, optional)*	3	Ignores all words with total frequency lower than this
epochs *(int, optional)*	30	Number of iterations (epochs) over the *corpus*
dm *({1,0}, optional)*	1	Defines the training algorithm. If *dm* = 1, 'distributed memory' (PV-DM) is used. Otherwise, *distributed bag of words* (PV-DBOW) is employed

The terms of the MeSH hierarchy were adopted for antimicrobial resistance, considering that the terms MeSH is a controlled vocabulary of biomedical terms whose elements are assigned to a document by indexers (specialists in biomedical subjects) based on its context. They contain high-density document-wide information that cannot be deduced from the title or abstract that PubMed returns using keywords (NLM: Medical Subject Headings (MeSH)).

Then, a list of PMCIDs (unique identifiers provided by PubMed Central to each document) is generated to be used to access the full texts of articles through the E-Fetch utility.

Named entity recognition and knowledge discovery

The entity recognition and knowledge discovery process consist of two steps: in the first step (Fig. 1C), the Doc2Vec unsupervised learning algorithm from the Gensim library was used, which implements the Paragraph Vector–Distributed Memory model, to obtain the embedding of the retrieved documents (dense representation of a sequence of words). Table 2 displays the parameters used in the algorithm.

In the second step (Fig. 1D), the pre-trained model was used, capable of predicting whether a set of documents {Doc_1, Doc_2, Doc_3…Doc_n} belong to the context of a central document Doc_0, to infer the similarity of the documents to the AMR context,

Figure 2 Evaluation of the proposed method.   Full-size ⬚ DOI: 10.7717/peerj.13351/fig-2

represented by 4,290 terms extracted from CARD and the Gene Ontology Database (*Ashburner et al., 2000*), calculating the cosine distance (*Nguyen & Bai, 2011*) between them. The resulting value varies in the range between −1 and 1 where the higher the number, the greater the similarity with the context.

Finally, each of the scientific articles was automatically labeled (Fig. 1E), considering the arithmetic average of the cosine similarity, values above the average were defined as relevant and below the mean as irrelevant.
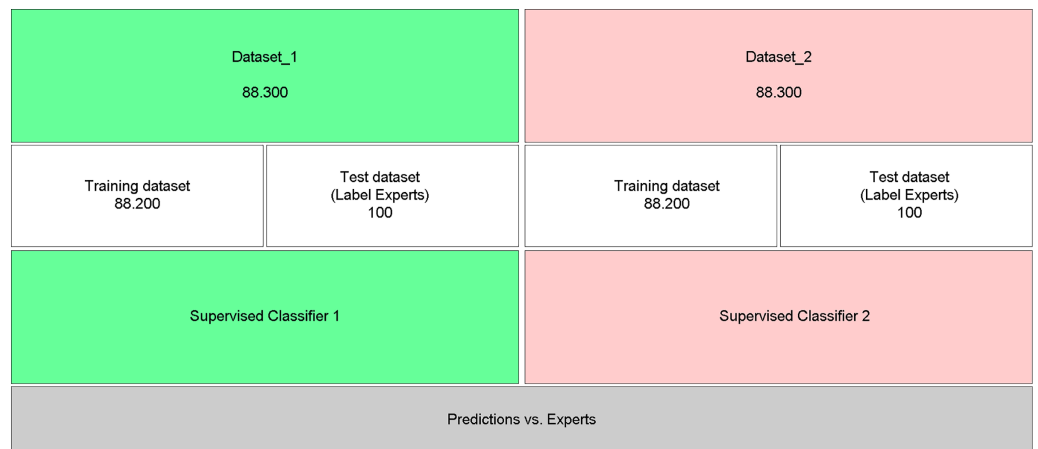
Evaluation of the proposed method

To evaluate the classification performance of the proposed TM approach (Fig. 2A), the same dataset was labeled using the Bag of words text representation model (Fig. 2B), adopting the specific vocabulary for AMR as a vector of features, and the documents were classified as relevant or irrelevant through the arithmetic mean of the weights of the words, obtained with the TF-IDF weighting method, similar to the methodology proposed in *Suomela & Andrade (2005)*. Finally, the two (automatically) labeled databases were compared with a test database labeled by experts (Fig. 2C), who independently labeled the articles as relevant or irrelevant. Only the samples where the three experts converged on the labels were included in the test dataset.

Predictions with automatically labeled data

To evaluate the efficiency of the proposed approach, which uses neural embeddings for labeling, the generated datasets (Fig. 3: Dataset_1 and Fig. 3: Dataset_2) were used as input to the supervised classifier SVM (*Boser, Guyon & Vapnik, 1992*; *Drucker et al., 1997*). The classification performance was evaluated through the analysis of the Precision, Recall,

Figure 3 Performance of predictions with automatically labeled data.
Full-size ▣ DOI: 10.7717/peerj.13351/fig-3

Accuracy, and F-Measure metrics (*Grandini, Bagli & Visani, 2020*), calculated from the test base labeled by experts and not used to train the SVM models (Fig. 3).

For the SVM classifier, the feature/attribute vector (AMR vocabulary with 4,290 words) was weighted using the TF-IDF technique, and cross-validation, with 5-folds (default value of the adopted algorithm), was applied to explore the combination of parameters for determining the best model, as the effectiveness of the SVM depends on the kernel selection, which is the function that will be used by the algorithm, in the margin parameter (C), which determines a balance between maximizing the margin and minimizing classification errors, and the Gamma parameter, when the chosen kernel is Gaussian (or RBF) (*Syarif, Prugel-Bennett & Wills, 2016*), adopted in this experiment.

Data avaliability

The dataset, script, and models generated by this work are available at https://github.com/engbiopct/TextMiningAMR, under CC BY 4.0 Copyright license, with information regards the workflow adopted, a short step-by-step guide to the readers reproduce this experiment and the complementary materials.

RESULTS AND DISCUSSION

Labeling

A collection of 88,300 scientific articles on antimicrobial resistance was retrieved from Pubmed Central, using the terms of reference of the MeSH (Medical Subject Headings, developed at the National Library of Medicine) hierarchical vocabulary referring to the AMR domain (https://meshb.nlm.nih.gov/record/ui?ui=D004352).

The retrieved dataset was submitted to the PV-DM text representation model, with the embedding of the documents obtained. The similarity of the documents to the AMR context (Table S1) was inferred, and the label "relevant" was automatically assigned (class 0) to all articles whose cosine distance value was equal to or greater than the arithmetic average of the cosine distances of the entire *corpus*, and the label "irrelevant" (class 1) to all

articles with a cosine distance value lower than that referred to average, resulting in 43,136 records labeled relevant and 45,164 labeled irrelevant (Table S2).

The same initial dataset was submitted to the Bag of words text representation model in order to obtain the weights of the words in the documents according to the AMR dictionary and thus automatically assign the label "relevant" (class 0) to all articles with weight equal to or greater than the arithmetic average of the weights, and the label "irrelevant" (class 1) to all articles with a value lower than the mean, resulting in 45,946 records labeled as relevant and 42,354 labeled as irrelevant (Table S3).

With the two labeled datasets, the results were compared with a test dataset labeled by experts, consolidated with a total of 62 scientific articles, 15 labeled as relevant and 47 labeled as irrelevant (Table S4).

In the comparison, the proposed method labeled 44 articles according to the experts, which represents 71% of hits in general, with 80% of hits for the relevant label and 68% of hits for the irrelevant label. As for the baseline method, there were only 26 labels according to the experts, which represents 42% of overall performance, with 66% of correctness for the relevant label and 34% of correctness for the irrelevant label.

The proposed approach presents a superior performance about the baseline, which despite its simplicity, efficiency, and often surprising precision, does not take into account the order and semantics of the words, that is, the distances between them. This means that the words "mighty", "strong" and "Paris" are equally distant. Although semantically, "powerful" is closer to "strong" than "Paris" (*Le & Mikolov, 2014*), characteristics present in the PV-DM model and fundamental human skills in manual data labeling tasks.

Classification

The two labeled datasets were submitted to the supervised SVM classifier, excluding the test dataset labeled by experts from training.

Figure 4 presents the confusion matrix of the SVM_1 classifier, trained with data from dataset 1 (Fig. 3: Dataset_1). There is a high degree of precision both in terms of true positives (relevant articles classified as relevant) and true negatives (non-relevant articles classified as non-relevant).
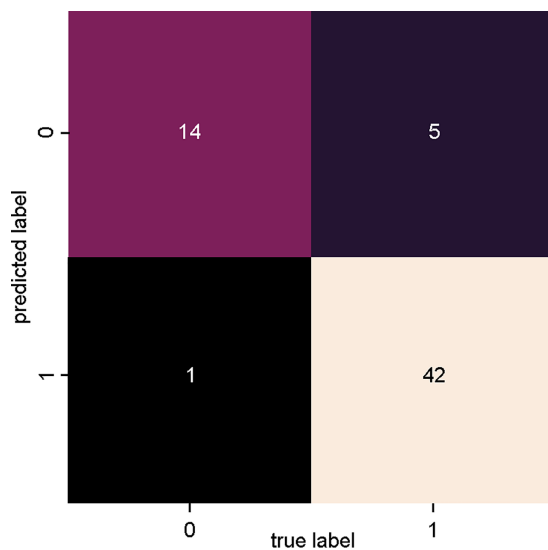
Only one article was incorrectly classified as relevant (false positive), and five articles were incorrectly classified as irrelevant (false negative).

Figure 5 presents the confusion matrix of SVM_2, trained with data from dataset 2 (Fig. 3: Dataset_2), where a lower degree of precision is observed for both true positives and true negatives in relation to SVM_1. With this classifier, however, there was an increase in the number of incorrect classifications, with 33 articles incorrectly classified as relevant (false positives) and six articles incorrectly classified as irrelevant (false negatives).

Table 3 presents the results of the evaluation metrics: precision, recall, accuracy and the f-measure, obtained based on the results of the confusion matrix of the two classifiers. The results of the SVM_1 classifier were superior to the SVM_2 classifier in all evaluated metrics.
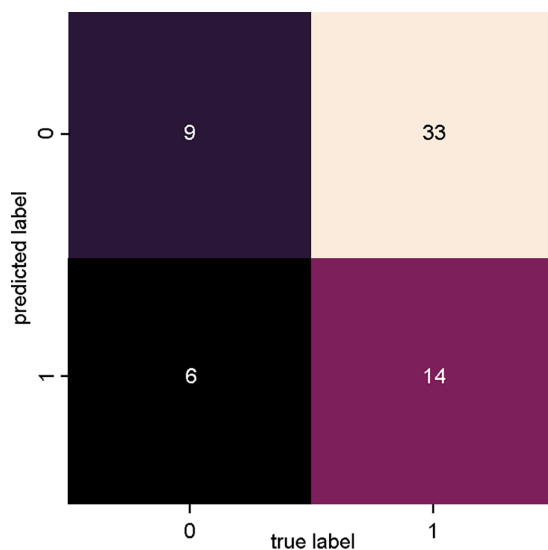
Accuracy, a metric that represents the overall performance of a model, reached 90% of accuracy and the f-measure, which is a harmonic average between the precision and recall

Figure 4 SVM classifier confusion matrix for dataset_1 (PV-DM).

Full-size ⬚ DOI: 10.7717/peerj.13351/fig-4

Figure 5 SVM classifier confusion matrix for dataset_1 (Bag of Words).

Full-size ⬚ DOI: 10.7717/peerj.13351/fig-5

Table 3 Classifier performance assessment.

	Class	Precision	Recall	F1-score	Support	Accuracy
SVM_1 (Doc2Vec+Mean)	0	0.74	0.93	0.82	15	0.90
	1	0.98	0.89	0.93	47	
SVM_2 (TF-IDF+Mean)	0	0.21	0.60	0.32	15	0.37
	1	0.70	0.30	0.42	47	

Fortunato Costa et al. (2022), *PeerJ*, DOI 10.7717/peerj.13351

9/16

Table 4 Results of labeling and classification steps *vs* experts.		
	Relevant (%)	**Irrelevant (%)**
Labeling		
Dataset_1	80	68
Dataset_2	66	34
Classification		
SVM_1	93	89
SVM_2	60	29

metrics and that can be used as a single measure to represent the quality in the text mining (*Rodriguez-Esteban, 2009*) reached 82% success rate for the positive class and 93% for the negative class.

The results show that the best performance was obtained with the database labeled by the PV-DM model.

Table 4 presents the percentage of correct predictions, both in the labeling and in the classification stage, in comparison with the data labeled by experts and validates the hypothesis that the use of Paragraph Vector, Distributed Representations of Sentences and Documents associated with similarity with a specific context is able not only to perform the binary classification of large volumes of data but also to optimize the percentage of hits of supervised classifiers. The SVM_2 classifier showed a reduction in the number of hits compared to the Labeling step, although we adopted the same attribute vector and the same representation in both experiments (bag of words, weighted with TF-IDF).

CONCLUSIONS

The proposed TM approach proved capable of identifying and prioritizing documents in the AMR context, as well as predicting the relevance of new documents in the same context. For this, we used the TM steps summarized in *Fleuren & Alkema (2015)*, plus some specifics of the proposal such as (1) use of the MeSH hierarchy; (2) use of full text; (3) use of domain-specific dictionaries (CARD and Gene Ontology), fundamental in this process, as it facilitated the detection of similarity of articles to the AMR context and (4) adoption of unsupervised learning for better representation of texts. Additionally, we submitted the labeled bases to the SVM classifier to evaluate their performance in comparison to the test base labeled by human experts.

The proposed approach efficiently identifies scientific articles relevant to the AMR context. Therefore, it is a valuable tool to automate information capture processes from robust bibliographic reference databases such as Pubmed Central, as well as to accelerate the screening of documents with biocuration potential, facilitating the other stages of the biocuration process.

This work presents a new set of pre-trained document embeddings in the AMR domain and a base labeled for relevance according to similarity to CARD and Gene Ontology, which, in future work, can be used as input to other algorithms of supervised learning and

for biocuration tools aimed at the identification and normalization of bioentities, detection of annotation events and filling out specific databases for AMR. This proposal is limited to the previous existence of an accurate database representing the main terms related to the target.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Grant Disclosures

Competing Interests

Rommel T. Ramos is an Academic Editor for PeerJ.

Author Contributions

- Kelle Fortunato Costa conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Fabrício Almeida Araújo analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Jefferson Morais analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Carlos Renato Lisboa Frances analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Rommel T.J. Ramos conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:
The data is available at GitHub: https://github.com/engbiopct/TextMiningAMR.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13351#supplemental-information.

REFERENCES

Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Mégane Bouchard, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran H-K, Werfalli RE,

Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur AG. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research* **48(D1)**:D517–D525 DOI 10.1093/nar/gkz935.

Annual Reports for NLM Program and Services. 2016. National Library of Medicine–NIH. *Available at* https://www.nlm.nih.gov/pubs/reports.html.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Becker KG, Hosack DA, Dennis G Jr, Lempicki RA, Bright TJ, Cheadle C, Engel J. 2003. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* **4**:2061 DOI 10.1186/1471-2105-4-61.

Bengio Y, Schwenk H, Senécal JS, Morin F, Gauvain JL. 2006. A neural probabilistic language models. In: *Innovations in Machine Learning*, Berlin: Springer, 137–186.

Boser BE, Guyon IM, Vapnik VN. 1992. A training algorithm for optimal margin classifiers, Pittsburgh: ACM Press, 144–152.

Brancotte B, Biton A, Bernard-Pierrot I, Radvanyi F, Reyal F, Cohen-Boulakia S. 2011. Gene List significance at-a-glance with GeneValorization. *Bioinformatics* **27(8)**:1187–1189 DOI 10.1093/bioinformatics/btr073.

Bush K, Jacoby GA. 2010. Updated functional classification of β-lactamases. *Antimicrobial Agents and Chemotherapy* **54(3)**:969–976 DOI 10.1128/AAC.01009-09.

Chaix E, Deléger L, Bossy R, Nédellec C. 2018. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology* **81(2)**:63–75 DOI 10.1016/j.fm.2018.04.011.

Chen C-H. 2017. Improved TF-IDF in big news retrieval: an empirical study. *Pattern Recognition Letters* **93(4)**:113–122 DOI 10.1016/j.patrec.2016.11.004.

Chen H, Sharp BM. 2004. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**:147 DOI 10.1186/1471-2105-5-147.

Collobert R, Weston J. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning*, New York: ACM, 160–167.

De S, Zhang Y, Garner JR, Alex Wang S, Becker KG. 2010. Disease and phenotype gene set analysis of disease-based gene expression in mouse and human. *Physiological Genomics* **42A(2)**:162–167.

Douglas SM, Montelione GT, Gerstein M. 2005. PubNet: a flexible system for visualizing literature derived networks. *Genome Biology* **6(9)**:R80 DOI 10.1186/gb-2005-6-9-r80.

Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. 1997. *Support vector regression machines: advances in neural information processing systems*. Burlington: Morgan Kaufmann Publishers, 155–161.

Fernandez JM, Hoffmann R, Valencia A. 2007. iHOP web services. *Nucleic Acids Research* **35(Web Server issue)**:W21–W26 DOI 10.1093/nar/gkm298.

Fleuren WWM, Alkema W. 2015. Application of text mining in the biomedical domain. *Methods* **74(2)**:97–106 DOI 10.1016/j.ymeth.2015.01.015.

Fleuren WW, Verhoeven S, Frijters R, Heupers B, Polman J. 2011. CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Research* **39(Suppl. 2)**:W450–W454 DOI 10.1093/nar/gkr310.

Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA. 2009. MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Research* **37**:W141–W146 DOI 10.1093/nar/gkp353.

Fontelo P, Liu F, Ackerman M. 2005. ask MEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Medical Informatics and Decision Making* **5**:5 DOI 10.1186/1472-6947-5-5.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. 2013. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41(Database issue)**:D808–D815 DOI 10.1093/nar/gks1094.

Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T. 2013. DeViSE: a deep visual-semantic embedding model. In: *Advances in Neural Information Processing Systems.*

Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal* **9(1)**:207–216 DOI 10.1038/ismej.2014.106.

Giglia E. 2011. Quertle and KNALIJ: searching PubMed has never been so easy and effective. *European Journal of Physical and Rehabilitation Medicine* **47(4)**:687–690.

Glynn RW, Kerin MJ, Sweeney KJ. 2010. Authorship trends in the surgical literature. *British Journal of Surgery* **97(8)**:1304–1308 DOI 10.1002/bjs.7109.

Grandini M, Bagli E, Visani G. 2020. Metrics for multi-class classification: an overview. *ArXiv preprint* DOI 10.48550/arXiv.2008.05756.

Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M. 2014. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy* **58**:212–220 DOI 10.1128/AAC.01310-13.

Hirschman L, Burns GA, Krallinger M, Arighi C, Cohen KB, Valencia A, Wu CH, Chatr-Aryamontri A, Dowell KG, Lourenco A, Nash R, Veuthey A-L, Wiegers T, Winter AG. 2012. Text mining for the biocuration workflow. *Database* **2012**:bas020 DOI 10.1093/database/bas020.

Hokamp K, Wolfe KH. 2004. PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Research* **32(Suppl. 2)**:W16–W19 DOI 10.1093/nar/gkh453.

Huang KC, Chiang IJ, Xiao F, Liao CC, Liu CC, Wong JM. 2013. PICO element detection in medical text without metadata: are first sentences enough? *Journal of Biomedical Informatics* **46(5)**:940–946 DOI 10.1016/j.jbi.2013.07.009.

Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, Rovira P, Abdo Z, Jones KL, Ruiz J, Belk KE, Morley PS, Boucher C. 2017. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Research* **45(D1)**:D574–D580 DOI 10.1093/nar/gkw1009.

Le Q, Mikolov T. 2014. Distributed representations of sentences and documents. In: *International Conference on Machine Learning.* 1188–1196.

Lewis J, Ossowski S, Hicks J, Errami M, Garner HR. 2006. Text similarity: an alternative way to search MEDLINE. *Bioinformatics* **22(18)**:2298–2304 DOI 10.1093/bioinformatics/btl388.

Li C, Jimeno-Yepes A, Arregui M, Kirsch H, Rebholz-Schuhmann D. 2013. PCorral—interactive mining of protein interactions from MEDLINE. *Database* **2013**:bat030 DOI 10.1093/database/bat030.

Liu B, Pop M. 2009. ARDB—antibiotic resistance genes database. *Nucleic Acids Research* **37(Database)**:D443–D447 DOI 10.1093/nar/gkn656.

Manning CD, Raghavan P, Schütze H. 2008. *Introduction to information retrieval.* First Edition. Cambridge: Cambridge University Press.

Mikolov T, Le QV, Sutskever I. 2013. Exploiting similarities among languages for machine translation. *Available at* https://arxiv.org/abs/1309.4168.

Munk P, Knudsen BE, Lukjancenko O, Duarte ASR, Van Gompel L, Luiken REC, Smit LAM, Schmitt H, Garcia AD, Hansen RB, Petersen TN, Bossers A, Ruppé E, Lund O, Hald T, Pamp Sünje J, Vigre Håkan, Heederik D, Wagenaar JA, Mevius D, Aarestrup FM. 2018. Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nature Microbiology* **3**:898–908 DOI 10.1038/s41564-018-0192-9.

Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, Retailleau P, Iorga BI. 2017. Beta-lactamase database (BLDB)-structure and function. *Journal of Enzyme Inhibition and Medicinal Chemistry* **32(1)**:917–919 DOI 10.1080/14756366.2017.1344235.

Nguyen H, Bai L. 2011. Cosine similarity metric learning for face verification. In: Kimmel R, Klette R, Sugimoto A, eds. *Computer Vision C ACCV 2010: Lecture Notes in Computer Science.* Vol. 6493. Berlin: Springer, 709–720.

Pafilis E, O'Donoghue SI, Jensen LJ, Horn H, Kuhn M, Brown NP, Schneider R. 2009. Reflect: augmented browsing for the life scientist. *Nature Biotechnology* **27(6)**:508–510 DOI 10.1038/nbt0609-508.

Paik JH. 2013. A novel TF-IDF weighting scheme for effective ranking. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 343–352.

Perez-Iratxeta C, Bork P, Andrade MA. 2001. XplorMed: a tool for exploring MEDLINE abstracts. *Trends in Biochemical Sciences* **26(9)**:573–575 DOI 10.1016/S0968-0004(01)01926-0.

Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. 2006. AliBaba: PubMed as a graph. *Bioinformatics* **22(19)**:2444–2445 DOI 10.1093/bioinformatics/btl408.

Plikus MV, Zhang Z, Chuong CM. 2006. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics* **7**:2424 DOI 10.1186/1471-2105-7-424.

Raja K, Subramani S, Natarajan J. 2013. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database* **2013**:bas052 DOI 10.1093/database/bas052.

Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. 2008. Text processing through Web services: calling Whatizit. *Bioinformatics* **24(2)**:296–298 DOI 10.1093/bioinformatics/btm557.

Review on Antimicrobial Resistance. 2016. Antimicrobial resistance: TACKLING DRUG-resistant infections globally: final report and recommendations. *Available at* https://apo.org.au/node/63983.

Rodriguez-Esteban R. 2009. Biomedical text mining and its applications. *PLoS Computational Biology* **5(12)**:e1000597 DOI 10.1371/journal.pcbi.1000597.

Ruppe E, Ghozlane A, Tap J, Pons N, Alvarez A-S, Maziers N, Cuesta T, Hernando-Amado S, Clares I, Martínez JL, Coque TM, Baquero F, Lanza VF, Máiz L, Goulenok T, de Lastours V, Amor N, Fantin B, Wieder I, Andremont A, van Schaik W, Rogers M, Zhang X, Willems RJL, de Brevern AG, Batto J-M, Blottière Hé M, Léonard P, Véronique L, Letur A, Levenez F, Weiszer K, Haimet F, Doré J, Kennedy SP, Ehrlich SD. 2019. Prediction of the

intestinal resistome by a three-dimensional structure-based method. *Nature Microbiology* **4**:112–123 DOI 10.1038/s41564-018-0292-6.

Salton G, Wong A, Yang CS. 1975. A vector space model for automatic indexing. *Communications of the ACM* **18(11)**:613–620 DOI 10.1145/361219.361220.

Sayers EW, Agarwala R, Bolton EE, Rodney BJ, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Bradley HJ, Kim S, Kimchi A, Kitts PA, Lathrop S, Lu Z, Madden TL, Marchler-Bauer A, Phan L, Schneider VA, Schoch CL, Pruitt KD, Ostll J. 2019. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **47(D1)**:D23–D28 DOI 10.1093/nar/gky1069.

Smalheiser NR, Zhou W, Torvik VI. 2008. Anne O'Tate: a tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of Biomedical Discovery and Collaboration* **3**:2 DOI 10.1186/1747-5333-3-2.

Socher R, Chen D, Manning CD, Ng AY. 2013a. Reasoning with neural tensor networks for knowledge base completion. In: *Advances in Neural Information Processing Systems*.

Soldatos TG, O'Donoghue SI, Satagopam VP, Jensen LJ, Brown NP, Barbosa-Silva A, Schneider R. 2010. Martini: using literature keywords to compare gene sets. *Nucleic Acids Research* **38(1)**:26–38 DOI 10.1093/nar/gkp876.

Srivastava A, Singhal N, Goel M, Virdi JS, Kumar M. 2014. CBMAR: a comprehensive beta-lactamase molecular annotation resource. *Database* **2014**:bau111 DOI 10.1093/database/bau111.

States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD. 2009. MiSearch adaptive pubMed search tool. *Bioinformatics* **25(7)**:974–976 DOI 10.1093/bioinformatics/btn033.

Suomela BP, Andrade MA. 2005. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics* **6**:75 DOI 10.1186/1471-2105-6-75.

Syarif I, Prugel-Bennett A, Wills G. 2016. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika* **14(4)**:1502 DOI 10.12928/telkomnika.v14i4.3956.

Thai QK, Bos F, Pleiss J. 2009. The lactamase engineering database: a critical survey of TEM sequences in public databases. *BMC Genomics* **10(1)**:390 DOI 10.1186/1471-2164-10-390.

Thai QK, Pleiss J. 2010. SHV lactamase engineering database: a reconciliation tool for SHV beta-lactamases in public databases. *BMC Genomics* **11(1)**:563 DOI 10.1186/1471-2164-11-563.

Tran TT, Munita JM, Arias CA. 2015. Mechanisms of drug resistance: daptomycin resistance. *Annals of the New York Academy of Sciences* **1354(1)**:32–53 DOI 10.1111/nyas.12948.

Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S. 2011. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **27(13)**:i111–i119 DOI 10.1093/bioinformatics/btr214.

Wallace JC, Port JA, Smith MN, Faustman EM. 2017. FARME DB: a functional antibiotic resistance element database. *Database* **2017(3)**:baw165 DOI 10.1093/database/baw165.

Wei CH, Kao H-Y, Lu Z. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* **41(W1)**:W518–W522 DOI 10.1093/nar/gkt441.

Wright GD. 2011. Molecular mechanisms of antibiotic resistance. *Chemical Communications* **47(14)**:4055–4061 DOI 10.1039/c0cc05111j.

Xuan W, Dai M, Mirel B, Wilson J, Athey B, Watson SJ, Meng F. 2007. An active visual search interface for Medline. *Computational Systems Bioinformatics Conference* **6**:359–369 DOI 10.1142/9781860948732_0036.

Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, Tiedje JM, Zhang T, Wren JL. 2018. ARGs-OAP v2.0 with an expanded SARG database and hidden Markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics* **34**:2263–2270 DOI 10.1093/bioinformatics/bty053.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* **67**:2640–2644 DOI 10.1093/jac/dks261.

Zhila A, Yih WT, Meek C, Zweig G, Mikolov T. 2013. Combining heterogeneous models for measuring relational similarity. In: *NAACL HLT*.

Zou W, Socher R, Cer D, Manning C. 2013. Bilingual word embeddings for phrase based machine translation. In: *Conference on Empirical Methods in Natural Language Processing*.