

## Article

# A Comparative Analysis of Active Learning for Biomedical Text Mining

Usman Naseem <sup>1</sup>, Matloob Khushi <sup>1,\*</sup>, Shah Khalid Khan <sup>2</sup>, Kamran Shaukat <sup>3</sup> and Mohammad Ali Moni <sup>4</sup>

<sup>1</sup> School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia; usman.naseem@sydney.edu.au

<sup>2</sup> School of Engineering, RMIT University, Carlton, VIC 3053, Australia; s3680269@student.rmit.edu.au

<sup>3</sup> School of Electrical Engineering and Computing, The University of Newcastle, Newcastle, NSW 2308, Australia; kamran.shaukat@uon.edu.au

<sup>4</sup> UNSW Digital Health, WHO Center for eHealth, Faculty of Medicine, The University of New South Wales, Sydney, NSW 2052, Australia; m.moni@unsw.edu.au

\* Correspondence: matloob.khushi@sydney.edu.au

**Abstract:** An enormous amount of clinical free-text information, such as pathology reports, progress reports, clinical notes and discharge summaries have been collected at hospitals and medical care clinics. These data provide an opportunity of developing many useful machine learning applications if the data could be transferred into a learn-able structure with appropriate labels for supervised learning. The annotation of this data has to be performed by qualified clinical experts, hence, limiting the use of this data due to the high cost of annotation. An underutilised technique of machine learning that can label new data called active learning (AL) is a promising candidate to address the high cost of the label the data. AL has been successfully applied to labelling speech recognition and text classification, however, there is a lack of literature investigating its use for clinical purposes. We performed a comparative investigation of various AL techniques using ML and deep learning (DL)-based strategies on three unique biomedical datasets. We investigated random sampling (RS), least confidence (LC), informative diversity and density (IDD), margin and maximum representativeness-diversity (MRD) AL query strategies. Our experiments show that AL has the potential to significantly reducing the cost of manual labelling. Furthermore, pre-labelling performed using AL expedites the labelling process by reducing the time required for labelling.

**Keywords:** active learning; machine learning; biomedical natural language processing



**Citation:** Naseem, U.; Khushi, M.; Khan, S.K.; Shaukat, K.; Moni, M.A. A Comparative Analysis of Active Learning for Biomedical Text Mining. *Appl. Syst. Innov.* **2021**, *4*, 23. <https://doi.org/10.3390/asi4010023>

Academic Editor: Teen-Hang Meen

Received: 31 December 2020

Accepted: 8 March 2021

Published: 15 March 2021

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



**Copyright:** © 2021 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 wide-spread utilisation of capacity and digitising advancements, specifically the digitisation of clinical records, presents numerous information examination chances. Notwithstanding, to arrive at their maximum capacity, such investigation frameworks need to remove organised information from unstructured content reports. An expanding volume of unstructured clinical information about patients is put away electronically by clinics and medical services. Organised data is fundamental for applications, for example, reporting, reasoning, and retrieving, for instance, malignancy observations from medical reports and death certificates [1], checking radiology reports to forestall missed fractures [2], and clinical data retrieval [3]. Late advancements of Natural Language Processing (NLP) and information extraction (IE) have confronted fundamental difficulties in adequately catching valuable data from this free-text resources [4]. IE is a nontrivial interaction for extricating helpful, organised data like examples and different connections from unstructured info text.

One of the challenges is distinguishing cases of ideas that are alluded to in manners not captured inside current lexical assets and tackle uncertainty, polysemy, synonymy, and word order varieties. Moreover, the data introduced in clinical narratives are frequently unstructured, ungrammatical, and divided. Along these lines, standard NLP advances and frameworks cannot be straightforwardly applied to the clinical domain [5].

ML-based algorithms, rule-based and existing dictionary-based methods can be utilised to identify and extract the concepts from raw text corpus in finance, medical, and various other domains [6–10]. In the clinical domain, the ShARe/CLEF 2013 eHealth Evaluation Lab and the i2b2/VA challenge methodologies have been applied in shared tasks [11–13]. The results demonstrated that ML-based algorithms are scalable and usually beats the rule-based approaches.

A critical challenge is a clinical text contains domain expert words which requires domain expert efforts to presented rule-based methods or label huge corpora as training data for supervised ML-based methods. Usually, rule-based approaches are expensive because it needs domain experts and is a challenging task itself that can create error [14] and not adaptable or transferable to other tasks. The results of the supervised ML-based approached increases as the set of labelled data is used for training. Using crowdsourcing for labelling clinical data is not useful in the general domain; manual labelling is an expensive and labour-intensive task.

AL [15] and semi-supervised learning [16] are viable options in contrast to standard supervised ML methods and can reduce labelling costs. AL can prepare to accomplish an automated system with high adequacy and less labelling cost. Training an ML-based approach using a small subset of labelled data, selected randomly, leads to reduced effectiveness compared to when the model uses complete labelled data, while in AL, the aim is to reach high viability and effectiveness by training a small chunk of data.

AL is a human-in-the-loop technique with the capacity to radically decrease human inclusion contrasted with the conventional supervised ML techniques that require a massive amount of labelled data at the start. Figure 1 presents the overall general cycle of AL for extracting information from text. It is an iterative cycle, where informative samples from raw and unstructured text documents are chosen utilising a query strategy. A human annotator then labels these samples to extricate data and construct a supervised ML-based model at every iteration. The viability of AL techniques has been shown and decisively demonstrated in numerous spaces, for example, text classification, IE, and speech recognition [15].

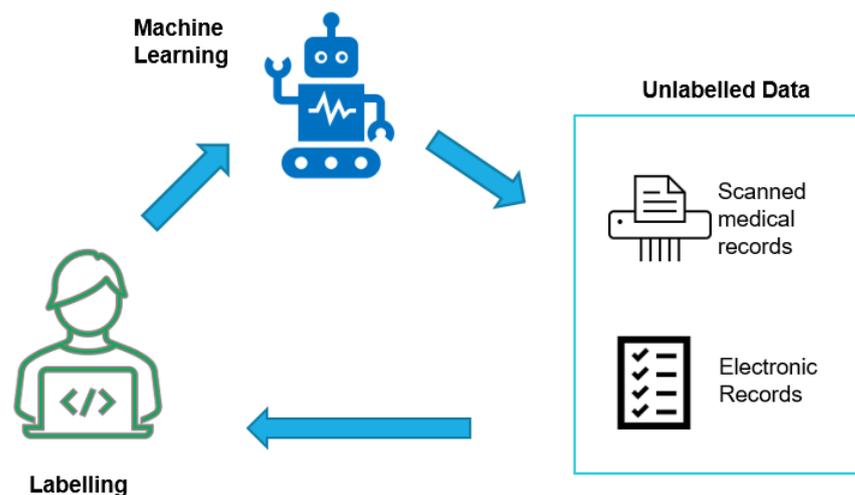


Figure 1. An overview of supervised machine learning using in active learning (AL).

Regardless of comparative findings on various tasks and domains, AL is not thoroughly investigated in biomedical tasks. Our research is based on the following research questions. RQ1: How AL can be used to reduce the labelling cost while maintaining the good quality of extracted information? RQ2: Which existing AL techniques perform well compared to other AL methods to reduce the labelling time?

RQ3: How can other ML approaches (i.e., representation learning and unsupervised learning) can produce effective information extraction while maintaining the quality and minimising the labelling effort?

Despite similar findings, the aim of our research is to provide a framework to the research community for extracting information from large amounts of unstructured biomedical documents by developing an AI-based framework that extracts high-quality concepts and reduces the burden of manual annotation.

## 2. Related Work

Expanding volumes of clinical information that can be presently digitised and put away in electronic medical records makes the extraction of information from clinical text progressively basic, especially in the region of NLP and ML. While numerous clinical assets and advances are presently accessible to encourage the preparing of clinical information, clinical data extraction remains challenging.

Recent studies focus on IE from biomedical literature, for example, books and scientific articles [17–20] and the subsequent gathering centres around IE from free-text clinical narratives delivered by clinical staff, for example, radiology and pathology reports or release synopses. Besides, other studies represent a more troublesome errand on account of the unstructured idea of the free content and the simple language used [6]. IE is a significant essential advance in extricating essential data from clinical records. The fundamental challenge is to create cost-productive methodologies that help automatic idea extraction from clinical free-text assets while guaranteeing the extracted ideas' high quality. Automatic handling of such volumes of information could incredibly profit clinical information systems.

### 2.1. Information Extraction from Biomedical Corpus

Extracting information from biomedical documents involves capturing words of natural language from raw and unstructured document which express the significant information within a given domain [14]. NLP-based techniques cannot be directly used to extraction information from biomedical corpus due to its ubiquitous, raw, and unstructured nature. Current methods can be divided into following techniques.

#### 2.1.1. Dictionary-Based Methods

Dictionary-based methods consist of matching a provided list of terms in a text and use patterns to extract structures like entities and text strings from a pre-defined dictionary. A large number of domain specific dictionaries are largely available which can be used to extract biomedical information. These include SNOMED CT [21] and UMLS [22]. Bashyam et al. [23] presented and demonstrated a lexical lookup approach for radiology reports to detect UMLS concepts. They showed that their method is 7 times faster than MetaMap in identifying the same concepts.

Dictionary-based techniques can be helpful in extracting information from free text with the help of dictionaries, and they can also normalise entities and be useful for both the syntactical and semantic level of information by associating the entities with terms in the dictionaries. These dictionary-based methods are useful but suffer from coverage issues, which makes their use limited in this domain.

#### 2.1.2. Rule-Based Methods

Rule-based methods have been generally evolved to extract entities in the biomedical domain [24]. Rule-based methods contain manually created rules to extract biomedical information from the corpus. Various techniques are used to define these rules, which are used to capture patterns within natural language [25].

Current databases have coverage issues and do not cover recently discovered elements; some helpful objective substances and biomedical-related data covered up in non-important settings may be missed and not extracted utilising a word reference-based methodology. Hamon and Graber [26] presented a rule-based method to extract biomedical information using existing terms, rules, and shallow parsing methods. Mack et al. [27] proposed BioTeks, a rule-based approach to capture biomedical information from biomedical corpus. These

methods are widely used in the biomedical domain; however, implementation requires domain expertise and they are not adaptable nor transferable to other domains [28].

## 2.2. Machine Learning (ML)

Machine learning (ML)-based methods are presented to address the shortcomings of the abovementioned techniques by making the machine learn and improve the performance [29]. Biomedical/clinical extraction can be classified as a labelling task sequence, referred to as a classification task in supervised learning of ML algorithms. Both support vector machines (SVMs) [30] and conditional random fields (CRFs) [31] are the methods mostly used in the classification for sequence labelling tasks. For another high level of sophisticated tasks, a large number and high quality of training data are required to train the models. Although a huge amount of data is available, the labelling cost is high and the task cumbersome. The AL technique is proposed to limit the required high volume of manual labelling of data. AL's main idea is to query and label those samples that carry useful information for the learning model compared to other available samples. It can attain better performance with less-annotated training data [15,32].

Semi-supervised learning is another approach to address annotated corpus [33]. It has been effectively applied to some real-world applications. An abundance of unlabelled examples is effortlessly accessible, while physically naming them is an escalated and costly errand. Self-training is a customarily utilised technique where unannotated text is annotated in an iterative interaction. The updated labelled set is utilised to retrain and refresh the fundamental classifier at every emphasis. This examination researches how to increase the learning model at every point by consolidating self-preparing into the AL process.

Representation learning refers to learning data representations to facilitate information extraction. The IE is fed into the training of the machine learning model [34]. Mikolov et al. [35] introduced a novel word embedding concept where words are represented in continuous vector representations of words based on their various dimensions of difference.

## 2.3. Natural Language Processing (NLP)

NLP is the intersection of computing science and linguistics that includes dissecting and understanding common human language from both speech and written texts. Over the years, NLP has been used in various applications such as email filtering [36], irony and sarcasm detection [37] document organisation [38], sentiment and opinion mining prediction [39–41], hate speech detection [42–44], question answering [45], content mining [46], biomedical text mining [47,48], and many more [8,49,50].

In biomedical named entity recognition (BioNER), Yao et al. [51] initially created embeddings of words on unlabelled texts of biological topics using neural networks, going on to establish a multi-layer neural network to obtain cutting edge output. Li et al. [52] mixed sentence vectors and twin word embeddings and utilised the BiLSTM on biomedical texts to identify domain-relevant entities. To identify drug entities, Zeng et al. [53] developed their model, BiLSTM-CRF. CNN was utilised to get the representation of features on a character level. This was done with representations on a word level and used as data to be fed to the BiLSTM-CRF for BioNER. In biomedical literature, many words can cause information redundancy whilst neural network models are trained for feature capture, preventing critical information from being obtained. This may cause the crucial areas not to focus on the BioNER models, and loss of information could occur. It is a salient focus to make models of neural networks attentive to areas of importance. In machine translation, Bahdanau et al. [54] suggest the attention focusing mechanism. Taking the decoder model into account, the focus can be made on the initial text's essential bits as it is decoded, reducing information loss. An attention-based BiLSTM-CRF model is used by Luo et al. [55] for BioNER on a document level. They optimise the tagging inconsistency

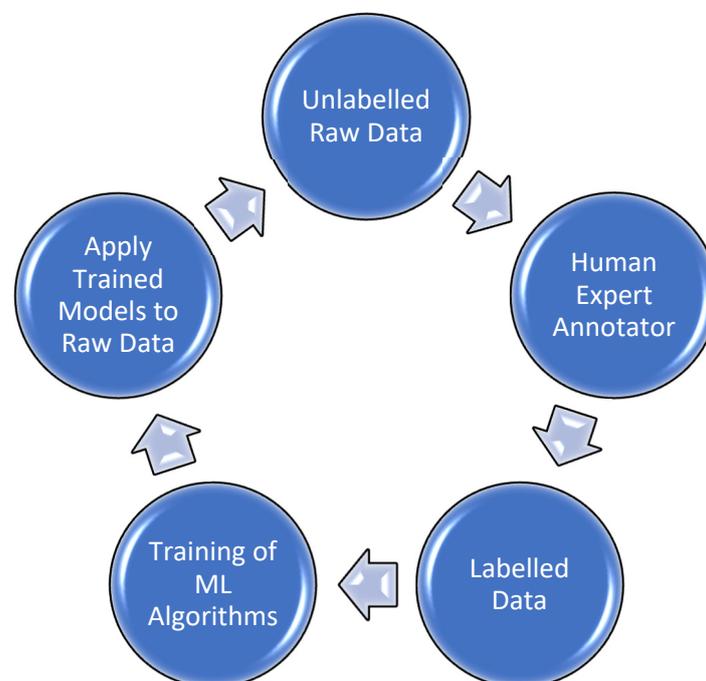
problem by using, between various sentences, mechanisms that are attention-focused. The best results are obtained on CHEMDNER and CDR corpora using this approach.

Several other works have investigated the benefit of contextual models in biomedical and clinical areas. Several researchers trained ELMo on biomedical corpora and presented BioELMo and found that BioELMo beats ELMo on BioNER tasks [56,57]. Along with their work, a pre-trained BioELMo model was published, enabling further clinical research. Beltagy et al. [58] released Scientific BERT (SciBERT), where BERT was trained on the scientific texts. In non-contextual embedding, BERT has been usually superior and better than ELMo. Similarly, innovative wireless connectivity techniques could be applied to the remote execution of these techniques [59–62]

Si et al. [63], trained the BERT on clinical notes corpora, using complex task-specific models to improve both traditional embedding and ELMo embedding i2b2 2010 and 2012 BioNER. Similarly, in another study, a new domain-specific language model, BioBERT [64], trained a BERT model on biomedical documents from PMC abstracts and articles from PubMed that resulted in improved BioNER results. Peng et al. [65] introduced Biomedical Language Understanding Evaluation (BLUE), a collection of resources for evaluating and analysing natural biomedical language representation models.

#### 2.4. Active Learning (AL)

AL algorithms are beneficial in ML, especially when we have large amounts of unannotated data. AL techniques use supervised ML methods in an iterative way. A human annotator is involved in the learning process and can drastically decrease the human involvement as demonstrated in Figure 2. Despite its strength, AL has not been fully explored for biomedical information extraction. AL's primary goal is to maximise the model's effectiveness by reducing the number of samples that need manual labelling. The main challenge is to find informative samples that are available to train a model, achieving the better performance and high effectiveness.



**Figure 2.** Schematic diagram of AL as an iterative process which help labelling the raw data.

#### 2.5. Active Learning in Clinical Domain

AL aims to reduce the costs and issues related to the manual annotation step in supervised ML methods. Decreasing the manual annotation burden becomes highly critical in the clinical domain because of qualified experts' high costs to annotate the clinical

free text. AL is used for various biomedical tasks [66], de-identifying clinical records [67], clinical text classification [68], and clinical named entity recognition [69]. Random sampling (RS), where samples are chosen randomly, is a commonly used AL technique.

Rosales et al. [70] presented an AL method to identify biomedical information to two groups. Their method outperformed the traditional methods. Chen et al. [66] presented a sampling technique established on the changes appearing in different learning models during AL. Another study on de-identifying of Swedish biomedical samples as a classification task was presented by Boström and Dalianis [67]. They presented the comparison on the performance of two AL methods against RS baseline methods. Recently, Chen et al. [69] proposed new AL query strategies that belong to uncertainty-based approaches and diversity-based approaches. Authors presented a comprehensive evaluation of current and new AL methods on biomedical tasks and found that uncertainty sample-based methods resulted in less effort being required to label the corpus as compared to diversity-based methods.

Considering the basic need of having cost-effective AL approaches for biomedical tasks, the highlighted limitations need to be addressed. Therefore, in this research, our aim is to address the cost needed for manual annotation using AL and representation learning.

### 3. Methodology

#### 3.1. Dataset

In this study, we used the following datasets.

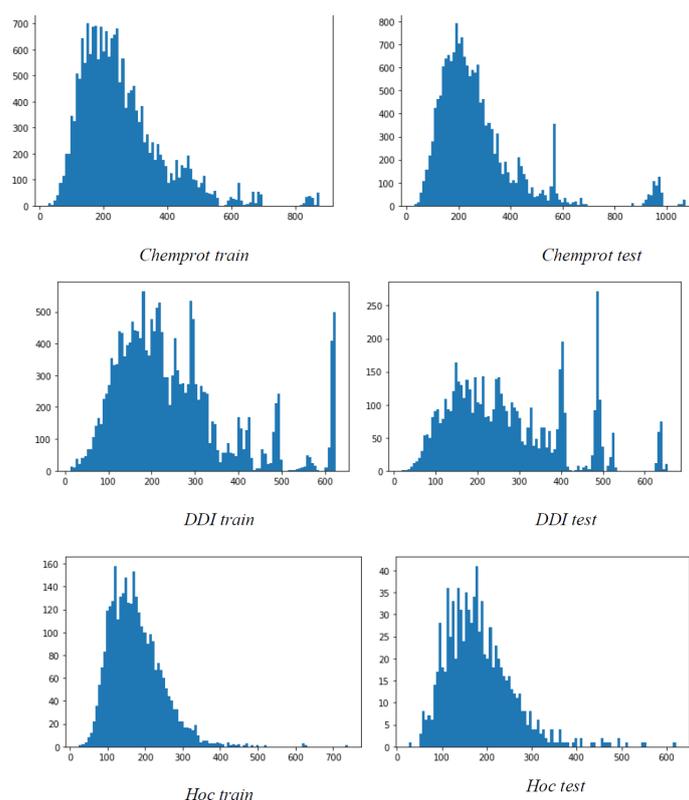
DDI extraction 2013 corpus is a collection of 792 texts selected from the DrugBank database and other 233 Medline abstracts [71]. The drug-drug interactions, including both pharmacokinetics and pharmacodynamic interactions, were annotated by two expert pharmacists with a substantial pharmacovigilance background. In our benchmark, we use 624 train files and 191 test files to evaluate the performance and report the micro-average F1-score of the four DDI types.

ChemProt consists of 1820 PubMed abstracts with chemical-protein interactions annotated by domain experts and was used in the BioCreative VI text mining chemical-protein interactions shared task [72]. We use the standard training, and test sets in the ChemProt shared task and evaluate the same five classes: CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9.

HoC (the Hallmarks of Cancers corpus) consists of 1580 PubMed abstracts annotated with ten currently known hallmarks of cancer [73]. Annotation was performed at the sentence level by an expert with 15+ years of experience in cancer research. We used 315 (20%) abstracts for testing and the remaining abstracts for training. Table 1 shows the name along with the task description of the dataset used in this study. Further, Figure 3 depicts the data analysis of the dataset used in our study.

**Table 1.** Dataset used.

Dataset	Task
DDI	Relation Extraction
ChemProt	Relation Extraction
HoC	Document Classification



**Figure 3.** Data analysis of the datasets used in this study.

### 3.2. Active Learning Query Strategies

#### 3.2.1. Random Sampling (RS)

The key idea for random sampling of AL is that it takes a small, random portion of the entire dataset to represent the entire dataset. Each member has an equal probability. During the AL application, random sampling is quite the most straightforward algorithm compared to other query strategies. It applies the random state and shuffles to achieve the random selection of the training and testing pools.

#### 3.2.2. Least Confidence (LC)

Least confidence is one of the methods belonging to uncertainty sampling, a query strategy that tries to determine the word's values by calculating the real uncertainty of the word.

#### 3.2.3. Informative Diversity and Density (IDD)

IDD is a method used to calculate the information density of an instance  $x$ . Unlike uncertainty sampling, IDD can lead us to take the structure of data into account.

#### 3.2.4. Margin

Margin is also belonging to uncertainty sampling; unlike LC, the margin is designed to measure the difference in probability of the first and second most likely prediction.

#### 3.2.5. Maximum Representativeness-Diversity (MRD)

Maximum representativeness diversity is a method that relies only on the similarity between samples and all other samples in unlabelled sets. The most representative is to mark various samples in the current batch and then add them to the training set. This method could prevent experts from waiting until the learning model is on the current set of tags, and then the next batch of samples selects tags using one of the above query strategies.

### 3.3. AL Query Strategies

There are many query strategies in AL; however, not all query strategies are invented for all situations. We pick up LC and margin because they are the most popular query strategies in other areas. We pick up RS because it is different from other algorithms as it picks up pools randomly. Then we choose IDD because IDD uses a different measure way compared with LC and margin. For the same reason, we pick up MRD to increase the variety of our query strategies schemes to get a better and reliable result for analysing.

### 3.4. Feature Extraction Methods

For feature extraction methods, we pick TF-IDF for feature extraction method in many areas. Then, we add FastText to compare with TF-IDF because TF-IDF only considers the frequency of a word in a document. FastText, consider more than that which can give our study result analysis some other aspect to analyse the performance. In the end, we decided to add BERT and ELMo and their extension into our study. Because BERT and ELMo are heavy methods compared to others and perform well, especially with NLP tasks in other areas than other methods. Therefore, we decided also to include this to analyse its performance with clinical datasets.

### 3.5. Machine Learning Methods

For ML methods, first of all, we determined to choose some the widely used method as the basic ML methods for our study is why we pick SVM, KNN, and NB; they are widely used methods in many different aspects of the dataset. Then, to make some comparison with SVM, KNN, and NB, we pick up some algorithms with different schemes compared to SVM, KNN, and NB. XGBoost and CatBoost are both gradient boosts based on decision trees. Random forest (RF) and AdaBoost are both ensemble functions. Furthermore, each of them is the most popular method in their area. Therefore, we finally pick these 7 methods as our ML methods to make our results more reliable by analysing different schemes' performance.

## 4. Results and Discussion

The results (Tables 2–13) show that the DDI dataset, which applies BERT for feature extraction, has the best performance in accuracy when we apply an SVM algorithm with an AL framework which builds based on MRD query strategies.

**Table 2.** Comparison of results with and without AL methods (TF-IDF) for DDI dataset.

TF-IDF	DDI	Supervised Learning	RS	LC	IDD	Margin	MRD
SVM	Accuracy	82.17	82.92	81.88	81.44	81.79	82.49
	F1	87.08	89.75	86.9	87	86.71	88.73
NB	Accuracy	81.08	82.99	83.04	82.61	83.01	81.84
	F1	86.34	90.69	90.7	89.25	90.71	88.78
KNN	Accuracy	64.92	74.99	68.84	67.45	69.87	70.6
	F1	61.42	75.71	66.78	64.91	68.72	69.63
XGBoost	Accuracy	82.87	82.62	82.85	79.93	83.06	82.07
	F1	89.82	88.56	88.63	84.58	89.03	88.05
Random forest	Accuracy	81.46	81.03	82.87	81.6	81.27	81.44
	F1	84.94	85.65	87.45	86.77	85.86	86.73
AdaBoost	Accuracy	78.11	83.01	82.83	80.56	82.9	82.38
	F1	81.31	82.1	90.31	86.32	90.53	89.78
CatBoost	Accuracy	81.18	90.71	91.43	89.93	90.5	89
	F1	86.01	87.4	90.8	89	90.1	90.21

**Table 3.** Comparison of results with and without AL methods (TF-IDF) for ChemProt dataset.

TF-IDF	ChemProt	without	RS	LC	IDD	Margin	MRD
SVM	Accuracy	77.40	79.30	81.88	81.44	81.79	79.12
	F1	80.83	86.21	86.9	87	86.71	85.45
NB	Accuracy	79.57	79.59	83.04	82.16	83.01	79.61
	F1	87.7	88.64	90.7	89.25	90.71	88.56
KNN	Accuracy	60.68	65.68	68.84	67.45	69.87	58.57
	F1	57.31	64.42	66.78	64.91	68.72	54.43
XGBoost	Accuracy	78.92	78.46	78.99	78.81	79.11	78.91
	F1	84.36	83.38	83.82	84.26	83.86	84.64
Random forest	Accuracy	78.85	78.32	78.55	78.5	78.6	78.58
	F1	83.81	84.28	83.04	84.42	83.49	84.39
AdaBoost	Accuracy	76.46	79.38	77.69	77.83	75.51	78.46
	F1	82.77	86.63	82.79	85.41	81.22	86.62
CatBoost	Accuracy	78.92	78.81	80.5	84.89	83.5	82.10
	F1	84.36	83.63	82.80	82.00	83.10	82.90

**Table 4.** Comparison of results with and without AL methods (TF-IDF) for HoC dataset.

TF-IDF	HoC	without	RS	LC	IDD	Margin	MRD
SVM	Accuracy	93.64	91.39	93.20	93.12	93.16	92.93
	F1	91.26	90.39	91.08	90.83	91.08	90.74
KNN	Accuracy	86.50	88.12	86.05	86.35	86.05	89.08
	F1	93.51	91.88	93.36	93.42	93.36	93.28
Random Forest	Accuracy	93.51	90.35	92.99	92.79	92.99	93.12
	F1	81.82	86.81	81.63	81.39	81.63	81.07
CatBoost	Accuracy	94.32	92.09	94.90	94.10	93.40	92.10
	F1	84.36	83.38	83.82	84.26	83.86	84.64
Random forest	Accuracy	78.85	78.32	78.55	78.50	78.60	78.58
	F1	83.81	84.28	83.04	84.42	83.49	84.39
AdaBoost	Accuracy	76.46	79.38	77.69	77.83	75.51	78.46
	F1	82.77	86.63	82.79	85.41	81.22	86.62
CatBoost	Accuracy	78.92	78.81	79.80	78.80	79.10	79.20
	F1	84.36	83.63	85.80	84.90	84.20	84.70

**Table 5.** Comparison of results with and without AL methods (FastText) for DDI dataset.

FastText	DDI	without	RS	LC	IDD	Margin	MRD
SVM	Accuracy	83.13	82.59	81.79	81.51	81.54	72.97
	F1	90.28	88.21	85.22	86.12	85.24	71.44
NB	Accuracy	83.01	82.93	83.15	83.01	82.99	83.02
	F1	90.71	90.48	90.13	90.71	90.21	90.27
KNN	Accuracy	76.46	73.75	73.35	78.31	73.35	73.12
	F1	77.53	74.81	73.59	81.54	73.59	73.09
XGBoost	Accuracy	83.41	82.72	83.54	82.55	83.47	83.35
	F1	89.99	89.67	89.83	89.11	90.30	89.54
Random forest	Accuracy	77.45	81.53	81.21	77.69	79.97	80.66
	F1	79.60	86.63	85.91	80.34	83.89	84.96

**Table 5.** *Cont.*

<b>FastText</b>	<b>DDI</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
AdaBoost	Accuracy	70.58	66.08	82.47	90.61	78.07	78.84
	F1	70.25	63.24	89.71	82.94	82.09	83.43
CatBoost	Accuracy	81.17	81.27	82.48	81.46	82.17	82.57
	F1	85.24	86.40	88.36	87.35	87.95	88.69

**Table 6.** Comparison of results with and without AL methods (FastText) for HoC dataset.

<b>FastText</b>	<b>HoC</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	F1	43.01	57.97	33.25	34.34	36.48	40.52
	F1	20.86	36.45	31.06	29.41	41.37	27.19
KNN	F1	43.43	40.04	36.89	39.72	39.09	32.64
	F1	41.04	45.61	41.75	40.09	49.05	27.77
Random forest	F1	23.47	28.08	22.87	26.68	23.95	22.04
	F1	36.92	32.71	37.75	40.23	45.01	30.39
CatBoost	F1	35.82	32.70	35.47	40.27	37.23	28.30
	F1	89.99	89.67	89.83	89.11	90.30	89.54
Random forest	Accuracy	77.45	81.53	81.21	77.69	79.97	80.66
	F1	79.60	86.63	85.91	80.34	83.89	84.96
AdaBoost	Accuracy	70.58	66.08	82.47	90.61	78.07	78.84
	F1	70.25	63.24	89.71	82.94	82.09	83.43
CatBoost	Accuracy	81.17	81.27	82.48	81.46	82.17	82.57
	F1	85.24	86.40	88.36	87.35	87.95	88.69

**Table 7.** Comparison of results with and without AL methods (FastText) for ChemProt dataset.

<b>FastText</b>	<b>HoC</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	Accuracy	79.51	79.37	77.64	78.99	79.34	79.06
	F1	88.35	87.82	84.07	86.85	87.97	85.73
NB	Accuracy	79.58	79.55	79.46	79.58	79.53	79.54
	F1	88.61	88.56	88.27	88.63	88.44	88.45
KNN	Accuracy	74.64	75.86	75.52	74.92	69.05	75.85
	F1	78.44	81.21	80.94	79.87	70.67	81.11
XGBoost	Accuracy	79.61	79.57	79.62	79.65	79.51	79.63
	F1	88.29	88.41	88.37	88.47	88.43	88.54
Random forest	Accuracy	67.89	68.41	76.20	74.21	76.39	77.73
	F1	68.68	69.73	82.43	79.06	82.49	84.66
AdaBoost	Accuracy	73.69	68.49	77.13	74.30	71.35	76.33
	F1	78.62	69.93	84.57	80.24	75.87	82.92
CatBoost	Accuracy	79.57	79.53	79.62	79.58	78.46	79.59
	F1	88.33	88.16	88.37	88.32	86.13	88.42

**Table 8.** Comparison of results with and without AL methods (BERT) for DDI dataset.

<b>BERT</b>	<b>HoC</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	Accuracy	82.24	83.37	82.95	81.64	79.78	83.72
	F1	86.59	89.57	85.58	85.53	81.94	89.48
NB	Accuracy	62.59	62.70	67.87	65.51	66.22	42.15
	F1	58.53	58.72	66.18	62.75	64.05	32.30
KNN	Accuracy	73.77	75.51	73.55	74.90	73.44	73.81
	F1	74.20	76.92	74.12	76.04	73.94	74.54
XGBoost	Accuracy	83.09	82.59	82.78	82.97	83.09	82.56
	F1	88.85	88.73	88.11	88.99	88.41	88.21
Random forest	Accuracy	75.89	78.93	79.43	78.25	79.90	79.43
	F1	77.80	82.30	83.64	81.87	84.37	82.78
AdaBoost	Accuracy	82.59	74.85	81.74	81.93	81.53	79.76
	F1	89.80	76.37	87.41	88.39	87.91	84.87
CatBoost	Accuracy	81.11	81.03	81.88	82.16	82.45	81.91
	F1	85.85	85.14	86.59	87.80	87.79	86.75

**Table 9.** Comparison of results with and without AL methods (BERT) for HoC dataset.

<b>BERT</b>	<b>HoC</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	F1	83.60	83.46	89.26	89.33	89.26	89.63
	F1	85.96	82.87	84.20	84.51	84.20	78.80
KNN	F1	82.81	82.22	81.96	81.56	81.96	81.86
	F1	86.80	85.24	86.40	85.94	86.40	86.29
Random forest	F1	83.69	83.43	83.69	82.40	83.69	84.17
	F1	94.65	86.64	95.67	91.50	95.67	91.87
CatBoost	F1	85.72	85.24	86.28	85.95	86.28	86.68
	F1	88.85	88.73	88.11	88.99	88.41	88.21
Random forest	Accuracy	75.89	78.93	79.43	78.25	79.90	79.43
	F1	77.80	82.30	83.64	81.87	84.37	82.78
AdaBoost	Accuracy	82.59	74.85	81.74	81.93	81.53	79.76
	F1	89.80	76.37	87.41	88.39	87.91	84.87
CatBoost	Accuracy	81.11	81.03	81.88	82.16	82.45	81.91
	F1	85.85	85.14	86.59	87.80	87.79	86.75

**Table 10.** Comparison of results with and without AL methods (BERT) for ChemProt dataset.

<b>BERT</b>	<b>ChemProt</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	Accuracy	79.59	79.58	78.68	79.20	79.44	79.50
	F1	88.63	88.63	84.80	87.00	87.41	88.01
NB	Accuracy	69.71	67.61	67.45	67.00	60.44	55.39
	F1	73.09	70.10	69.60	69.15	59.19	51.79
KNN	Accuracy	65.92	68.42	64.70	64.12	64.73	64.69
	F1	65.87	69.73	64.36	63.58	64.36	64.33
XGBoost	Accuracy	79.57	79.44	79.44	79.38	79.40	79.35
	F1	88.49	88.19	88.01	87.61	87.84	87.68
Random forest	Accuracy	76.62	76.11	76.73	76.18	77.01	76.26
	F1	82.88	82.31	83.21	82.18	83.62	82.27
AdaBoost	Accuracy	79.17	68.12	73.86	71.79	76.62	74.45
	F1	87.93	70.18	79.46	76.09	83.04	80.28
CatBoost	Accuracy	78.19	77.81	77.77	77.67	77.31	76.61
	F1	85.53	84.72	84.82	84.31	84.22	82.76

**Table 11.** Comparison of results with and without AL methods (ELMo) for DDI dataset.

<b>ELMo</b>	<b>DDI</b>	<b>without</b>	<b>RS</b>	<b>LC</b>	<b>IDD</b>	<b>Margin</b>	<b>MRD</b>
SVM	Accuracy	83.01	83.01	83.01	83.01	83.01	83.01
	F1	90.71	90.71	90.71	90.71	90.71	90.71
NB	Accuracy	38.21	44.61	50.56	55.06	56.22	56.29
	F1	28.45	36.23	48.32	47.97	50.28	49.46
KNN	Accuracy	78.43	79.60	79.33	80.56	79.12	80.78
	F1	79.11	82.54	81.26	84.26	82.77	84.38
XGBoost	Accuracy	82.38	83.13	82.17	82.79	83.16	82.87
	F1	88.23	90.20	88.65	90.62	90.66	89.94
Random forest	Accuracy	80.82	79.92	80.43	80.32	81.41	80.85
	F1	85.31	84.44	86.21	85.10	87.06	86.15
AdaBoost	Accuracy	77.56	78.01	79.76	78.15	80.99	80.16
	F1	81.32	82.40	83.53	82.45	86.53	85.32
CatBoost	Accuracy	80.45	82.42	81.43	82.97	83.02	82.90
	F1	87.34	88.71	89.54	90.35	90.42	89.34

**Table 12.** Comparison of results with and without AL methods (ELMo) for HoC dataset.

ELMo	HoC	without	RS	LC	IDD	Margin	MRD
SVM	F1	85.76	84.51	90.51	90.78	90.62	90.73
	F1	84.75	82.58	84.84	84.21	84.84	82.73
KNN	F1	84.06	82.95	81.86	82.19	83.91	81.64
	F1	88.31	87.10	88.08	87.94	88.12	87.95
Random forest	F1	78.45	78.56	79.94	74.27	79.94	79.92
	F1	98.10	90.95	91.49	94.81	93.94	93.09
CatBoost	F1	86.61	87.16	87.16	86.69	87.39	88.01
	F1	88.23	90.20	88.65	90.62	90.66	89.94
Random forest	Accuracy	80.82	79.92	80.43	80.32	81.41	80.85
	F1	85.31	84.44	86.21	85.10	87.06	86.15
AdaBoost	Accuracy	77.56	78.01	79.76	78.15	80.99	80.16
	F1	81.32	82.40	83.53	82.45	86.53	85.32
CatBoost	Accuracy	80.45	82.42	81.43	82.97	83.02	82.90
	F1	87.34	88.71	89.54	90.35	90.42	89.34

**Table 13.** Comparison of results with and without AL methods (ELMo) for ChemProt dataset.

ELMo	ChemProt	without	RS	LC	IDD	Margin	MRD
SVM	Accuracy	79.59	79.59	79.59	79.59	79.59	79.59
	F1	88.63	88.64	88.64	88.64	88.64	88.64
NB	Accuracy	37.77	38.59	48.91	50.18	50.46	45.55
	F1	28.23	29.20	41.98	42.98	43.42	37.91
KNN	Accuracy	62.52	64.89	63.57	63.76	62.34	67.18
	F1	60.38	63.78	62.08	62.64	60.08	67.35
XGBoost	Accuracy	79.83	79.72	79.96	79.65	79.81	79.75
	F1	88.00	87.56	87.26	87.47	87.27	87.45
Random forest	Accuracy	74.82	75.38	77.42	73.49	77.60	76.97
	F1	80.22	80.87	83.68	77.86	84.28	84.00
AdaBoost	Accuracy	72.88	68.11	77.02	75.58	76.72	75.50
	F1	76.40	69.51	83.92	81.45	82.68	76.88
CatBoost	Accuracy	76.78	76.23	78.45	78.34	78.79	76.94
	F1	82.01	81.92	84.61	84.92	85.17	83.05

The result shows a table DDI dataset, which applies BERT for feature extraction and has the best performance in accuracy when applying an SVM algorithm with an AL framework that builds based on MRD query strategies.

Almost all ML methods have good performance except KNN algorithms. Furthermore, in general, AL algorithms have slightly better performance than passive learning algorithms.

HoC datasets have a much clearer difference between different methods applied. In general, AL performs better than passive learning algorithms. For ML algorithms, we can see that XGBoost and CatBoost have relatively better performance than others. Furthermore, for query strategies, margins have overall better performance than others.

The following are answers to our research questions.

- CatBoost performs better than others in most situations after we summarise all results tables.
- In general, LC and margin have better performance than other query strategies after we summarise all result tables.
- Overall, AL performance is better; therefore, AL is more recommended than passive learning.

In addition to the above results, we also notice that CatBoost always performs stably in every situation where other classifiers somehow have some bad performance. The judgement of LC and margin performance is challenging since they still have similar performance in almost all cases.

For the first results, CatBoost, as described in the methodology, is part of the gradient boost based on DT. This structure gives CatBoost the ability to get more chances to recover the errors during the implementation of the entire CatBoost structure since the later tree will fix the error that occurred by the previous tree. At the same time, the CatBoost boosting scheme is modified to be more efficient than other gradient boost algorithms, such as XGBoost, which gives CatBoost stability when changing the hyperparameters, especially with extensive data. All these advantages make CatBoost, overall, have a better ability to perform better in our study. Besides other algorithms such as AdaBoost, SVM is very sensitive with the correlation between data and data, which lead them to make more mistakes during the training and prediction than other algorithms.

For the second result, both LC and margin belong to uncertainty sampling, which calculates the uncertainty between data to measure the value of the word to decide the query order. Therefore, we can consider these two methods as a similar scheme used for AL. Then, uncertainty sampling was invented to reduce classification errors, making them more able to reduce classification errors than other query strategies, which is also what our study aims for. At the same time, IDD and MRD focused more on one word to decide the values. This could be better than uncertainty sampling with efficiently pre-computed; however, we cannot develop an efficiently pre-computed IDD and MRD algorithm to test the performance due to the tight time for implementation. This lead IDD and MRD cannot perform better than LC and Margin.

For the third result, the main reason why we can achieve better performance since with fewer data trained by using AL than using passive learning is the unbalance of the dataset. More data does not mean more accuracy for text classification. There exist iterations for the classification, even with all valuable data. In this time, insufficient data can immediately cause errors and reduce the accuracy of the classification. Therefore, the most important thing for classification is not the number of training pools. The most important thing is, can you find out which data is valuable enough to train the classifier. The AL algorithm is invented to achieve this goal by applying different query strategies. Therefore, AL can perform better than passive learning. Graphical representation of results is shown in Figure 4.

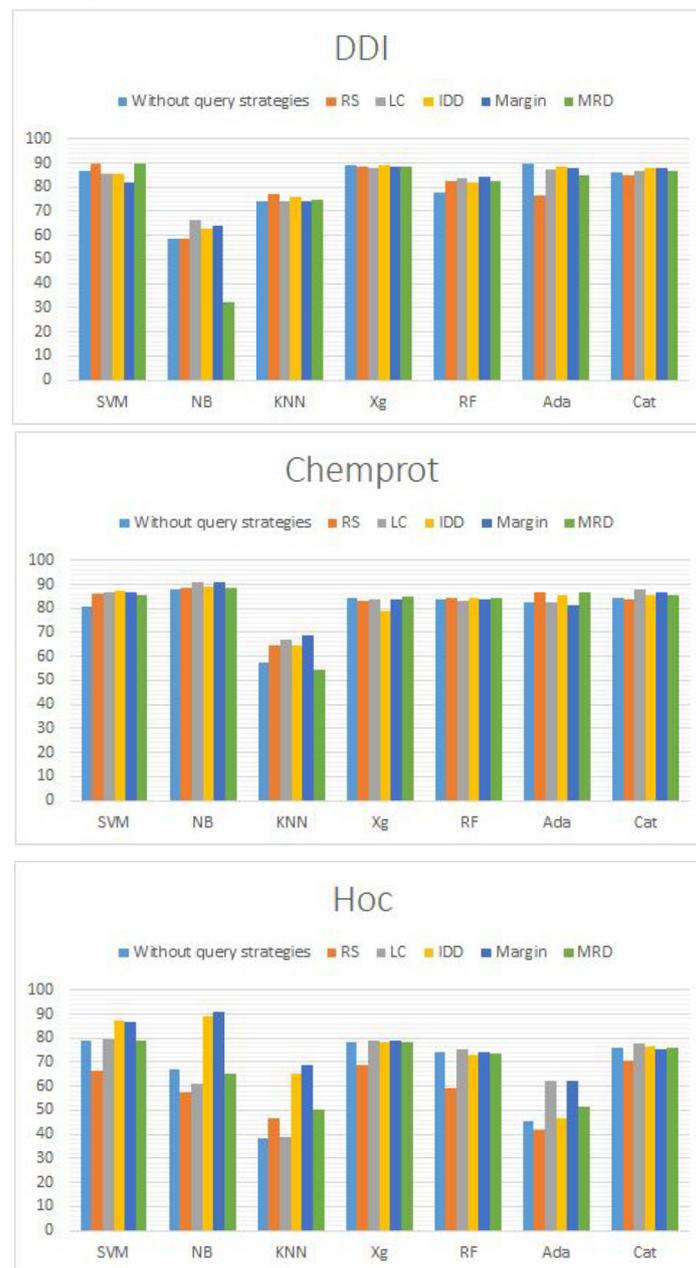


Figure 4. Graphical Representation of Results.

## 5. Conclusions

We conducted a simulated study to compare different AL algorithms for a clinical task. Our results showed that most AL algorithms outperformed the passive learning method when we assume equal annotation cost for each sentence. However, savings of annotation by AL were reduced when the length of sentences was considered. We suggest that the effectiveness of AL for clinical NER needs to be further evaluated by developing AL enabled annotation systems and conducting user studies.

We can conclude that AL is more recommended to test a clinical dataset classification with unlabelled data than passive learning. Compared to nowadays techniques to generate the health care outcomes, it will provide at least the same accuracy as before and even with less training dataset, which will significantly decrease the cost of collecting and labelling the dataset. Also, we can see that CatBoost makes a great performance combined with the uncertainty sampling AL framework. This also gives more options to choose when people want to implement the AL to text classification. Furthermore, the domain knowledge is not

so hard to understand since AL is still one part of ML; therefore, the required knowledge is only ML; once master this knowledge, the rest part is not hard to implement.

**Author Contributions:** Conceptualisation, U.N. and M.K.; methodology, U.N., M.K.; writing—original draft preparation, U.N.; writing—review and editing, K.S., S.K.K., M.A.M.; project administration, M.K.; 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:** The code and data are available from <https://github.com/usmann> (accessed on 14 March 2021).

**Acknowledgments:** Authors acknowledge and thank Junchi Li and Hengze Liu for their contribution to the project.

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

## References

1. Nguyen, A.N.; Moore, J.; O'Dwyer, J.; Philpot, S. Automated cancer registry notifications: Validation of a medical text analytics system for identifying patients with cancer from a state-wide pathology repository. *AMIA Annu. Symp. Proc.* **2016**, *2016*, 964. [PubMed]
2. Koopman, B.; Zuccon, G.; Waghlikar, A.; Chu, K.; O'Dwyer, J.; Nguyen, A.; Keijzers, G. Automated reconciliation of radiology reports and discharge summaries. *AMIA Annu. Symp. Proc.* **2015**, *2015*, 775.
3. Zuccon, G.; Koopman, B.; Nguyen, A.; Vickers, D.; Butt, L. Exploiting medical hierarchies for concept-based information retrieval. In Proceedings of the Seventeenth Australasian Document Computing Symposium, Dunedin, New Zealand, 5–6 December 2012; pp. 111–114.
4. Ohno-Machado, L.; Nadkarni, P.; Johnson, K. Natural language processing: Algorithms and tools to extract computable information from EHRs and from the biomedical literature. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 805. [CrossRef] [PubMed]
5. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 544–551. [CrossRef]
6. Meystre, S.M.; Savova, G.K.; Kipper-Schuler, K.C.; Hurdle, J.F. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb. Med. Inform.* **2008**, *17*, 128–144.
7. Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. *Appl. Syst. Innov.* **2021**, *4*, 9. [CrossRef]
8. Jaggi, M.; Mandal, P.; Narang, S.; Naseem, U.; Khushi, M. Text Mining of Stocktwits Data for Predicting Stock Prices. *Appl. Syst. Innov.* **2021**, *4*, 13. [CrossRef]
9. Singh, J.; Khushi, M. Feature Learning for Stock Price Prediction Shows a Significant Role of Analyst Rating. *Appl. Syst. Innov.* **2021**, *4*, 17.
10. Mukherjee, M.; Khushi, M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **2021**, *4*, 18.
11. Uzuner, Ö.; Goldstein, I.; Luo, Y.; Kohane, I. Identifying patient smoking status from medical discharge records. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 14–24. [CrossRef] [PubMed]
12. Suominen, H.; Salanterä, S.; Velupillai, S.; Chapman, W.W.; Savova, G.; Elhadad, N.; Pradhan, S.; South, B.R.; Mowery, D.L.; Jones, G.J.; et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*; Springer: Berlin, Germany, 2013; pp. 212–231.
13. Gurulingappa, H. Mining the Medical and Patent Literature to Support Healthcare and Pharmacovigilance. Ph.D. Thesis, Universitäts- und Landesbibliothek Bonn, Bonn, Germany, 2012.
14. Settles, B. Active Learning, volume 6 of Synthesis Lectures on Artificial Intelligence and Machine Learning. *Morgan Claypool* **2012**, 6.
15. Garla, V.; Taylor, C.; Brandt, C. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *J. Biomed. Inform.* **2013**, *46*, 869–875. [CrossRef] [PubMed]
16. Kholghi, M. Active Learning for Concept Extraction from Clinical Free Text. Ph.D. Thesis, Queensland University of Technology, Brisbane, Australia, 2017.
17. Leser, U.; Hakenberg, J. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings Bioinform.* **2005**, *6*, 357–369. [CrossRef] [PubMed]
18. Cho, H.; Lee, H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinform.* **2019**, *20*, 1–11. [CrossRef]

19. Kumar, P.; Gupta, A. Active learning query strategies for classification, regression, and clustering: A survey. *J. Comput. Sci. Technol.* **2020**, *35*, 913–945. [[CrossRef](#)]
20. Carvallo, A.; Parra, D.; Lobel, H.; Soto, A. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics* **2020**, *125*, 3047–3084. [[CrossRef](#)]
21. Cote, R.A.; Robboy, S. Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED). *JAMA* **1980**, *243*, 756–762. [[CrossRef](#)] [[PubMed](#)]
22. Lindberg, D.A.; Humphreys, B.L.; McCray, A.T. The unified medical language system. *Methods Inf. Med.* **1993**, *32*, 281.
23. Bashyam, V.; Divita, G.; Bennett, D.B.; Browne, A.C.; Taira, R.K. A normalized lexical lookup approach to identifying UMLS concepts in free text. *Stud. Health Technol. Inform.* **2007**, *129*, 545.
24. Spasić, I.; Sarafranz, F.; Keane, J.A.; Nenadić, G. Medication information extraction with linguistic pattern matching and semantic rules. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 532–535. [[CrossRef](#)] [[PubMed](#)]
25. Thapa, S.; Adhikari, S.; Naseem, U.; Singh, P.; Bharathy, G.; Prasad, M. Detecting Alzheimer’s Disease by Exploiting Linguistic Information from Nepali Transcript. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 17 November 2020; Springer: Bermany, Germany, 2020; pp. 176–184.
26. Hamon, T.; Grabar, N. Linguistic approach for identification of medication names and related information in clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 549–554. [[CrossRef](#)]
27. Mack, R.; Mukherjea, S.; Soffer, A.; Uramoto, N.; Brown, E.; Coden, A.; Cooper, J.; Inokuchi, A.; Iyer, B.; Mass, Y.; et al. Text analytics for life science using the unstructured information management architecture. *IBM Syst. J.* **2004**, *43*, 490–515. [[CrossRef](#)]
28. Esuli, A.; Marcheggiani, D.; Sebastiani, F. An enhanced CRFs-based system for information extraction from radiology reports. *J. Biomed. Inform.* **2013**, *46*, 425–435. [[CrossRef](#)]
29. Qazi, A.; Bhowmik, C.; Hussain, F.; Yang, S.; Naseem, U.; Adebayo, A.A.; Gumaei, A.; Al-Rakhami, M. Analyzing the Public Opinion as a Guide for Renewable-Energy Status in Malaysia: A Case Study. *IEEE Trans. Eng. Manag.* **2021**, 1–15. [[CrossRef](#)]
30. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
31. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), San Francisco, CA, USA, 28 June–1 July 2001.
32. Naseem, U.; Khushi, M.; Khan, S.K.; Waheed, N.; Mir, A.; Qazi, A.; Alshammari, B.; Poon, S.K. Diabetic Retinopathy Detection Using Multi-layer Neural Networks and Split Attention with Focal Loss. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 17 November 2020; Springer: Bermany, Germany, 2020; pp. 26–37.
33. Gan, H.; Li, Z.; Wu, W.; Luo, Z.; Huang, R. Safety-aware graph-based semi-supervised learning. *Expert Syst. Appl.* **2018**, *107*, 243–254. [[CrossRef](#)]
34. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [[CrossRef](#)]
35. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
36. Carreras, X.; Màrquez, L. Boosting Trees for Anti-Spam Email Filtering. *arXiv* **2001**, arXiv:cs/0109015.
37. Naseem, U.; Razzak, I.; Eklund, P.; Musial, K. Towards Improved Deep Contextual Embedding for the identification of Irony and Sarcasm. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
38. Hammouda, K.M.; Kamel, M.S. Efficient Phrase-Based Document Indexing for Web Document Clustering. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1279–1296. [[CrossRef](#)]
39. Naseem, U.; Khan, S.K.; Razzak, I.; Hameed, I.A. Hybrid Words Representation for Airlines Sentiment Analysis. In *AI 2019: Advances in Artificial Intelligence*; Liu, J., Bailey, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 381–392.
40. Naseem, U.; Razzak, I.; Musial, K.; Imran, M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Gener. Comput. Syst.* **2020**, *113*, 58–69. [[CrossRef](#)]
41. Naseem, U.; Razzak, I.; Khushi, M.; Eklund, P.W.; Kim, J. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Trans. Comput. Soc. Syst.* **2021**, 1–13. [[CrossRef](#)]
42. Naseem, U.; Khan, S.K.; Farasat, M.; Ali, F. Abusive Language Detection: A Comprehensive Review. *Indian J. Sci. Technol.* **2019**, *12*, 1–13. [[CrossRef](#)]
43. Naseem, U.; Razzak, I.; Hameed, I.A. Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. *Aust. J. Intell. Inf. Process. Syst.* **2019**, *15*, 69–76.
44. Naseem, U.; Musial, K. Dice: Deep intelligent contextual embedding for twitter sentiment analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 953–958.
45. Gupta, V.; Lehal, G. A Survey of Text Mining Techniques and Applications. *J. Emerg. Technol. Web Intell.* **2009**, *1*. [[CrossRef](#)]
46. Aggarwal, C.C.; Reddy, C.K. *Data Clustering: Algorithms and Applications*; CRC Prints: Boca Raton, FL, USA, 2013.
47. Naseem, U.; Khushi, M.; Reddy, V.; Rajendran, S.; Razzak, I.; Kim, J. BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition. *arXiv* **2020**, arXiv:2009.09223.

48. Naseem, U.; Musial, K.; Eklund, P.; Prasad, M. Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
49. Naseem, U.; Razzak, I.; Eklund, P.W. A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on twitter. *Multimed. Tools Appl.* **2020**, 1–28. [[CrossRef](#)]
50. Naseem, U.; Razzak, I.; Khan, S.K.; Prasad, M. A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models. *arXiv* **2020**, arXiv:2010.15036.
51. Yao, L.; Liu, H.; Liu, Y.; Li, X.; Anwar, M. Biomedical Named Entity Recognition based on Deep Neural Network. *Int. J. Hybrid Inf. Technol.* **2015**, *8*, 279–288. [[CrossRef](#)]
52. Li, L.; Jin, L.; Jiang, Y.; Huang, D. Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Springer: Cham, Switzerland, 2016.
53. Zeng, D.; Sun, C.; Lin, L.; Liu, B. LSTM-CRF for Drug-Named Entity Recognition. *Entropy* **2017**, *19*, 283. [[CrossRef](#)]
54. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
55. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2018**, *34*, 1381–1388. [[CrossRef](#)]
56. Jin, Q.; Dhingra, B.; Cohen, W.W.; Lu, X. Probing Biomedical Embeddings from Language Models. *arXiv* **2019**, arXiv:1904.02181.
57. Zhu, H.; Paschalidis, I.C.; Tahmasebi, A.M. Clinical Concept Extraction with Contextual Word Embedding. *arXiv* **2018**, arXiv:1810.10566.
58. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676.
59. Khan, S.K.; Farasat, M.; Naseem, U.; Ali, F. Performance evaluation of next-generation wireless (5G) UAV relay. *Wirel. Pers. Commun.* **2020**, *113*, 945–960. [[CrossRef](#)]
60. Khan, S.K.; Naseem, U.; Siraj, H.; Razzak, I.; Imran, M. The role of UAVs and mmWave in 5G: Recent advances, and Challenges. *Trans. Emerg. Telecommun. Technol.* **2020**, e4241. [[CrossRef](#)]
61. Khan, S.K.; Naseem, U.; Sattar, A.; Waheed, N.; Mir, A.; Qazi, A.; Ismail, M. UAV-aided 5G Network in Suburban, Urban, Dense Urban, and High-rise Urban Environments. In Proceedings of the 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 24–27 November 2020; pp. 1–4.
62. Khan, S.K.; Farasat, M.; Naseem, U.; Ali, F. Link-level Performance Modelling for Next-Generation UAV Relay with Millimetre-Wave Simultaneously in Access and Backhaul. *Indian J. Sci. Technol.* **2019**, *12*, 1–9.
63. Si, Y.; Wang, J.; Xu, H.; Roberts, K. Enhancing clinical concept extraction with contextual embeddings. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1297–1304. [[CrossRef](#)] [[PubMed](#)]
64. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *arXiv* **2019**, arXiv:1901.08746. [[CrossRef](#)] [[PubMed](#)]
65. 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.
66. Chen, Y.; Mani, S.; Xu, H. Applying active learning to assertion classification of concepts in clinical text. *J. Biomed. Inform.* **2012**, *45*, 265–272. [[CrossRef](#)]
67. Boström, H.; Dalianis, H. De-identifying health records by means of active learning. *Recall (micro)* **2012**, *97*, 90–97.
68. Figueroa, R.L.; Zeng-Treitler, Q.; Ngo, L.H.; Goryachev, S.; Wiechmann, E.P. Active learning for clinical text classification: Is it better than random sampling? *J. Am. Med. Inform. Assoc.* **2012**, *19*, 809–816. [[CrossRef](#)] [[PubMed](#)]
69. Chen, Y.; Lasko, T.A.; Mei, Q.; Denny, J.C.; Xu, H. A study of active learning methods for named entity recognition in clinical text. *J. Biomed. Inform.* **2015**, *58*, 11–18. [[CrossRef](#)]
70. Rosales, R.; Krishnamurthy, P.; Rao, R.B. Semi-supervised active learning for modeling medical concepts from free text. In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 13–15 December 2007; pp. 530–536.
71. Herrero-Zazo, M.; Segura-Bedmar, I.; Martínez, P.; Declerck, T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inform.* **2013**, *46*, 914–920. [[CrossRef](#)] [[PubMed](#)]
72. Krallinger, M.; Rabal, O.; Akhondi, S.A.; Pérez, M.P.; Santamaría, J.; Rodríguez, G. Overview of the BioCreative VI chemical-protein interaction Track. In Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, Bethesda, MD USA, 18–20 October 2017; Volume 1, pp. 141–146.
73. Baker, S.; Silins, I.; Guo, Y.; Ali, I.; Högborg, J.; Stenius, U.; Korhonen, A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* **2016**, *32*, 432–440. [[CrossRef](#)]