




Review

Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review

Cyrille YetuYetu Kesiku ^{*}, Andrea Chaves-Villota  and Begonya Garcia-Zapirain 

eVida Research Group, University of Deusto, Avda/Universidades 24, 48007 Bilbao, Spain

^{*} Correspondence: cyrille.kesiku@opendeusto.es

Abstract: The classification of biomedical literature is engaged in a number of critical issues that physicians are expected to answer. In many cases, these issues are extremely difficult. This can be conducted for jobs such as diagnosis and treatment, as well as efficient representations of ideas such as medications, procedure codes, and patient visits, as well as in the quick search of a document or disease classification. Pathologies are being sought from clinical notes, among other sources. The goal of this systematic review is to analyze the literature on various problems of classification of medical texts of patients based on criteria such as: the quality of the evaluation metrics used, the different methods of machine learning applied, the different data sets, to highlight the best methods in this type of problem, and to identify the different challenges associated. The study covers the period from 1 January 2016 to 10 July 2022. We used multiple databases and archives of research articles, including Web Of Science, Scopus, MDPI, arXiv, IEEE, and ACM, to find 894 articles dealing with the subject of text classification, which we were able to filter using inclusion and exclusion criteria. Following a thorough review, we selected 33 articles dealing with biological text categorization issues. Following our investigation, we discovered two major issues linked to the methodology and data used for biomedical text classification. First, there is the data-centric challenge, followed by the data quality challenge.

Keywords: text classification; biomedical document; natural language processing; biomedical text classification challenges

**Citation:** Kesiku, C.Y.;Chaves-Villotas, A.; Garcia-Zapirain, B. Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review. *Information* **2022**, *13*, 499. <https://doi.org/10.3390/info13100499>

Academic Editor: Ralf Krestel

Received: 15 September 2022

Accepted: 11 October 2022

Published: 17 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The focus on text data is increasing day by day in different fields. Generally in the healthcare field, patient information consists mostly of medical texts or notes taken by doctors and nurses. The classification of medical text in the process of extracting knowledge from medical data has gained momentum in recent times thanks to Natural Language Processing techniques. In this technique, the main approach is the recognition of a necessary pattern that explains a fact from the links between words and sentences in a text. These links give a semantic meaning and allow a good understanding of the information in the text. In health, this helps in the rapid search for the causes of a disease and correlates all the causes extracted from the text to predict the disease. Many other problems are treated by following this approach.

Since 2013 until today, NLP research has demonstrated its inescapable capabilities with very relevant models emerging every year probably. Techniques based on neural network architectures, very intuitive in classification and other important natural language processing tasks [1,2]. Many other problems in health care use text classification such as in the International Classification of Diseases (ICD), which is a medical classification list published by the World Health Organization, which defines the universe of diseases, disorders, injuries and other related health conditions as well as the standard of diagnosis classification [3,4].

In this systematic review, we examine the different articles on patient medical text classification from 1 January 2016 to 10 July 2022, in order to identify the relevant challenges in biomedical text classification. The knowledge gained in this study will clearly map out the methodologies and techniques for future research work. In this study, we seek to answer the questions in the Table 1.

Table 1. Research questions and purpose.

	Question	Purpose
Q1	What are the best NLP methods used in medical text classification?	To Describe the best methods used in the medical classification framework based on the evaluation metrics. And identify the challenges
Q2	How are medical text classification datasets constructed?	To study the composition and description of medical texts in the classification task.
Q3	In terms of data, what are the most common problems that medical text classification can solve?	To understand and highlight the common problems and challenges addressed in medical text-based problem solving.
Q4	What are the mostly used evaluation metrics of medical document classification?	To identify the different mostly metrics used in medical document classification

2. Material and Methods

The major purpose of our systematic study is to highlight current issues that text classification systems have to cope with in order to analyze biological text information. The insights discovered in this study will be utilized as a starting point for future research in this area. Table 1 outlines the main questions we hoped to address by the conclusion of this research. To achieve this review system, we have merged the methodologies employed by Sofia Zahia et al. in [5] and those by Urdaneta-Ponte et al. in [6]. On the basis of these strategies, we shall produce our review article.

2.1. Data Collection

The articles in the databases were chosen using a variety of methodologies and eligibility criteria which are briefly presented in the following subsections. We initially applied the filter of papers collected from various databases, followed by the filters based on the qualifying criteria. Each metric was used to pick publications that were relevant to our research.

2.1.1. Searched Databases

Several databases were used to conduct the literature search, including Web of Science, Arxiv, IEEE, ACM, Scopus, and MDPI. The selection time of articles was limited from 1 January 2016 to 10 July 2022. Several factors influenced our choice of publications, including the search terms, which covered studies published on biomedical text classification as well as image-text classification.

2.1.2. Search Terms

Several terms were used to look for works on biomedical text classification task; some of these terms were combined to enhance the search in multiple databases. The terms chosen for the selections were: "text classification", "medical text", "medical document", "healthcare", "patient records", "text prediction", "nursing notes", "Natural Language Processing", "text-image", "biomedical text classification", "nursing text prediction", "prediction", "classification", "image", "text", "Machine learning", "transformers", "LSTM", "GRU", "clinical" and "clinical text".

2.1.3. Inclusion Criteria

The initial stage in the selection procedure was to look through titles and abstracts to discover papers that fulfilled the needed criteria. Then duplicates were removed. Because medical record classification encompasses numerous applications, such as the detection and classification of text in nursing picture notes, relevant matching publications were obtained and classified.

2.1.4. Exclusion Criteria

The following exclusion criteria were applied to select the papers: date of publication, type of publication, ranking of the journal in case the paper was published in an international journal, type of problem studied in the paper, and finally the number of citations of the paper.

Figure 1 depicts the revised flowchart for PRISMA in [6]. This systematic review's data gathering approach followed a logical progression until only 33 publications were deemed appropriate for analysis. Each database indicated in Section 2.1.1 as a source of publications was recognized, along with a total of 894 papers for the selection process. Following the identification stage, a screening was conducted to eliminate duplicate documents. Certainly, a journal or conference-published work may be archived in at least one research database. In this stage, 708 papers were retained after screening. The last step was to apply the eligibility criteria to select the best articles according to the Sections 2.1.3 and 2.1.4 Table 2. In the first screening, 97 articles were kept and 611 were rejected based on their titles; in the subsequent screening, 48 papers were retained and 49 were rejected based on their abstracts. After a thorough reading of each manuscript, 33 were ultimately chosen for study, while 15 were discarded.

Table 2. Exclusion criterion description.

Criteria	Description
Date	The publications included for this research were screened between 1 January 2016 and 10 July 2022. The quantity of relevant articles to filter dictated the selection of this range. Given the fast advancement of deep learning models and machine learning.
Type of publications	filtering was performed on two categories of publications, papers published at international conferences and articles published in international journals.
Ranking	To determine the finest articles, we used the ranking count of papers published in journals systematically. This criteria was not applied to papers presented at conferences. We examined the rankings Q1, Q2, and Q3 for the publications in the journals.
Type of problem	Only articles on biomedical text or image-text classification were evaluated for this criteria.
Citations	This criteria was given less weight, particularly for articles published recently, such as those from 2021 and 2022

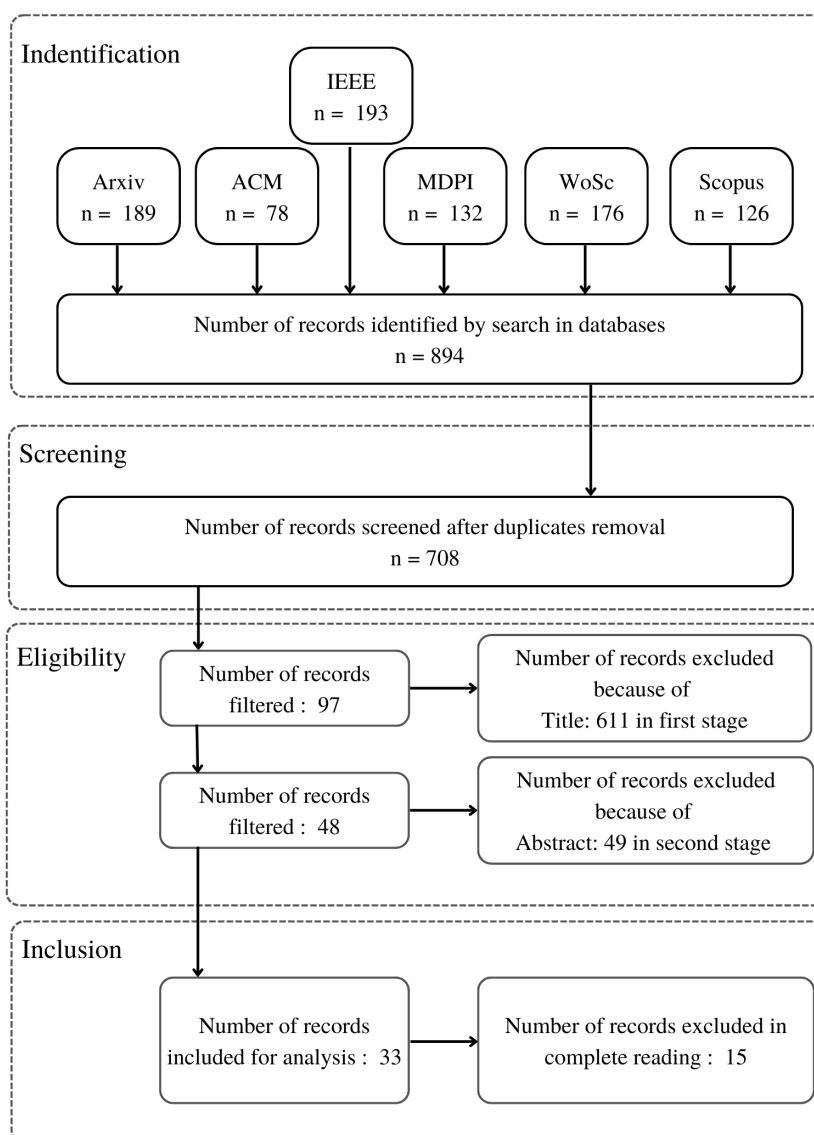


Figure 1. Paper selection flow diagram for text classification in biomedical domain.

2.2. Quality Metrics

The quality metrics in Table 3 were used to evaluate the relevance of each paper selected during the analysis see Table A2. The total score of 15 for the papers published in the international journal and the total score of 11 for those published in an international conference because the metric M11 for the ranking does not concern the conference papers. The following ratings have been defined for both types of papers: For journal papers, the paper is rating Excellent if the score is between 13–15; between 10–12 very good; between 7–9 good; between 3–6 sufficient and between 0–2 deficient. For the conference papers, the paper is rating Excellent if the score is between 9–11; between 6–8 very good; between 4–5 good, if this 3, sufficient and between 0–2 deficient.

Table 3. Quality metrics of paper selection.

Category Metric	Metric	Description	Value	Weight
Metrics based on the text content of the paper (5 points)	M1	Provide a clear and balanced summary according to the context of the problem solved in the paper	(No/Yes) [0,1]	1
	M2	Provide details of the model's performance metrics and the entire evaluation process	[0,1]	1
	M3	Implement one or more medical text classification models	[0,1]	1
	M4	compares the results with other similar work and presents the limitations of this work	[0,1]	1
	M5	Contains a deep and rich discussion on the result obtained	[0,1]	1
Other quality metrics (10 points)	M6	Innovation	[0,1]	1
	M7	The dataset used in the research is a benchmark or it has been made publicly available	[No,Yes] (0–1)	1
	M8	Performance (Accuracy)	Regarding the performance, if the percentage of quality of result is between 60–70% (0.5), between 71–80% (1), between 81–90% (1.5) and 91% + (2) otherwise (0)	2
	M9	Citation	If the paper is cited 0 times (0), 1–4 times (0.5) and cited 6+ (1)	1
	M10	Availability of source code	[0,1]	1
	M11	Journal ranking	If rank = Q1 then (4), rank = Q2 then (3) rank = Q3 then (2) and if rank = Q4 then 1	4

3. Results

All the papers selected following the steps of the flow diagram in Figure 1 were included in the analysis. Table A1 summarizes the selections made in this review paper. All the metrics of Table 3 were applied to evaluate the selected papers, and the result of this evaluation is in Table 4. The whole evaluation process is presented in Table A2 in the Appendix A. In addition, to answer the questions in Table 1, the evaluation of the different text classification databases used in each selected paper was conducted in order to discover new challenges in the data and their influence in building the models. Finally, an evaluation of the frequency distribution of the selected papers by location, publication database, and type (Journal/Conference) was done, followed by an evaluation of the frequency distribution by ranking and year of publication.

3.1. Quality Metric Result

To make sure that the evaluations of each article's quality parameters were correct, the right measurements were taken using the defined indicators Table 3. Each article was ranked on a scale from deficient to excellent based on how much it added to our systematic review. They were judged based on the degree of innovation, the details of the proposal, validation, results and analysis, ranking, and the number of citations. Table A2 shows how each article did in terms of the score, and Table 4 gives a full summary of Table A2.

Table 4. Metric result.

Score	No Journal	No Conference	Total
Excellence	3	6	9
Very good	14	7	21
Good	2	1	3
Sufficient	0	0	0
Deficient	0	0	0
Total	19	14	33

3.2. Text Classification Methods Performance According Datasets Used

The best approaches in relation to the database were identified in two ways, based on one of the performance indicators such as accuracy, precision, recall, and F1-score. First, papers that utilized the same datasets were grouped together, and then all publications were grouped together. Two datasets that used more than one publication were identified such as: **MIMIC III** with five papers and **AIM** had two papers. With MIMIC III, the BioBERT approach in [7] has an Accuracy of 90.05% and is regarded the best method for this classification, whereas the LSTM method in [8] gets a score of 91.00%. In the publication [9], the BiGRU technique achieves 97.73% of accuracy on the AIM dataset. The synthesis is shown in Table 5.

Table 5. Performance of the most frequent text classification methods and database used.

Methods	Dataset	Accuracy	Precision	Recall	F1-Score
TAGS [10]	MIMIC-III dataset	82.00%	-	-	-
SWAM-text CNN [11]	MIMIC-III dataset	-	-	-	60.00%
BioBERT [7]	MIMIC-III database	90.05%	77.37%	-	48.63%
BERT-base [12]	MIMIC III dataset	82.30%	-	-	82.20%
LSTM [8]	MIMIC-III dataset	-	-	-	91.00%
QC-LSTM; BiGRU [13]	AIM dataset	96.72%	-	-	-
BiGRU [9]	AIM dataset	97.73%	-	-	-

Considering only the performance measurement values of the different classification techniques in general [14], without basing them on the direct comparison with the data used and their statistical distribution, the problem to be solved, we observe that, BERT-based technique in [15–17], GRU [9,13], BiGRU [9,13] and LSTM [8,18] produced a good performance on most of the problems studied in the different papers identified in our study. In addition, the methods that present the good performance but represented only once in the papers studied, are Random forest [19], CNN-MHA-BLSTM [20], Double-channel (DC-LSTM) [21], MT-MI-BiLSTM-ATT [22] and QC-LSTM [13] Table 6.

Table 6. Performance obtained on different text classification methods used in each paper.

Methods	Dataset	Accuracy	Precision	Recall	F1-Score
TAGS [10]	MIMIC-III	82.00%	-	-	-
SWAM-text CNN [11]	MIMIC-III full dataset; MIMIC-III 50 dataset	-	-	60.00%	-
BioBERT [7]	MIMIC-III database	90.05%	77.37%	-	48.63%
BERT-base [12]	PubMed abstract; MIMIC III	82.30%	-	-	82.20%
LSTM [8]	MIMIC-III; CSU dataset	-	-	-	91.00%
QC-LSTM; BiGRU [13]	Hallmarks dataset; AIM dataset	96.72%	-	-	-
BiGRU [9]	TCM—Traditional Chinese medicine dataset; CCKS dataset; Hall-marks—corpus dataset; AIM—Activating invasion and metastasis dataset	97.73%	-	-	-
Conv-LSTM [23]	EMR text data (benchmark)	83.30%	-	-	-
MT-MI-BiLSTM-ATT [22]	EMR data set comes from a hospital (benchmark)	93.00%	-	-	87.00%
SVM (Sigmoid Kernel) [24]	EMR data from outpatient visits during 2017 to 2018 at a public hospital in Surabaya City, Indonesia (benchmark)	88.40%	81.28%	76.46%	78.80%
BERT [15]	THUCNews; iFLYTEK	96.63%	96.64%	96.63%	96.61%
BERT-based [16]	COVID-19 fake news dataset” by Sumit Bank; extremist-non-extremist dataset	99.71%	98.82%	97.84%	98.33%
LSTM [18]	SQuAD	-	98.00%	98.00%	98.00%
MedTS [25]	MIMICSQL	88.00%	-	-	-
CNN [26]	DingXiangyisheng’s question and answer module (benchmark)	86.28%	-	-	-
CRNN [27]	iDASH dataset; MGH dataset	-	-	-	84.50%
Double-channel (DC-LSTM) [21]	cMedQA medical diagnosis dataset; Sentiment140 Twitter dataset	97.20%	91.80%	91.80%	91.00%
CNN Based model [28]	EMR Progress Notes from a medical center (benchmark)	58.00%	58.20%	57.90%	58.00%
BidirLSTM [29]	clinical nursing shift notes (benchmark)	-	-	-	-
Random forest [19]	Text dataset from NHLS-CDW	95.25%	94.60%	95.69%	95.34%
SVM [30]	Medical records from from digital health (benchmark)	80.00%	-	-	-
CNN-MHA-BLSTM [20]	EMR texte dataset (benchmark)	91.99%	-	-	92.03%
MLP [31]	EMR dataset (benchmark)	82.00%	-	-	82.00%
MobileNetV2 [32]	RVL-CDIP dataset	-	-	-	82.00%
Med2Vec [33]	CHOA dataset	-	-	91.00%	-
biGRU [34]	RCV1/RCV2 dataset	-	-	-	84.00%
Capsule+LSTM [35]	Chinese electronic medical record dataset	-	-	-	73.51%
BioLinkBERT [36]	MedQA-USMLE; MMLU-professional medicine	50.00%	-	-	-
Bert-based [17]	Harvard obesity 2008 challenge dataset	94.70%	-	-	-

3.3. Frequency Result According Geographical Distribution and Type of Publication

As we can see in Table 7, several studies based on text classification were carried out in Asia with a percentage of 51.5, which is half of all the papers analyzed in our research. With 6.1 percent, Africa has a low representation papers, whereas America and Europe both have 21.2 percent. It is also shown in this study that it has 57.6% of papers published in journals and 42.4% published in conferences. In Figure 2, we present the different frequencies of the selected papers according to regions, continents, search database and type of publications. The most frequented region with published studies on medical text classification was the Eastern Asia region. In addition, among the search databases Web Of Science was the most frequented database after filtering.

Table 7. Number and frequency of research database, conference or journal and by geographical distribution of publication.

Parameters	Category	Frequency	
		No. Papers	Percentage
Location	Southern Africa	1	3.0%
	Africa	1	3.0%
	Eastern Asia	13	39.4%
	Southern Asia	2	6.1%
	Western Asia	1	3.0%
	South-Eastern Asia	1	3.0%
	Asia	17	51.5%
	Northern Europe	3	9.1%
	Eastern Europe	1	3.0%
	Southern Europe	2	6.1%
	Western Europe	2	6.1%
	Europe	8	24.3%
	Northern America	7	21.2%
	America	7	21.2%
Database	Arxiv	7	21.2%
	ACM	2	6.1%
	MDPI	2	6.1%
	WoSc	10	30.3%
	Scopus	4	12.1%
	IEEE	8	24.2%
	Type of publication	Conference	14
Journal		19	57.6%

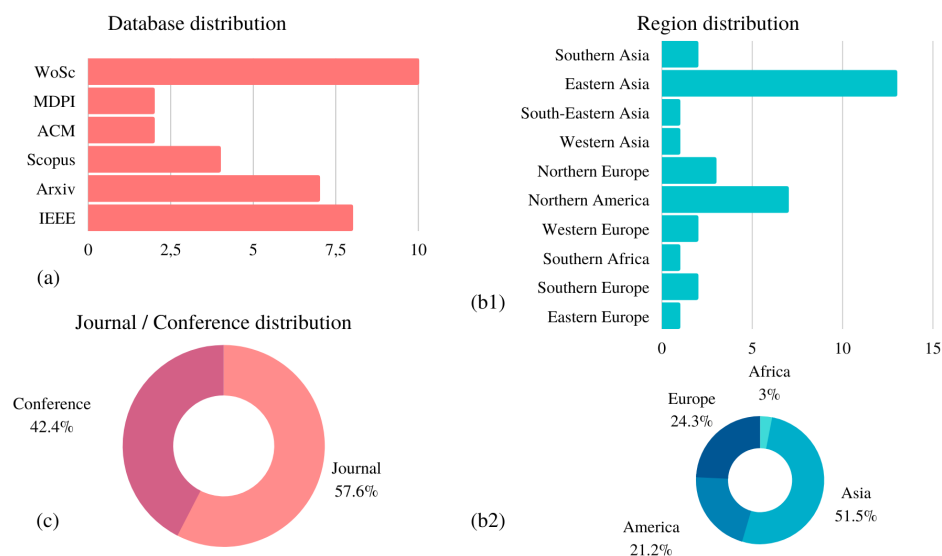


Figure 2. Frequency of research database, conference or journal and by geographical distribution of publication. (a) is the distribution of the different databases from which we collected papers in this study. (b1,b2) represent respectively the distribution of the selected papers by region and by continent. (c) the distribution of papers by conference and journal.

3.4. Paper Publication Map by Country

The map in Figure 3 describes the degree of contribution of countries in Artificial Intelligence (NLP) in biomedical text classification from 1 January 2016 to 10 July 2022. China largely dominates in this study, followed by the USA, this result coincides with the result published in Artificial Intelligence Index Report 2022 [37].

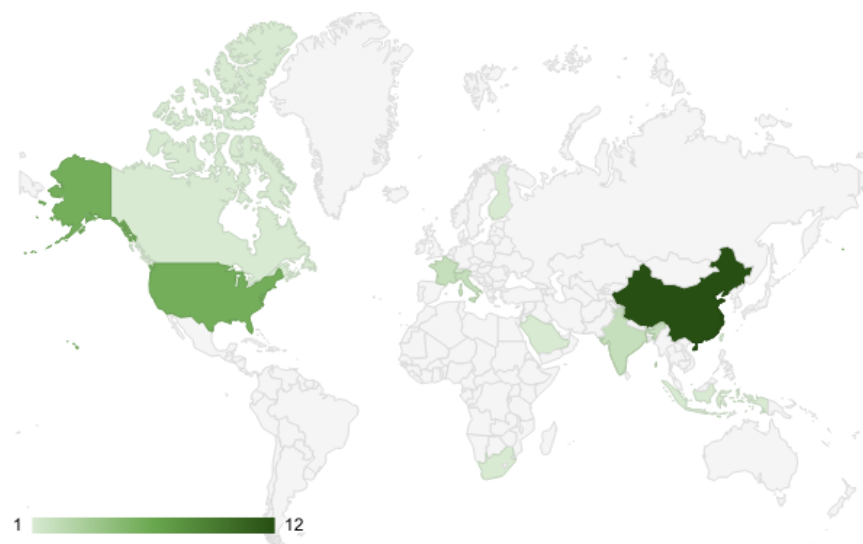


Figure 3. Degree of contribution of countries in Artificial Intelligence (NLP) in biomedical text classification.

3.5. Frequency Result According Year and Journal Ranking

Table 8 shows the frequency and number of papers per year and per ranking. As mentioned above, the time range considered for the selection of articles for analysis in this systematic review was from 1 January 2016 to 10 July 2022. The year 2020 counted 11 papers and represents 33.3% of the papers compared to other years with low representatives in the classification of biomedical texts. In addition, the ranking was considered as one of the major eligibility criteria of papers for analysis in the case of papers published in journals. All the papers whose category is none are published in an international conference.

Considering the ranking, more of the selected papers, i.e., 12 out of 19 papers published in journals, were of Q1. Figure 4 presents the different frequencies in the analysis for the year and the ranking distribution.

Table 8. Number and frequency of year and paper ranking.

Parameters	Category	Frequency	
		No. Papers	Percentage
Year	2016	1	3.0%
	2017	3	9.1%
	2018	1	3.0%
	2019	5	15.1%
	2020	11	33.3%
	2021	8	24.3%
	2022	4	12.1%
Paper ranking	Q1	12	36.4%
	Q2	3	9.1%
	Q3	3	9.1%
	Conference	15	45.5%

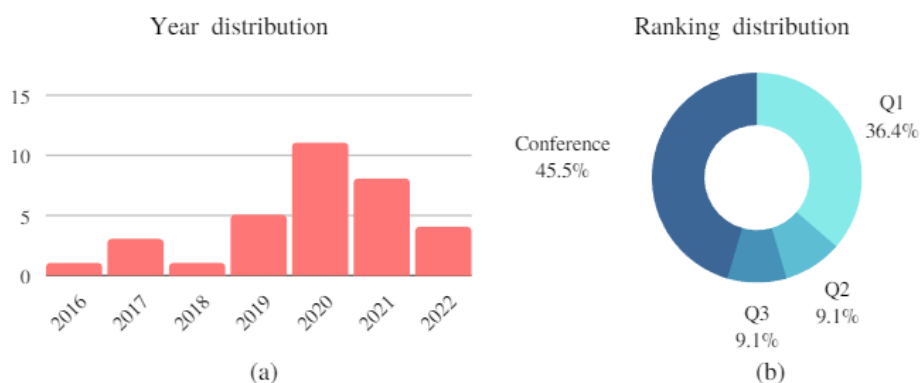


Figure 4. (a) Represent the frequency by year and (b) the distribution by conference and paper ranking.

4. Discussion

Text classification in biomedical field plays an important role in the rapid search (diagnosis) of a disease from the patient record, hospital administration, and even the treatment appropriate for a specific case, as the volume of patient medical records continues to increase significantly. Each year, new classification methods with high classification accuracy are proposed, while the performance of older [38–40] NLP methods is enhanced through the use of alternative approaches such as optimization and other type of algorithm based on transformers architecture [12,40–42] and XLNet [43], data-centric technique and many others. The data-centric technique presents a challenge in enhancing the performance of biomedical text classification methods [44]. The observation is that the majority of methods have been pre-trained with text databases in a generic context without any prior specificity. In other words, a model that has been pre-trained with biomedical data will adapt better when re-trained with new data for a biomedical domain. In this context, we discuss the data-centric problem, which must be a key consideration when developing models tailored to specific case. Another challenge in the classification of biomedical texts is the data quality. We found two kinds of datasets in the articles we looked at: those made public by research institutes and labs [7,9,13,15–17,21], and those that any reference

(benchmark) could use without more information. When training the models to give good results, it is important to think about how good the data [45] are. This quality can be made sure of by thinking about the whole process of collecting and preprocessing the data until it is ready to be used for classification tasks.

Before performing the classification task, biomedical texts can be derived from a variety of sources [46,47]. We find data in medical reports that are already in text form, as well as notes taken by doctors or nurses during consultations that are scanned images. Depending on the context of the problem and the goal to be achieved, several approaches can be used with these types of data. Alternatively, the data can be represented in both formats, or a radio image is accompanied by text that explains the image. Depending on the expected result, several methods can be combined in the text classification process in image-text data [13]. To complete these tasks, methods based on CNN architectures [48,49] are frequently used [13,50].

The classification of biomedical texts is involved in several important problems that physicians are expected to solve. These issues can sometimes be large challenges in multiple steps. This can be conducted in diagnosis [11,28], patient treatment [11], or even effective representations of concepts such as diagnoses, drugs, procedure codes, and patient visits [33], as well as in the quick search of a document or disease classification [23]. Pathologies from clinical notes [23] and much more. In all of these ways, it is harder to classify texts in the biomedical field than in other fields in general. This is because biomedical texts include both medical records and medical literature, which are both important sources of clinical information. However, medical texts have hard-to-understand medical terms and measurements that cause problems with high dimensionality and a lack of data [9]. All of these problems are very important when it comes to the task of classifying biomedical text.

In the biomedical text classification task, as in most classification problems in general [51], the model evaluation metrics are the same. In all the papers studied in our systematic review, the metrics identified are Accuracy, Recall, F1-score, Precision, Average precision, Average recall, and Average F1-score. These metrics are the most commonly used to evaluate text classification models. As in this study, the different methods used in each paper analyzed, used at least one of these metrics except for one paper [52] which used Spearman.C.C. metric [53].

5. Conclusions and Perspectives

This study discusses the various challenges in the task of biomedical text classification by focusing on several aspects such as the challenge in method performance, discovering the structure of biomedical data for the classification task, listing the various problems and challenges that text classification can solve in the biomedical domain, and finally reviewing the most commonly used metrics to evaluate biomedical text classification methods. We discovered two significant issues linked to the approaches utilized for biomedical text classification by reviewing the various literature chosen for examination in this research. First, there is the data-centric, which is explained by the fact that most transfer learning using pre-trained techniques employ a dataset of broad text classification settings. However, the biomedical issue includes various medical words that may be classified as process, therapy, medicine, and diagnosis. Because the contextual representation of medical language is quite poor in the general context, this already creates a contextual challenge when training to generate the best outcomes. This necessitates only training models on a huge number of biological data in order to execute transfer learning more correctly. There are certain approaches that are exclusively trained with biomedical databases, such as BioBERT [7] and BioLinkBERT [17], but the task remains to study as many ways as possible with just biomedical databases to enhance biomedical text classification outcomes. This is the first problem that affects how well the text classification methods work in biomedical domain.

Another issue to consider is data quality. We found two types of datasets in the articles we looked at: those made public by research institutes and labs, and those accessed by any reference (benchmark) without more information. The quality of the data is a key factor to

consider while training the models to deliver good outcomes. This quality may be assured by considering the whole collecting and pre-processing process until the data set is ready as an usable source for classification tasks. Several other challenges can be described by taking into account several aspects that we have not addressed in this work. Some of the challenges we have discussed are the most common ones in our overall study.

In the perspective, to significantly advance research in the biomedical field, it is preferable to make well-preserved and verified data more widely available in order to assist research and overcome data quality [54–56] in biomedical classification challenges. Because of domain drifts among different institutes, the cooperation between research laboratories, universities and other research entities, would be an action to be strengthened in order to create a great network of scientific sharing of scarce resources such as data to advance research. Joint work sessions between domain experts should be a good procedure to validate the dataset as a common resource for scientific research of text classification. Finally, a policy of simplification of data sharing, which is often confidential, would be an essential point among many others to be defined to answer the problem of data deficiency. Most of the models used in the papers selected in this study are based on Deep learning. The interpretability of robust models is an important aspect in clinical research. Accuracy and reliability are also important aspects in biomedical research field. Whether one uses simple models based on statistical learning or robust models based on Deep learning, whatever their performance, the interpretability and reliability aspect would be very important to take into account, to validate the results for a clinical research.

Author Contributions: Conceptualization, C.Y.K., A.C.-V. and B.G.-Z.; methodology, C.Y.K. and A.C.-V.; formal analysis, C.Y.K.; investigation, C.Y.K.; writing—original draft preparation, C.Y.K.; writing—review and editing, C.Y.K., A.C.-V. and B.G.-Z.; supervision, B.G.-Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank and acknowledge the eVida research group of the University of Deusto, recognized by the Basque Government with the code IT1536-22, and the ICCRF Colon Cancer Challenge for its untiring support and commitment to providing us with the resources necessary to carry out the study, until its finalization.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Summary of papers filtering aspects.

P.	Year	J/C	Loc	Database	Methods	Dataset	Best Method	Metric (Best)	Rank	Cite
[10]	2020	J	India	Scopus	Fuzzy similarity-based data cleansing approach, supervised multi-label classification models, MLP, KNN, KNN as OvR, AGgregated using fuzzy Similarity (TAGS)	MIMIC-III	TAGS	Ac: 82.0	Q1	18
[23]	2020	J	India	IEEE	MLP, ConvNet, LSTM, Bi-LSTM, Conv-LSTM, Seg-GRU	EMR text data (benchmark)	Conv-LSTM	Ac: 83.3	Q1	7
[22]	2020	J	China	IEEE	BiLSTM, CNN, CRF layer. In particular, they used BiLSTM and CNN to learn text features and CRF as the last layer of the model; MT-MI-BiLSTM-ATT	EMR data set comes from a hospital (benchmark)	MT-MI-BiLSTM-ATT	Ac: 93.0 F1: 87.0	Q1	3
[57]	2021	C	China	IEEE	ResNet; BERT-BiGRU; ResNet-BERTBiGRU	Text-image data (benchmark)	ResNet-BERTBiGRU	Mavg.P: 98.0 Mavg.R: 98.0 Mavg.F1: 98.0	None	0
[58]	2021	C	Indonesia	IEEE	SVM (Linear Kernel); SVM (Polynomial Kernel); SVM (RBF Kernel); SVM (Sigmoid Kernel)	EMR data from outpatient visits during 2017 to 2018 at a public hospital in Surabaya City, Indonesia (benchmark)	SVM (Sigmoid Kernel)	R: 76.46; P: 81.28; F1: 78.80; Ac: 91.0	None	0
[24]	2020	C	China	IEEE	GM; Seq2Seq; CNN; LP; HBLA-A (This model can be seen as a combination of BERT and BiLSTM.)	ARXIV Academic Paper Dataset (AAPD); Reuters Corpus Volume I (RCV1-V2)	BLA-A	Micro-P: 90.6; Micro-R: 89.2; Micro-F1: 89.9	None	1
[15]	2021	C	China	IEEE	Text CNN; BERT; ALBERT	THUCNews; iFLYTEK	BERT	Ac: 96.63; P: 96.64; R: 96.63; F1: 96.61	None	0
[16]	2022	J	Saudi Arabia	Scopus	BERT-base; BERT-large; RoBERTa-base; RoBERTa-large; DistilBERT; ALBERT-base-v2; XLM-RoBERTa-base; Electra-small; and BART-large	COVID-19 fake news dataset" by Sumit Bank; extremist-non-extremist dataset	BERT-base	Ac: 99.71; P: 98.82; R: 97.84; F1: 98.33	Q3	3
[18]	2020	C	UK	WoSc	LSTM; Multilingual; BERT-base; SCIBERT; SCIBERT 2.0	SQuAD	LSTM	P: 98.0; R: 98.0; F1: 98.0	None	10

Table A1. Cont.

P.	Year	J/C	Loc	Database	Methods	Dataset	Best Method	Metric(best)	Rank	Cite
[13]	2021	J	China	WoSc	CNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, logistic regression, naïve Bayesian classifier (NBC), SVM, and BiGRU. QC-LSTM; BiGRU	Hallmarks dataset; AIM dataset	QC-LSTM	AC: 96.72	Q3	1
[25]	2021	J	China	WoSc	Seq2Seq; SQLNet; PtrGen; Coarse2Fine; TREQS; MedTS	MIMICSQL	MedTS	AC: 88.0	Q2	0
[26]	2029	J	China	WoSc	CNN; LSTM	DingXiangyisheng’s question and answer module (benchmark)	CNN	AC: 86.28	Q1	1
[27]	2027	J	USA	WoSc	Tf-Idf CRNN	iDASH dataset; MGH dataset	CRNN	AUC: 99.1; F1: 84.5	Q1	59
[21]	2027	J	China	WoSc	CNN; LSTM; CNN-LSTM; GRU; DC-LSTM	cMedQA medical diagnosis dataset; Sentiment140 Twitter dataset	DC-LSTM	Ac: 97.2; P: 91.8; R: 91.8; F1: 91.0	Q3	1
[28]	2020	J	Taiwan	WoSc	CNN; CNN Based model	EMR Progress Notes from a medical center (benchmark)	CNN Based model	Ac: 58.0; P: 58.2; R: 57.9; F1: 58.0	Q1	2
[9]	2019	J	China	Scopus	CNN; RCNN; LSTM; AC-BiLSTM; SVM; Logistic-Regression	TCM—Traditional Chinese medicine dataset; CCKS dataset; Hallmarks—corpus dataset; AIM—Activating invasion and metastasis dataset	BiGRU	Hallmarks, Ac: 75.72; TCM, Ac: 89.09; CCKS, Ac: 93.75; AIM, Ac: 97.73	Q2	15
[59]	2021	J	China	WoSc	RoBERTa; ALBERT; transformers-sklearn based	TrialClassification, BC5CDR, DiabetesNER, and BIOSSES	transformers-sklearn based	Mavg-F1: 89.03	Q1	2
[7]	2020	C	UK	WoSc	BioBERT; Bert	MIMIC-III database	BioBERT	Ac: 90.05; Precision: 77.37; F1: 48.63	None	0
[29]	2019	C	Finland	IEEE	BidirLSTM, LSTM, CNN, fastText, BoWLinearSVC, RandomForest, Word Heading Embeddings, Most Common, Random	clinical nursing shift notes (benchmark)	BidirLSTM	Avg-R: 54.35	None	3

Table A1. Cont.

P.	Year	J/C	Loc	Database	Methods	Dataset	Best Method	Metric(best)	Rank	Cite
[19]	2021	J	South Africa	MDPI	Random forest, SVMLinear, SVMRadial	text dataset from NHLS-CDW	Random forest	F1: 95.34; R: 95.69 P: 94.60 Ac: 95.25	Q2	2
[30]	2020	J	Italy	Scopus	SVM	Medical records from from digital health (benchmark)	SVM	Mavg-P: 88.6; Mavg-Ac: 80.0	Q1	27
[8]	2020	J	UK	WoSc	LSTM; LSTM-RNNs; SVM, Decision Tree; RF	MIMIC-III; CSU dataset	LSTM	F1: 91.0	Q1	13
[20]	2020	C	USA	ACM	CNN-MHA-BLSTM; CNN, LSTM	EMR texte dataset (benchmark)	CNN-MHA-BLSTM	Ac: 91.99; F1: 92.03	None	22
[31]	2019	C	USA	IEEE	MLP	EMR dataset (benchmark)	MLP	Ac: 82.0; F1: 82.0	None	1
[12]	2019	C	USA	Arxiv	BERT-base, ELMo, BioBERT	PubMed abstract; MIMIC III	BERT-base	Ac: 82.3	None	0
[32]	2020	J	France	Arxiv	MLP, CNN CNN 1D, MobileNetV2, MobileNetV2 (w/ DA)	RVL-CDIP dataset	MobileNetV2	F1: 82:0	Q1	55
[33]	2016	C	USA	ACM	Med2Vec	CHOA dataset	Med2Vec	R: 91.0	None	378
[52]	2018	C	Canada	Arxiv	word2vec, Hill, dict2vec	MENd dataset; SV-d dataset	word2vec	Spearman.C.C: 65.3	None	37
[34]	2017	C	Switzerland	Arxiv	biGRU, GRU, DENSE	RCV1/RCV2 dataset	biGRU	F1: 84.0	None	34
[11]	2021	J	China	Arxiv	Logistic regression; SWAM-CAML; SWAM-text CNN	MIMIC-III full dataset; MIMIC-III 50 dataset	SWAM-text CNN	F1: 60.0	Q1	6
[35]	2022	J	China	MDPI	LSTM, CNN, GRU, Capsule+GRU, Capsule+LSTM	Chinese electronic medical record dataset	Capsule+LSTM	F1: 73.51	Q2	2
[36]	2022	C	USA	Arxiv	BERTtiny; LinkBERTtiny, GPT-3, BioLinkBERT, UnifiedQA	MedQA-USMLE; MMLU-professional medicine	BioLinkBERT	Ac: 50.0	None	4
[17]	2022	J	USA	Arxiv	CNN, LSTM, RNN, GRU, Bi-LSTM, Transformers, Bert-based	Harvard obesity 2008 challenge dataset	Bert-based	Ac: 94.7	Q1	0

Table A2. Results of the application of the eligibility criteria to the filtered papers.

P.	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	Result
[10]	1	1	1	0	1	1	1	1	1	0	4	12/15
[23]	1	1	1	0	1	1	1	1	1	0	4	12/15
[22]	1	1	1	0	1	1	0	2	1	0	4	12/15
[57]	1	1	1	0	1	0	0	2	0	0	Conf.	6/11
[58]	1	1	0	0	1	0	0	2	0	0	Conf.	5/11
[24]	1	1	1	1	1	0	1	2	1	0	Conf.	9/11
[15]	1	1	1	0	1	0	0	2	0	0	Conf.	6/11
[16]	1	1	1	0	1	0	0	2	0.5	0	2	8.5/15
[18]	1	1	1	1	1	0	1	2	1	0	Conf.	9/11
[13]	1	1	1	1	1	0	1	2	0.5	1	2	11.5/15
[25]	1	1	1	0	1	0	1	2	1	0	3	11/15
[26]	1	1	1	1	1	0	1	1.5	0.5	0	4	12/15
[27]	1	1	1	1	1	0	1	2	1	1	4	14/15
[21]	1	1	1	1	1	0	1	2	0.5	0	3	11/15
[28]	1	1	0	1	1	0	1	0.5	0.5	0	4	9/15
[9]	1	1	1	1	1	0	1	2	1	0	3	12/15
[59]	1	1	1	1	1	0	1	1.5	0.5	0	4	12/15
[7]	1	1	1	1	0	0	1	1.5	0	0	Conf.	6.5/11
[29]	1	1	1	1	1	0	1	0	0.5	0	Conf.	6.5/11
[19]	1	1	1	1	1	0	1	2	0.5	0	3	11.5/15
[30]	1	1	0	1	1	1	0	1.5	1	0	4	11.5/15
[8]	1	1	1	1	1	0	0	2	1	0	4	12/15
[20]	1	1	1	1	1	1	0	2	1	0	Conf.	9/11
[31]	1	1	0	1	1	1	0	1.5	0.5	0	Conf.	7/11
[12]	1	1	1	1	1	0	0	0	0	1	Conf.	6/11
[32]	1	1	1	1	1	0	2	1.5	1	1	4	14.5/15
[33]	1	1	0	1	1	1	2	1	1	1	Conf.	10/11
[52]	1	1	1	1	1	1	0.5	1	1	1	Conf.	9.5/11
[34]	1	1	1	1	1	1	1.5	1	1	1	Conf.	10.5/11
[11]	1	1	1	1	1	0	0	1	1	1	4	12/15
[35]	1	1	1	1	1	0	1	1	0.5	1	3	11/15
[36]	1	1	1	1	1	1	1	0	0.5	1	Conf.	8/11
[17]	1	1	1	1	1	0	1	2	0	1	4	13/15

References

1. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
3. World Health Organization. The International Classification of Diseases, 10th Revision. 2015. Available online: <https://icd.who.int/browse10/2015/en> (accessed on 4 August 2021).
4. Chen, P.; Wang, S.; Liao, W.; Kuo, L.; Chen, K.; Lin, Y.; Yang, C.; Chiu, C.; Chang, S.; Lai, F. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. *JMIR Med. Inform.* **2021**, *9*, e23230. [[CrossRef](#)] [[PubMed](#)]
5. Zahia, S.; Zafirain, M.B.; Sevillano, X.; González, A.; Kim, P.J.; Elmaghraby, A. Pressure injury image analysis with machine learning techniques: A systematic review on previous and possible future methods. *Artif. Intell. Med.* **2020**, *102*, 101742. [[CrossRef](#)] [[PubMed](#)]

6. Urdaneta-Ponte, M.C.; Mendez-Zorrilla, A.; Oleagordia-Ruiz, I. Recommendation Systems for Education: Systematic Review. *Electronics* **2021**, *10*, 1611. [CrossRef]
7. Amin-Nejad, A.; Ive, J.; Velupillai, S. LREC Exploring Transformer Text Generation for Medical Dataset Augmentation. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Palais du Pharo, Marseille, France, 11–16 May 2020; Available online: <https://aclanthology.org/2020.lrec-1.578> (accessed on 4 August 2021).
8. Venkataraman, G.R.; Pineda, A.L.; Bear Don't Walk, O.J., IV.; Zehnder, A.M.; Ayyar, S.; Page, R.L.; Bustamante, C.D.; Rivas, M.A. FasTag: Automatic text classification of unstructured medical narratives. *PLoS ONE* **2020**, *15*, e0234647. [CrossRef]
9. Qing, L.; Linhong, W.; Xuehai, D. A Novel Neural Network-Based Method for Medical Text Classification. *Future Internet* **2019**, *11*, 255. [CrossRef]
10. Gangavarapu, T.; Jayasimha, A.; Krishnan, G.S.; Kamath, S. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowl.-Based Syst.* **2020**, *190*, 105321. [CrossRef]
11. Hu, S.; Teng, F.; Huang, L.; Yan, J.; Zhang, H. An explainable CNN approach for medical codes prediction from clinical text. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 256. [CrossRef]
12. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv* **2019**, arXiv:1906.05474.
13. Prabhakar, S.K.; Won, D.O. Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. *Comput. Intell. Neurosci.* **2021**, *2021*, 9425655. [CrossRef]
14. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical Transformers for Long Document Classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 838–844. [CrossRef]
15. Fang, F.; Hu, X.; Shu, J.; Wang, P.; Shen, T.; Li, F. Text Classification Model Based on Multi-head self-attention mechanism and BiGRU. In Proceedings of the 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 11–13 December 2021; pp. 357–361. [CrossRef]
16. Qasim, R.; Bangyal, W.H.; Alqarni, M.A.; Ali, Almazroi, A. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *J. Healthc. Eng.* **2022**, *2022*, 3498123. [CrossRef]
17. Lu, H.; Ehwerhemuepha, L.; Rakovski, C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med. Res. Methodol.* **2022**, *22*, 181. [CrossRef]
18. Schmidt, L.; Weeds, J.; Higgins, J. Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks. *arXiv* **2020**, arXiv:2001.11268.
19. Achilonu, O.J.; Olago, V.; Singh, E.; Eijkemans, R.M.J.C.; Nimako, G.; Musenge, E. A Text Mining Approach in the Classification of Free-Text Cancer Pathology Reports from the South African National Health Laboratory Services. *Information* **2021**, *12*, 451. [CrossRef]
20. Shen, Z.; Zhang, S. A Novel Deep-Learning-Based Model for Medical Text Classification. In Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition (ICCP 2020), Xiamen, China, 30 October–1 November 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 267–273. [CrossRef]
21. Liang, S.; Chen, X.; Ma, J.; Du, W.; Ma, H. An Improved Double Channel Long Short-Term Memory Model for Medical Text Classification. *J. Healthc. Eng.* **2021**, *2021*, 6664893. [CrossRef]
22. Wang, S.; Pang, M.; Pan, C.; Yuan, J.; Xu, B.; Du, M.; Zhang, H. Information Extraction for Intestinal Cancer Electronic Medical Records. *IEEE Access* **2020**, *8*, 125923–125934. [CrossRef]
23. Gangavarapu, T.; Krishnan, G.S.; Kamath, S.; Jeganathan, J. FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 1151–1169. [CrossRef]
24. Cai, L.; Song, Y.; Liu, T.; Zhang, K. A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. *IEEE Access* **2020**, *8*, 152183–152192. [CrossRef]
25. Pan, Y.; Wang, C.; Hu, B.; Xiang, Y.; Wang, X.; Chen, Q.; Chen, J.; Du, J. A BERT-Based Generation Model to Transform Medical Texts to SQL Queries for Electronic Medical Records: Model Development and Validation. *JMIR Med. Inform.* **2021**, *9*, e32698. [CrossRef]
26. Liu, K.; Chen, L. Medical Social Media Text Classification Integrating Consumer Health Terminology. *IEEE Access* **2019**, *7*, 78185–78193. [CrossRef]
27. Weng, W.H.; Waghlikar, K.B.; McCray, A.T.; Szolovits, P.; Chueh, H.C. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 155. [CrossRef]
28. Hsu, J.-L.; Hsu, T.-J.; Hsieh, C.-H.; Singaravelan, A. Applying Convolutional Neural Networks to Predict the ICD-9 Codes of Medical Records. *Sensors* **2020**, *20*, 7116. [CrossRef]
29. Moen, H.; Hakala, K.; Peltonen, L.M.; Suhonen, H.; Ginter, F.; Salakoski, T.; Salanterä, S. Supporting the use of standardized nursing terminologies with automatic subject heading prediction: A comparison of sentence-level text classification methods. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 81–88. [CrossRef]
30. Chintalapudi, N.; Battineni, G.; Canio, M.D.; Sagaró, G.G.; Amenta, F. Text mining with sentiment analysis on seafarers' medical documents. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100005. ISSN 2667-0968. [CrossRef]

31. Al-Doulat, A.; Obaidat, I.; Lee, M. Unstructured Medical Text Classification using Linguistic Analysis: A Supervised Deep Learning Approach. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019; pp. 1–7. [CrossRef]
32. Audebert, N.; Herold, C.; Slimani, K.; Vidal, C. Multimodal Deep Networks for Text and Image-Based Document Classification. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Würzburg, Germany, 16–20 September 2020. [CrossRef]
33. Choi, E.; Bahadori, M.T.; Searles, E.; Coffey, C.; Thompson, M.; Bost, J.; Tejedor-Sojo, J.; Sun, J. Multi-layer Representation Learning for Medical Concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1495–1504. [CrossRef]
34. Pappas, N.; Popescu-Belis, A. Multilingual hierarchical attention networks for document classification. *arXiv* **2017**, arXiv:1707.00896.
35. Zhang, Q.; Yuan, Q.; Lv, P.; Zhang, M.; Lv, L. Research on Medical Text Classification Based on Improved Capsule Network. *Electronics* **2022**, *11*, 2229. [CrossRef]
36. Yasunaga, I.; Leskovec, J.; Liang, P. LinkBERT: Pretraining Language Models with Document Links. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 8003–8016.
37. Zhang, D.; Mishra, S.; Brynjolfsson, E.; Etchemendy, J.; Ganguli, D.; Grosz, B.; Lyons, T.; Manyika, J.; Niebles, J.C.; Sellitto, M.; et al. "The AI Index 2022 Annual Report," *AI Index Steering Committee*; Stanford Institute for Human-Centered AI, Stanford University: Stanford, CA, USA, 2022.
38. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning (PMLR), Beijing, China, 22–24 June 2014; pp. 1188–1196.
39. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv* **2016**, arXiv:1607.01759.
40. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf> (accessed on 10 October 2022).
41. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arxiv:1907.11692.
42. Abreu, J.; Fred, L.; Macêdo, D.; Zanchettin, C. Hierarchical Attentional Hybrid Neural Networks for Document Classification. In Proceedings of the International Conference on Artificial Neural Networks, Munich, Germany, 17–19 September 2019. [CrossRef]
43. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* **2019**, arxiv:1906.08237.
44. Fries, J.A.; Weber, L.; Seelam, N.; Altay, G.; Datta, D.; Garda, S.; Kang, M.; Su, R.; Kusa, W.; Cahyawijaya, S.; et al. BigBio: A Framework for Data-Centric Biomedical Natural Language Processing. *arXiv* **2022**, arXiv:2206.15076.
45. Zunic, A.; Corcoran, P. Spasic ISentiment Analysis in Health and Well-Being: Systematic Review. *JMIR Med. Inform.* **2020**, *8*, e16023. [CrossRef] [PubMed]
46. Aattouchi, I.; Elmendili, S.; Elmendili, F. Sentiment Analysis of Health Care: Review. *E3s Web Conf.* **2021**, *319*, 01064. [CrossRef]
47. Tai, K.S.; Socher, R.; Manning, C.D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *arXiv* **2015**, arxiv:1503.00075.
48. Nii, M.; Tsuchida, Y.; Kato, Y.; Uchinuno, A.; Sakashita, R. Nursing-care text classification using word vector representation and convolutional neural networks. In Proceedings of the 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), Otsu, Japan, 27–30 June 2017; pp. 1–5.
49. Qian, Y.; Woodland, P.C. Very Deep Convolutional Neural Networks for Robust Speech Recognition. *arXiv* **2016**, arXiv:1607.01759.
50. Zhang, Y.; Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
51. Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11. [CrossRef]
52. Bosc, T.; Vincent, P. Auto-Encoding Dictionary Definitions into Consistent Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1522–1532. [CrossRef]
53. Spearman, C. 'General Intelligence,' Objectively Determined and Measured. *Am. J. Psychol.* **1904**, *15*, 201–292. [CrossRef]
54. Zhan, X.; Wang, F.; Gevaert, O. Reliably Filter Drug-Induced Liver Injury Literature With Natural Language Processing and Conformal Prediction. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5033–5041. [CrossRef]
55. Rathee, S.; MacMahon, M.; Liu, A.; Katritsis, N.; Youssef, G.; Hwang, W.; Wollman, L.; Han, N. DILL: An AI-based classifier to search for Drug-Induced Liver Injury literature. *bioRxiv* **2022**. [CrossRef]
56. Oh, J.H.; Tannenbaum, A.R.; Deasy, J.O. Automatic identification of drug-induced liver injury literature using natural language processing and machine learning methods. *bioRxiv* **2022**. [CrossRef]
57. Chen, Y.; Zhang, X.; Li, T. Medical Records Classification Model Based on Text-Image Dual-Mode Fusion. In Proceedings of the 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 28–31 May 2021; pp. 432–436. [CrossRef]

-
58. Jamaluddin, M.; Wibawa, A.D. Patient Diagnosis Classification based on Electronic Medical Record using Text Mining and Support Vector Machine. In Proceedings of the 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 18–19 September 2021; pp. 243–248. [[CrossRef](#)]
 59. Yang, F.; Wang, X.; Ma, H.; Li, J. Transformers-sklearn: A toolkit for medical language understanding with transformer-based models. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 90. [[CrossRef](#)]