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], **Rommel T J Ramos** [Corresp. 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

Corresponding Author: Rommel T J Ramos
Email address: rommelramos@ufpa.br

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.

1 # Text Mining for Identification of Biological Entities
2 # Related to Antibiotic Resistant Organisms

4 Kelle Fortunato da Costa[1], Fabricio Almeida Araujo[2,3], Jefferson Morais[4], Renato Francês[5] and
5 Rommel Ramos[6]

7 1 Universidade Federal do Pará, Belém, Pará, Brasil
8 2 Computational Biology, Universidade Federal do Pará, Belém, Pará, Brasil
9 3 Universidade Federal do Rural da Amazônia, Belém, Pará, Brasil
10 4 Universidade Federal do Pará, Belém, Pará, Brasil
11 5 Universidade Federal do Pará, Belém, Pará, Brasil
12 6 Biological Science Institute, Universidade Federal do Pará, Belém, Pará, Brasil

14 Corresponding Author:
15 Rommel Ramos
16 R. Augusto Corrêa, 01 - Guamá, Belém - PA, 66075-110, Brasil
17 E-mail address: rommelramos@ufpa.br

Abstract

21 Antimicrobial resistance is a significant public health problem worldwide. In recent years, the
22 scientific community has been intensifying efforts to combat this problem; many experiments
23 have been developed, and many articles are published in this area. However, the growing volume
24 of biological literature increases the difficulty of the biocuration process due to the cost and time
25 required. Modern text mining tools with the adoption of artificial intelligence technology are
26 helpful to assist in the evolution of research. In this article, we propose a text mining model
27 capable of identifying and ranking prioritizing scientific articles in the context of antimicrobial
28 resistance. We retrieved scientific articles from the PubMed database, adopted machine learning
29 techniques to generate the vector representation of the retrieved scientific articles, and identified
30 their similarity with the context. As a result of this process, we obtained a dataset labeled
31 "Relevant" and "Irrelevant" and used this dataset to implement one supervised learning
32 algorithm to classify new records. The model's overall performance reached 90% accuracy and
33 the f-measure (harmonic mean between the metrics) reached 82% accuracy for positive class and
34 93% for negative class, showing quality in the identification of scientific articles relevant to the
35 context. The dataset, scripts and models are available at
36 https://github.com/engbiopct/TextMiningAMR.

37

Introduction

39   Antibiotics are the most successful drugs of the last 100 years, responsible for saving countless
40   lives and enabling modern medical procedures that would otherwise be unthinkable. However,
41   all antibiotics derived from secondary or fully synthetic microbial metabolism products are
42   subject to resistance [1].
43   Antimicrobial resistance (AMR) has been increasingly recognized as an important public health
44   problem worldwide, considering that infections caused by multidrug-resistant organisms (MDR)
45   result in a significant increase in mortality and cause a tremendous economic burden [2]. In
46   addition to the costs of rising hospital admission rates, it is estimated that by 2050 there will be
47   an economic loss of $100 trillion in global antibiotic production [3].
48   In recent years, the scientific community has intensified efforts to combat this problem by
49   making available a wide range of public databases specific to AMR, such as: National Database
50   of Antibiotic Resistant Organisms (NDARO) [4], Comprehensive Antibiotic Resistance
51   Database (CARD) [5], Resfinder [6], ResfinderFG [7], Resfams [8], Antibiotic Resistance Genes
52   Database (ARDB) [9], MEGARes [10], Antibiotic Resistance Gene Annotation (ARG-ANNOT)
53   [11], Mustard [12], Functional Antibiotic Resistance Metagenomic Element (FARME database)
54   [13], SARG (v2) [14], Lahey list of β-lactamases [15], β-Lactamase Database (BLDB) [16],
55   Lactamase Engineering Database (LacED) [17, 18], Comprehensive β-Lactamase Molecular
56   Annotation Resource (CBMAR) [19], among others, which are frequently used as reference
57   databases, with gene sequences related to antimicrobial resistance and metadata that enrich the
58   characterization of sequences. Due to the increase in the volume of literature related to biological
59   and health sciences, the curation process has become challenging for researchers and biocurators
60   who use these databases as a source of research, mainly due to the time required to locate
61   relevant information about biological entities related to antibiotic-resistant organisms. Even
62   queries in specialized databases in biomedical literature such as PubMed (scientific and medical
63   abstracts/citations), PubMed Central (full-text journal articles), NLM Catalog (index of NLM
64   collections), Books (books and reports), and MeSH (ontology used for PubMed indexing) [20],
65   tends to make document selection difficult due to a large amount of retrieved items.
66   In this context, the adoption of text mining (TM) techniques are viable alternatives [21], as they
67   can help in different stages of the standard biocuration workflow. According to [22], the steps
68   are:
69   • Selection: search for articles relevant to the curation.
70   • Identification and standardization of bioentities: detection of mentions of bioentities relevant to
71   the curatorship; for example, genes, proteins, or small molecules, linked to unique identifiers
72   from databases such as UniProt, EntrezGene, or ChEBI.
73   • Detection of annotation events: identification and encoding of events, such as descriptions of
74   protein-protein interactions, characterizations of gene products in terms of cellular location,
75   molecular function, involvement in the biological process and phenotypic effect.
76   • Evidence qualifier association: association of experimental evidence that supports the
77   annotation event performed due to biocuration efforts.
78   • Completion and verification of the database record.

79 TM technologies combine knowledge resources such as controlled vocabularies, taxonomies, and
80 ontologies with linguistic analysis and machine learning to deal with language variations and
81 extract not only terms from the text but also relationships between terms [23].
82 In the last decade, several applications in the biomedical area were developed [24 - 50] using one
83 or more of the following TM steps: (i) retrieving textual resources relevant to a particular subject
84 of interest, a process known as information retrieval (IR), (ii) detect the occurrence of specific
85 keywords of interest and the relationships between these keywords and (iii) infer new
86 relationships based on known facts, and this step is called knowledge discovery (KD) [51].
87 One of the most used machine learning techniques in knowledge discovery, especially in
88 document screening, is text classification. However, supervised classification requires the prior
89 labeling of a training set, a non-trivial task for human curators, as it requires a lot of time and
90 effort.
91 In this sense, Suomela and Andrade [52] propose a methodology for automatic (binary)
92 classification of large volumes of data, adopting the word counting technique, known as the bag
93 of words (BOW) [53], a textual representation that composes the vector space model [54], where
94 documents are converted into vectors of words. A weighting scheme is applied to each word, which
95 can be a simple word count or a metric such as Term Frequency - Inverse Document Frequency
96 (tf -idf) [53][55][56] and based on the arithmetic mean of the weights of these words, text
97 summaries are classified as relevant or irrelevant.
98 This methodology was implemented from specific abstracts of an area of interest, extracted from
99 Pubmed, and inspired the development of the MedlineRanker application [57] and served as a
100 baseline for this study, which instead of the bag of words, adopts a representation approach based
101 on neural networks [58], called Paragraph Vector – Distributed Memory (PV-DM) [59], capable
102 of revealing semantic characteristics between documents, a property that makes this approach
103 useful for many natural language processing (NLP) tasks and justifies its wide use in works
104 involving natural language understanding [60; 61], machine translation [62; 63], image
105 comprehension [64] and relational extraction [65].
106 The study of antimicrobial resistance genes is essential to public health. However, there is
107 challenging to handle and extract the amount of available data of scientific and medical
108 manuscripts from public databases without computational methods. Thus, this work proposes an
109 unsupervised learning-based TM approach for ranking the relevance of articles on AMR context
110 to generate a set of training, accurate enough to generalize new data, maximizing the efficiency of
111 the supervised classifiers.

112

Materials & Methods

1. Labeling pipeline

115 Figure 1 shows the TM steps implemented in this work in order to label the data.
116 **Figure 1.** Proposed TM model. Steps (A) and (B) include retrieving the information. Steps (C)
117 and (D) include the recognition of entities and the discovery of knowledge, resulting in a metric
118 (cosine similarity) responsible for determining the binary classification performed in step (E).
119 1.1. Information retrieval

120 An Application Programming Interface (API) was implemented to retrieve a collection of
121 relevant articles in the Drug Resistance, and Microbial domain through the Pubmed Central
122 (PMC) database, which contains free full-text files of the Library's of Medicine and the US
123 National Institutes of Health (NIH/NLM) biomedical literature are available. In the API, the E-
124 Search and E-Fetch tools from the E-utilities package were used, which provide a structured
125 interface for accessing the Entrez system, the NCBI database system, which currently includes
126 38 databases, covering a variety of biomedical data, including nucleotide and protein sequences,
127 gene records, three-dimensional molecular structures, and biomedical literature [66]. Table 1
128 presents the set of parameters incorporated into the search (E-Search).
129 **Table 1**. Parameters *E-Search* PubMed Central
130 The terms of the MeSH hierarchy were adopted for antimicrobial resistance, considering that the
131 terms MeSH is a controlled vocabulary of biomedical terms whose elements are assigned to a
132 document by indexers (specialists in biomedical subjects) based on its context. They contain
133 high-density document-wide information that cannot be deduced from the title or abstract that
134 PubMed returns using keywords [67].
135 Then, a list of PMCIDs (unique identifiers provided by PubMed Central to each document) is
136 generated to be used to access the full texts of articles through the E-Fetch utility.
137
138 1.2. Named entity recognition and Knowledge Discovery
139 The entity recognition and knowledge discovery process consist of two steps: in the first step
140 (Figure 1-C), the Doc2Vec unsupervised learning algorithm from the Gensim library was used,
141 which implements the Paragraph Vector – Distributed Memory model, to obtain the embedding
142 of the retrieved documents (dense representation of a sequence of words). Table 2 displays the
143 parameters used in the algorithm.
144 **Table 2**. Parameters Doc2Vec algorithm
145 In the second step (Figure 1-D), the pre-trained model was used, capable of predicting whether a
146 set of documents $\{Doc_1, Doc_2, Doc_3...Doc_n\}$ belong to the context of a central document Doc_0, to
147 infer the similarity of the documents to the AMR context, represented by 4,290 terms extracted
148 from CARD and the Gene Ontology Database [68], calculating the cosine distance [69] between
149 them. The resulting value varies in the range between -1 and 1 where the higher the number, the
150 greater the similarity with the context.
151 Finally, each of the scientific articles was automatically labeled (Figure 1-E), considering the
152 arithmetic average of the cosine similarity, values above the average were defined as relevant
153 and below the mean as irrelevant.
154 2. Evaluation of the proposed method
155 To evaluate the classification performance of the proposed TM approach (Figure 2 - A), the same
156 dataset was labeled using the Bag of words text representation model (Figure 2 - B), adopting the
157 specific vocabulary for AMR as a vector of features, and the documents were classified as
158 relevant or irrelevant through the arithmetic mean of the weights of the words, obtained with the
159 TF-IDF weighting method, similar to the methodology proposed in [52]. Finally, the two
160 (automatically) labeled databases were compared with a test database labeled by experts (Figure

161   2 - C), who independently labeled the articles as relevant or irrelevant. Only the samples where
162   the three experts converged on the labels were included in the test dataset.
163   **Figure 2.** Evaluation of the proposed method.
164
165   3.   Predictions with automatically labeled data
166   To evaluate the efficiency of the proposed approach, which uses neural embeddings for labeling,
167   the generated datasets (Figure 3 - Dataset_1 and Figure 3 - Dataset_2) were used as input to the
168   supervised classifier SVM [70; 71]. The classification performance was evaluated through the
169   analysis of the Precision, Recall, Accuracy, and F-Measure metrics [72], calculated from the test
170   base labeled by experts and not used to train the SVM models (Figure 3).
171   **Figure 3.** Performance of predictions with automatically labeled data.
172   For the SVM classifier, the feature/attribute vector (AMR vocabulary with 4,290 words) was
173   weighted using the TF-IDF technique, and cross-validation, with 5-folds (default value of the
174   adopted algorithm), was applied to explore the combination of parameters for determining the
175   best model, as the effectiveness of the SVM depends on the kernel selection, which is the
176   function that will be used by the algorithm, in the margin parameter (C), which determines a
177   balance between maximizing the margin and minimizing classification errors, and the Gamma
178   parameter, when the chosen kernel is Gaussian (or RBF) [73], adopted in this experiment.
179
180   4.   Data Avaliability
181   The dataset, script, and models generated by this work are available at
182   https://github.com/engbiopct/TextMiningAMR, under CC BY 4.0 Copyright license, with
183   information regards the workflow adopted, a short step-by-step guide to the readers reproduce
184   this experiment and the complementary materials.
185

Results and Discussion

187   1.   Labeling
188   A collection of 88,300 scientific articles on antimicrobial resistance was retrieved from Pubmed
189   Central, using the terms of reference of the MeSH (Medical Subject Headings, developed at the
190   National Library of Medicine) hierarchical vocabulary referring to the AMR domain
191   (https://meshb.nlm.nih.gov/record/ui?ui=D004352).
192   The retrieved dataset was submitted to the PV-DM text representation model, with the
193   embedding of the documents obtained. The similarity of the documents to the AMR context
194   (Table S1) was inferred, and the label "relevant" was automatically assigned (class 0) to all
195   articles whose cosine distance value was equal to or greater than the arithmetic average of the
196   cosine distances of the entire corpus, and the label "irrelevant" (class 1) to all articles with a
197   cosine distance value lower than that referred to average, resulting in 43,136 records labeled
198   relevant and 45,164 labeled irrelevant (Table S2).
199   The same initial dataset was submitted to the Bag of words text representation model in order to
200   obtain the weights of the words in the documents according to the AMR dictionary and thus

201   automatically assign the label "relevant" (class 0) to all articles with weight equal to or greater
202   than the arithmetic average of the weights, and the label "irrelevant" (class 1) to all articles with
203   a value lower than the mean, resulting in 45,946 records labeled as relevant and 42,354 labeled
204   as irrelevant (Table S3).
205   With the two labeled datasets, the results were compared with a test dataset labeled by experts,
206   consolidated with a total of 62 scientific articles, 15 labeled as relevant and 47 labeled as
207   irrelevant (Table S4).
208   In the comparison, the proposed method labeled 44 articles according to the experts, which
209   represents 71% of hits in general, with 80% of hits for the relevant label and 68% of hits for the
210   irrelevant label. As for the baseline method, there were only 26 labels according to the experts,
211   which represents 42% of overall performance, with 66% of correctness for the relevant label and
212   34% of correctness for the irrelevant label.
213   The proposed approach presents a superior performance about the baseline, which despite its
214   simplicity, efficiency, and often surprising precision, does not take into account the order and
215   semantics of the words, that is, the distances between them. This means that the words "mighty",
216   "strong" and "Paris" are equally distant. Although semantically, "powerful" is closer to "strong"
217   than "Paris" [59], characteristics present in the PV-DM model and fundamental human skills in
218   manual data labeling tasks.
219
220   2.   Classification
221   The two labeled datasets were submitted to the supervised SVM classifier, excluding the test
222   dataset labeled by experts from training.
223   Figure 4 presents the confusion matrix of the SVM_1 classifier, trained with data from dataset 1
224   (Figure 3 - Dataset_1). There is a high degree of precision both in terms of true positives
225   (relevant articles classified as relevant) and true negatives (non-relevant articles classified as
226   non-relevant).
227   Only 1 article was incorrectly classified as relevant (false positive), and 5 articles were
228   incorrectly classified as irrelevant (false negative).
229   **Figure 4.** SVM classifier confusion matrix for dataset_1 (PV-DM)
230   Figure 5 presents the confusion matrix of SVM_2, trained with data from dataset 2 (Figure 3 -
231   Dataset_2), where a lower degree of precision is observed for both true positives and true
232   negatives in relation to SVM_1. With this classifier, however, there was an increase in the
233   number of incorrect classifications, with 33 articles incorrectly classified as relevant (false
234   positives) and 6 articles incorrectly classified as irrelevant (false negatives).
235   **Figure 5.** SVM classifier confusion matrix for dataset_1 (Bag of Words)
236   Table 3 presents the results of the evaluation metrics: precision, recall, accuracy and the f-
237   measure, obtained based on the results of the confusion matrix of the two classifiers. The results
238   of the SVM_1 classifier were superior to the SVM_2 classifier in all evaluated metrics.
239   **Table 3.** Classifier Performance Assessment
240   Accuracy, a metric that represents the overall performance of a model, reached 90% of accuracy
241   and the f-measure, which is a harmonic average between the precision and recall metrics and that

242   can be used as a single measure to represent the quality in the text mining [74] reached 82%
243   success rate for the positive class and 93% for the negative class.
244   The results show that the best performance was obtained with the database labeled by the PV-
245   DM model.
246   **Table 4.** Results of Labeling and Classification steps vs Experts
247   Table 4 presents the percentage of correct predictions, both in the labeling and in the
248   classification stage, in comparison with the data labeled by experts and validates the hypothesis
249   that the use of Paragraph Vector, Distributed Representations of Sentences and Documents
250   associated with similarity with a specific context is able not only to perform the binary
251   classification of large volumes of data but also to optimize the percentage of hits of supervised
252   classifiers. The SVM_2 classifier showed a reduction in the number of hits compared to the
253   Labeling step, although we adopted the same attribute vector and the same representation in both
254   experiments (bag of words, weighted with TF-IDF).
255

Conclusions

257   The proposed TM approach proved capable of identifying and prioritizing documents in the
258   AMR context, as well as predicting the relevance of new documents in the same context. For
259   this, we used the TM steps summarized in [51], plus some specifics of the proposal such as (1)
260   Use of the MeSH hierarchy; (2) Use of full text; (3) Use of domain-specific dictionaries (CARD
261   and Gene Ontology), fundamental in this process, as it facilitated the detection of similarity of
262   articles to the AMR context and (4) Adoption of unsupervised learning for better representation
263   of texts. Additionally, we submitted the labeled bases to the SVM classifier to evaluate their
264   performance in comparison to the test base labeled by human experts.
265   The proposed approach efficiently identifies scientific articles relevant to the AMR context.
266   Therefore, it is a valuable tool to automate information capture processes from robust
267   bibliographic reference databases such as Pubmed Central, as well as to accelerate the screening
268   of documents with biocuration potential, facilitating the other stages of the biocuration process.
269   This work presents a new set of pre-trained document embeddings in the AMR domain and a
270   base labeled for relevance according to similarity to CARD and Gene Ontology, which, in future
271   work, can be used as input to other algorithms of supervised learning and for biocuration tools
272   aimed at the identification and normalization of bioentities, detection of annotation events and
273   filling out specific databases for AMR. This proposal is limited to the previous existence of an
274   accurate database representing the main terms related to the target.
275
276

Acknowledgements

278

References

280   [1] G. D. Wright, Molecular mechanisms of antibiotic resistance. Chem. Commun. (Camb.) 47,
281   4055–061 (2011). 10.1039/c0cc05111j pmid:21286630 doi:10.1039/c0cc05111j.

282 [2] Tran TT, Munita JM, Arias CA. Mechanisms of drug resistance: daptomycin resistance. Ann
283 N Y Acad Sci. 2015; 1354:32–53.

284 [3] Review on Antimicrobial Resistance. 2016. Antimicrobial resistance: TACKLING DRUG-
285 Resistant infections globally: final report and recommendations. accessed march 11, 2020.

286 [4] NLM Annual Report 2016 - National Library of Medicine – NIH.

287 [5] Alcock et al . CARD 2020: antibiotic resistome surveillance with the comprehensive
288 antibiotic resistance database, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020,
289 Pages D517–D525, https://doi.org/10.1093/nar/gkz935.

290 [6] Zankari E et al. Identification of acquired antimicrobial resistance genes. J. Antimicrob.
291 Chemother 67, 2640–2644 (2012).

292 [7] Munk P et al. Abundance and diversity of the faecal resistome in slaughter pigs and broilers
293 in nine European countries. Nat. Microbiol 3, 898–908 (2018).

294 [8] Gibson MK, Forsberg KJ & Dantas G Improved annotation of antibiotic resistance
295 determinants reveals microbial resistomes cluster by ecology. ISME J 9, 207–216 (2015).

296 [9] Liu B & Pop M ARDB—antibiotic resistance genes database. Nucleic Acids Res 37, D443–
297 D447 (2009).

298 [10] Lakin SM et al. MEGARes: an antimicrobial resistance database for high throughput
299 sequencing. Nucleic Acids Res 45, D574–D580 (2017).

300 [11] Gupta SK et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance
301 genes in bacterial genomes. Antimicrob. Agents Chemother 58, 212–220 (2014).

302 [12] Ruppe E et al. Prediction of the intestinal resistome by a three-dimensional structure-based
303 method. Nat. Microbiol 4, 112–123 (2019)

304 [13] Wallace JC, Port JA, Smith MN & Faustman EM FARME DB: a functional antibiotic
305 resistance element database. Database (Oxford) 2017, baw165 (2017).

306 [14] Yin X et al. ARGs-OAP v2.0 with an expanded SARG database and hidden Markov models
307 for enhancement characterization and quantification of antibiotic resistance genes in
308 environmental metagenomes. Bioinformatics 34, 2263–2270 (2018).

309 [15] Bush K & Jacoby GA Updated functional classification of β-lactamases. Antimicrob.
310 Agents Chemother 54, 969–976 (2010).

311 [16] Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, Retailleau P, Iorga BI.
312 Beta-lactamase database (BLDB) - structure and function. J Enzyme Inhib Med Chem. 2017
313 Dec;32(1):917-919. doi: 10.1080/14756366.2017.1344235. PMID: 28719998; PMCID:
314 PMC6445328.

315 [17] Thai QK, Bos F & Pleiss J The lactamase engineering database: a critical survey of TEM
316 sequences in public databases. BMC Genomics 10, 390 (2009).

317 [18] Thai QK & Pleiss J SHV lactamase engineering database: a reconciliation tool for SHV
318 beta-lactamases in public databases. BMC Genomics 11, 563 (2010).

319 [19] Srivastava A, Singhal N, Goel M, Virdi JS & Kumar M CBMAR: a comprehensive beta-
320 lactamase molecular annotation resource. Database (Oxford) 2014, bau111 (2014).

321 [20] Sayers, Eric W et al. "Database resources of the National Center for Biotechnology
322 Information." Nucleic acids research vol. 47,D1 (2019): D23-D28. doi:10.1093/nar/gky1069
323 [21] Hirschman L, Burns GA, Krallinger M, Arighi C, Cohen KB, Valencia A, Wu CH,
324 Chatr-Aryamontri A, Dowell KG, Huala E, et al. Text mining for the biocuration workflow,
325 Database, 2012, vol. 2012 April 18 (doi:10.1093/database/bas020; epub ahead of print).
326 [22] Wei CH, et al. PubTator: a web-based text mining tool for assisting biocuration, Nucleic
327 Acids Res., 2013, vol. 41 Web server(pg. W518-W522).
328 [23] Chaix Estelle, Deléger Louise, Bossy Robert, Nédellec Claire. Text mining tools for
329 extracting information about microbial biodiversity in food. Food Microbiology. 2018 doi:
330 10.1016/j.fm.2018.04.011.
331 [24] W.W. Fleuren et al., Nucleic Acids Res. 39 (Web Server issue) (2011) W450–W454.
332 [25] P. Fontelo, F. Liu, M. Ackerman, BMC Med. Inf. Decis. Mak. 5 (2005) 5.
333 [26] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Trends Biochem. Sci. 26 (9) (2001) 573–575.
334 [27] J. Lewis et al., Bioinformatics 22 (18) (2006) 2298–2304.
335 [28] J.F. Fontaine et al., Nucleic Acids Res. 37 (Web Server issue) (2009) W141– W146.
336 [29] D.J. States et al., Bioinformatics 25 (7) (2009) 974–976.
337 [30] K.C. Huang et al., J. Biomed. Inf. 46 (5) (2013) 940–946.
338 [31] K. Hokamp, K.H. Wolfe, Nucleic Acids Res 32 (Web Server issue) (2004) W16–W19.
339 [32] M.V. Plikus, Z. Zhang, C.M. Chuong, BMC Bioinf. 7 (2006) 424.
340 [33] K.G. Becker et al., BMC Bioinf. 4 (2003) 61.
341 [34] S.M. Douglas, G.T. Montelione, M. Gerstein, Genome Biol. 6 (9) (2005) R80.
342 [35] B. Brancotte et al., Bioinformatics 27 (8) (2011) 1187–1189.
343 [36] S. De et al., Physiol. Genomics 42A (2) (2010) 162–167.
344 [37] N.R. Smalheiser, W. Zhou, V.I. Torvik, J. Biomed. Discov. Collab. 3 (2008) 2.
345 [38] H. Chen, B.M. Sharp, BMC Bioinf. 5 (2004) 147.
346 [39] C. Li et al., Database (oxford) 2013 (2013) bat030.
347 [40] R.W. Glynn, M.J. Kerin, K.J. Sweeney, Br. J. Surg. 97 (8) (2010) 1304–1308.
348 [41] W. Xuan et al., Comput. Syst. Bioinf. Conf. 6 (2007) 359–369.
349 [42] E. Giglia, Eur. J. Phys. Rehabil. Med. 47 (4) (2011) 687–690.
350 [43] Y. Tsuruoka et al., Bioinformatics 27 (13) (2011) i111–i119.
351 [44] J.M. Fernandez, R. Hoffmann, A. Valencia, Nucleic Acids Res. 35 (Web Server issue)
352 (2007) W21–W26.
353 [45] K. Raja, S. Subramani, J. Natarajan, Database (Oxford) 2013 (2013) bas052.
354 [46] E. Pafilis et al., Nat. Biotechnol. 27 (6) (2009) 508–510.
355 [47] D. Rebholz-Schuhmann et al., Bioinformatics 24 (2) (2008) 296–298.
356 [48] C. Plake et al., Bioinformatics 22 (19) (2006) 2444–2445.
357 [49] T.G. Soldatos et al., Nucleic Acids Res. 38 (1) (2010) 26–38.
358 [50] A. Franceschini et al., Nucleic Acids Res. 41 (Database issue) (2013) D808–D815.
359 [51] Fleuren, W. W. M. & Alkema, W. Application of text mining in the biomedical domain.
360 Methods 74, 97–106 (2015).

361 [52] Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large
362 training set using text indexing, BMC Bioinformatics, 2005, vol. 6 pg. 75
363 [53] C. D. Manning, P. Raghavan, H. Sch¨utze, Introduction to Information Retrieval, 1st
364 Edition, Cambridge University Press, Cambridge, 2008.
365 [54] SALTON, Gerard; WONG, Anita; YANG, Chung-Shu. A vector space model for automatic
366 indexing. Communications of the ACM, v. 18, n. 11, p. 613-620, 1975.
367 [55] J. H. Paik, A novel tf-idf weighting scheme for effective ranking, in: Proceedings of the
368 International ACM SIGIR Conference on Research and Development in Information Retrieval,
369 2013, pp. 343–352.
370 [56] C.-H. Chen, Improved tfidf in big news retrieval: An empirical study, Pattern Recognition
371 Letters 93 (2017) 113–122.
372 [57] Jean-Fred Fontaine, Adriano Barbosa-Silva, Martin Schaefer, Matthew R. Huska, Enrique
373 M. Muro, Miguel A. Andrade-Navarro, MedlineRanker: ranking flexível de literatura biomédica,
374 Nucleic Acids Research , Volume 37, Issue suppl_2, 1 de julho 2009, Pages W141 – W146,
375 https://doi.org/10.1093/nar/gkp353
376 [58] Bengio, Yoshua, Schwenk, Holger, Sen´ecal, Jean-S´ebastien, Morin, Fr´ederic, and
377 Gauvain, Jean-Luc. A Neural probabilistic language models. In Innovations in Machine
378 Learning, pp. 137–186. Springer, 2006.
379 [59] LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In:
380 International Conference on Machine Learning. [S.l.: s.n.], 2014. p. 1188–1196.
381 [60]Collobert, Ronan andWeston, Jason. A unified architecture for natural language processing:
382 Deep neural networks with multitask learning. In Proceedings of the 25th International
383 Conference on Machine Learning, pp. 160– 167. ACM, 2008.
384 [61] Zhila, A., Yih, W.T., Meek, C., Zweig, G., and Mikolov, T. Combining heterogeneous
385 models for measuring relational similarity. In NAACL HLT, 2013.
386 [62] Mikolov, Tomas, Le, Quoc V., and Sutskever, Ilya. Exploiting similarities among languages
387 for machine translation. CoRR, abs/1309.4168, 2013b.
388 [63] Zou, Will, Socher, Richard, Cer, Daniel, and Manning, Christopher. Bilingual word
389 embeddings for phrasebased machine translation. In Conference on Empirical Methods in
390 Natural Language Processing, 2013.
391 [64] Frome, Andrea, Corrado, Greg S., Shlens, Jonathon, Bengio, Samy, Dean, Jeffrey, Ranzato,
392 Marc'Aurelio, and Mikolov, Tomas. DeViSE: A deep visual-semantic embedding model. In
393 Advances in Neural Information Processing Systems, 2013.
394 [65]Socher, Richard, Chen, Danqi, Manning, Christopher D., and Ng, Andrew Y. Reasoning
395 with neural tensor networks for knowledge base completion. In Advances in Neural Information
396 Processing Systems, 2013a.
397 [66] NCIBI Homepage, https://www.ncbi.nlm.nih.gov/books/NBK3827/, acessado em
398 25/04/2021.
399 [67] NLM: Medical Subject Headings (MeSH).
400 [68] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000.

401 [69] H. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in
402 Computer Vision C ACCV 2010, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette,
403 and A. Sugimoto, Eds. Springer Berlin Heidelberg, 2011, vol. 6493, pp. 709–720.
404 [70] BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A Training Algorithm for Optimal Margin
405 Classifiers. In: ANNUAL WORKSHOP ON COMPUTACIONAL LEARNING, 5, 1992,
406 Pittsburgh. ACM Press. Pittsburgh: Haussler D, jul 1992. p.144-152 .
407 [71] DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A.; VAPNIK, V. Support
408 vector regression machines. Advances in neural information processing systems, Morgan
409 Kaufmann Publishers, p. 155–161, 1997.
410 [72] PANG, B.; LEE, L. Opinion mining and sentiment analysis. Foundations and Trends in
411 Information Retrieval, v. 2, n. 1-2, p. 1–135, 2008.
412 [73] SYARIF, Iwan; PRUGEL-BENNETT, Adam; WILLS, Gary. SVM parameter optimization
413 using grid search and genetic algorithm to improve classification performance. Telkomnika, v.
414 14, n. 4, p. 1502, 2016.
415 [74] Rodriguez-Esteban, R. (2009) Biomedical text mining and its applications. PLoS Comput.
416 Biol. 5, e1000597.

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).

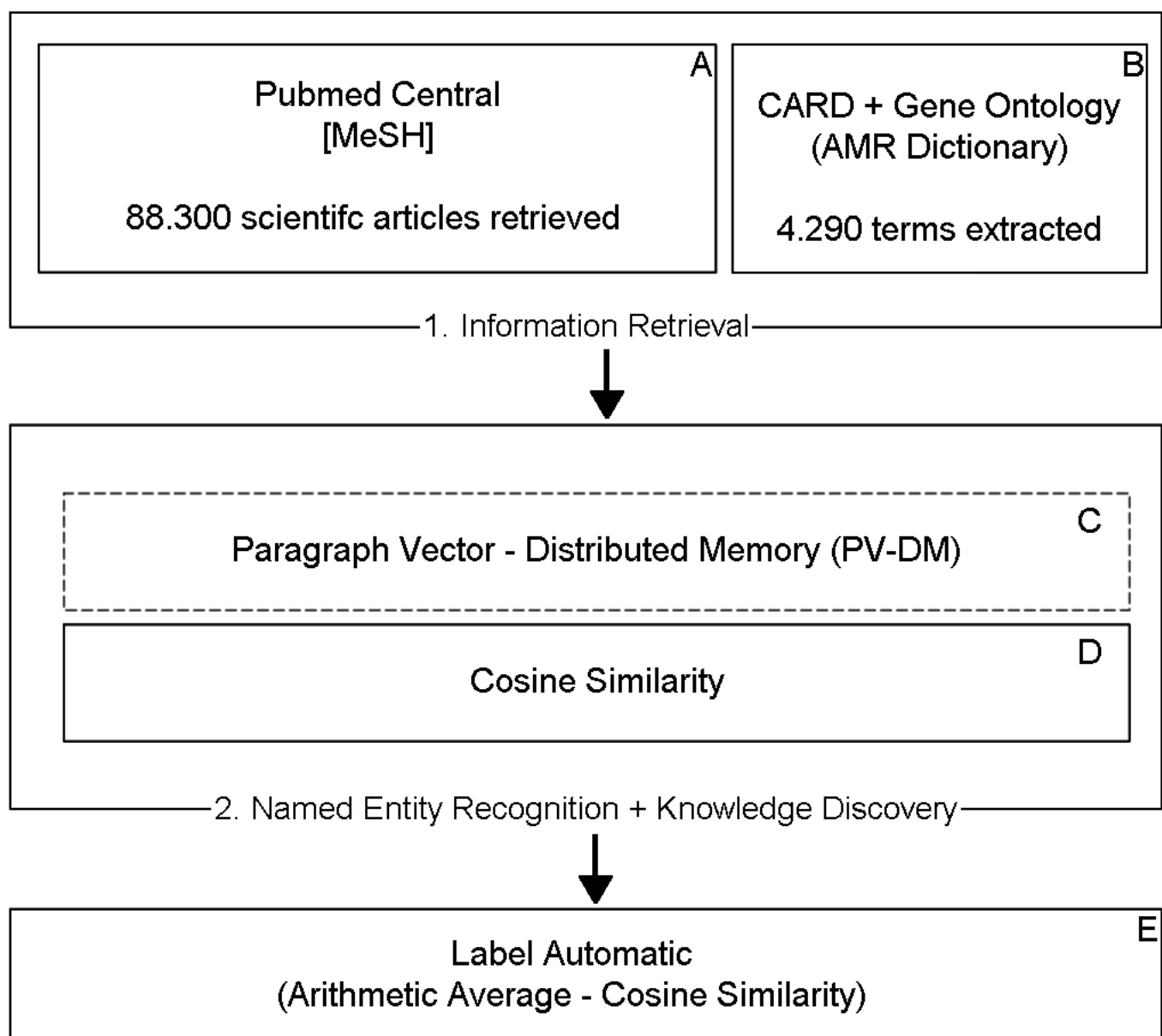



Figure 2

Evaluation of the proposed method

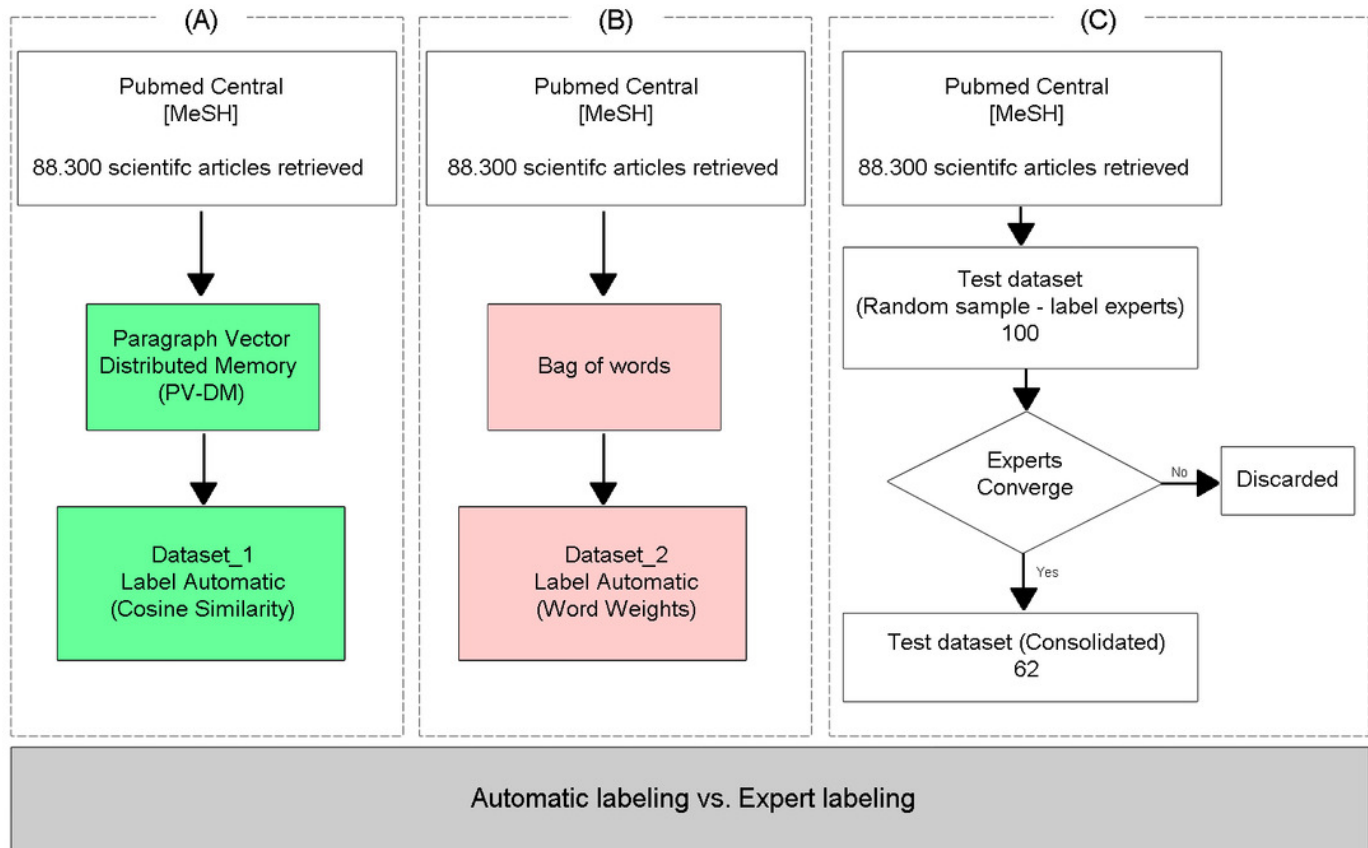



Figure 3

Performance of predictions with automatically labeled data

Figure 4

SVM classifier confusion matrix for dataset_1 (PV-DM)

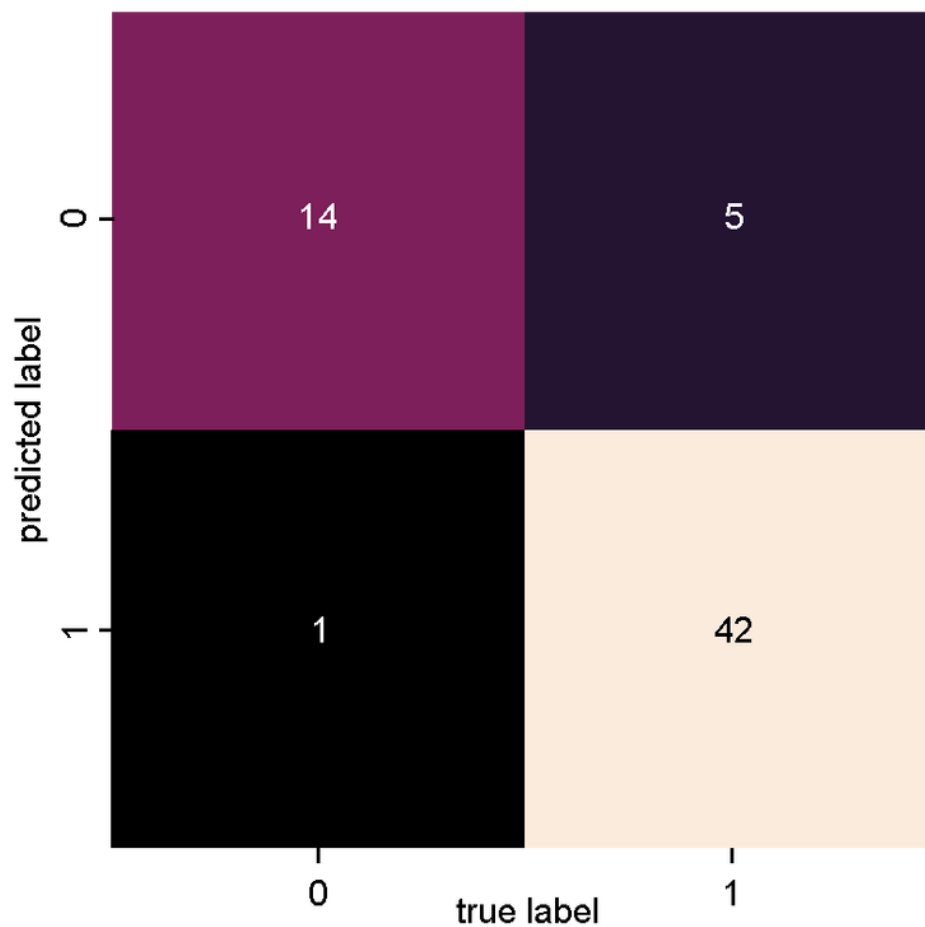
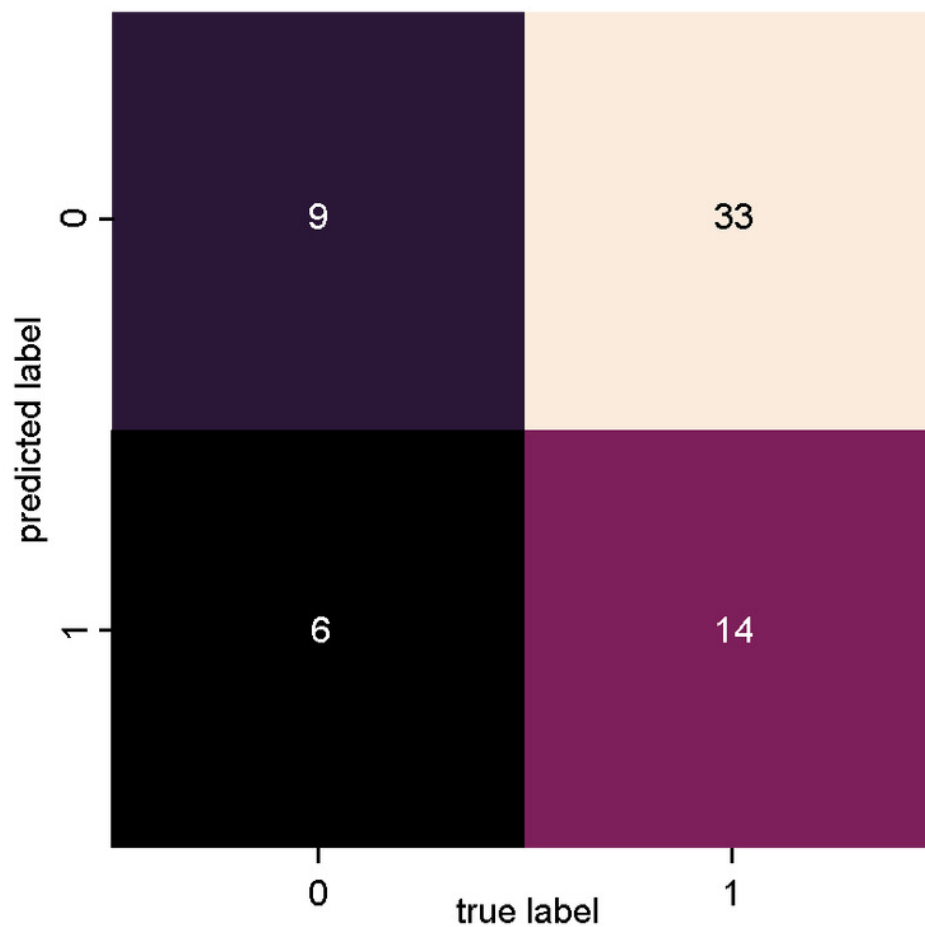



Figure 5

SVM classifier confusion matrix for dataset_1 (Bag of Words)

Parameters *E-Search* PubMed Central

1 **Table 1.** Parameters *E-Search* PubMed Central

Parameters	Value
URL	https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=""&term ="" AND ""[filter]
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

Parameters Doc2Vec algorithm

1   **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.

Classifier Performance Assessment

1   **Table 3.** Performance evaluation of the classifiers SVM

	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	

Results of Labeling and Classification steps vs Experts

1   **Table 4.** Labeling and Classification vs Experts

	Relevant	Irrelevant
Labeling		
Dataset_1	80%	68%
Dataset_2	66%	34%
Classification		
SVM_1	**93%**	**89%**
SVM_2	60%	29%