# Coventry University

## Faculty of Engineering, Environment and Computing

## School of Computing, Electronics and Mathematics

**Data Science and Computational Intelligence**

7151CEM- Computing Individual Research Project

**Citation Screening using Deep Learning**

Author: Mohnish Chaudhary

SID: 11909183

Supervisor: Dr. Xiaorui Jiang

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science and Computational Intelligence

Academic Year: 2022/23

## Declaration of Originality

I declare that this project is all my own work and has not been plagiarised in whole or in part from any other source, unless properly attributed. The elements of the project that are based on published sources such as scientific papers, books, magazines, or the internet have been cited in the references section. I agree to preserve an electronic copy of this project in order to detect and prevent plagiarism.

## Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk

## Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below

Signed: Mohnish Chaudhary Date:7/12/2022

| | |
|---|---|
| First Name | Mohnish |
| Last Name | Chaudhary |
| Student ID number | 11909183 |
| Ethics Application Number | P142661 |
| 1st Supervisor Name | Dr. Xiaorui Jiang |
| 2nd Supervisor Name | Dr. Soroush Abolfathi |

# Table of Contents

## Abstract:

A systematic review is a procedure in which researchers summarise the medical literature by synthesising the findings of multiple primary studies that are related to one another and important to a subject, and which are useful in decision making by giving evidence. In the process of systematic review, the literature is first filtered by looking at abstracts, which is known as citation screening, and then the researchers study the entire text of the publications in the next step. This process necessitates researchers going through each abstract in the initial stage of the systematic study, which takes a long time. Several studies have attempted to shorten this time and assist researchers in focusing on the most important aspects of a systematic study. Most research for citation screening has employed traditional machine learning techniques like SVM and logistic regression which requires selection of features. One significant disadvantage of this strategy is that it necessitates the identification of the appropriate collection of features, which is not a simple operation. There have been few research studies in which deep learning has been employed for this job. Deep learning has the extra benefit of learning the key features without us having to offer this information to the model. In the study, we are attempting to use a state-of-the-art transformer-based model that has demonstrated success on most NLP tasks. We attempted BioLinkBERT and BiomedNLP-PubMedBERT-base-uncased-abstract (hereafter referred as PubMedBERT) models, which are explicitly trained on PubMed abstracts. So, these models have the extra benefit of having previous knowledge of the PubMed content, and by fine tuning, we are adjusting the models to execute on the task of citation filtering. We have used 23 publicly available systematic study datasets from this study. There are also other constraints to the dataset utilised in the study, one of which is that it is relatively tiny; to go around this, we tried reasoning through language translation. PICO has shown to be useful in capturing the key components of an abstract for medical publications. In this experiment, we used PICO-based characteristics to investigate if they can aid in citation screening. Our study shows that using transformer-based models filter out 48%of citations and for the datasets which are bigger than number is close to 70%.

## Acknowledgements

I'd like to express my sincere gratitude to my project supervisor, Dr. Xiaorui Jiang, for his guidance and assistance during the project. Dr. Xiaorui assisted me by providing suggestions and ideas anytime I got stuck in the project. He also offered past studies and research connected to the notion so that I could fully comprehend the concept and apply it to this project. He helped me overcome conceptual barriers, difficulties, and challenges that I had encountered during the process. He assisted me in running all the codes on a powerful GPU, which aided in conducting all of the experiments in parallel. This project could not have been accomplished without his assistance since a model of this scale cannot be trained on standard laptop computers.

# 1. Introduction

Evidence-based medicine refers to the utilisation of published literature as the most recent evidence for patient care decision making. Previously, physicians were obligated to search the primary literature for data relevant to their patients. The present notion of evidence-based medicine is based on filtration of articles in the form of systematic reviews, which has evolved as the discipline has advanced. Systematic reviews compare several kinds of medications used to treat illnesses, such as oral opioids and oestrogen replacement therapy.

For each systematic review, thousands of papers must be examined, triaged, and summarised. Before conducting any systematic review, the topic that has to be answered by that review is identified, as are the parameters that the article must meet in order to be included in that research. For systematic review, a set of conditions in the form of inclusion and exclusion criteria is defined. This criterion might be as basic as literature from a given language, a conference, or something more precise like which age group the research was performed on, or the population count on which the trials were conducted. After identifying the inclusion and exclusion criteria, the following step is to filter the citations based on these criteria. In the first stage, the abstracts of the papers are evaluated to assess whether they meet the inclusion requirements. Only papers with abstracts that meet the inclusion requirements are processed in the second phase. In the second part, reviewers read the whole text of the article and filter out any particles that do not meet the inclusion criteria in any manner. The reviewers keep the systematic review detailed, displaying the search methods, the articles that were reviewed by abstracts and the ones that were fully reviewed, leading to the shortlisted articles that meet the inclusion criteria with sufficient evidence and should be included in the final review.

## 1.1. Research Problem

Systematic review is a time-consuming procedure that might take anywhere from 6 months to a year to complete. Every year, around 4000 systematic studies are conducted, and according to certain research, public systematic reviews require updating within two years after completion. Systematic review is a time-consuming procedure that must be updated on a regular basis due to new publications in the market on the same topic. To further emphasise the issue, Medline alone has over 13 million references and over 4800 biomedical and health papers, making manual screening even more challenging. Automating this process would be extremely advantageous to the medical community. In this project, we will attempt to automate the first stage of a systematic review by evaluating citations based on the content of the abstract and title, such that the reviewers have to just go through a subset of filtered abstracts and not all the abstracts for systematic review. The figure 1 depicts an overview of the systematic review method.

The automated citation screening process will help the researchers in following ways

- Researchers will be able to find candidates for systematic reviews using the automated citation screening procedure, which will assist them in determining whether they have enough freshly created literature and it is time to amend the current systematic studies with the new publications.
- Because the solution automates the initial phase of the systematic review, it speeds up the process because less work is required by humans to filter out the citations for the systematic review.
- Because the algorithm generates a confidence score along with the included papers, it may be useful for researchers to prioritise articles based on this score.

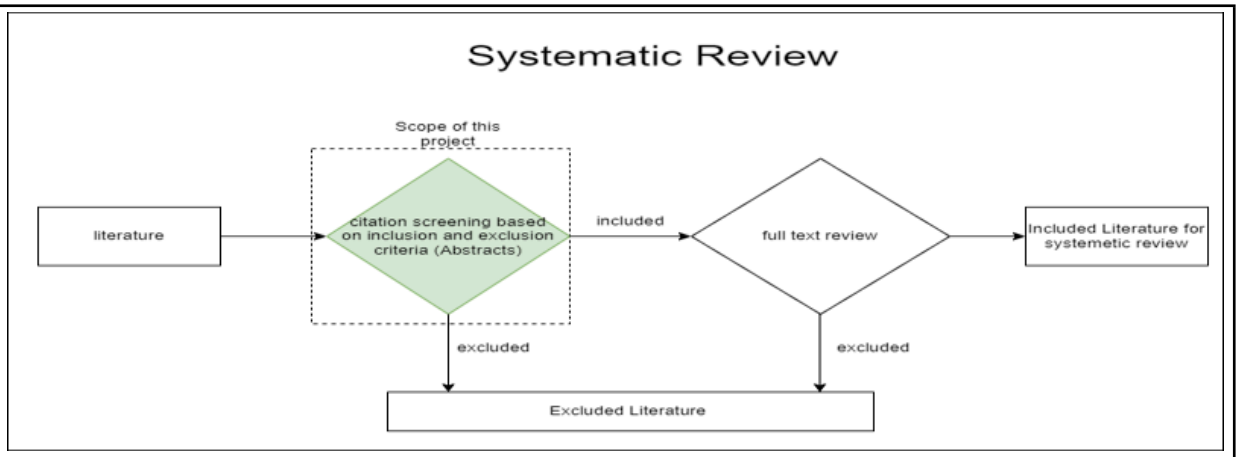


*Figure 1:Systematic Review process*

In the study we are trying to automate the first part of the systematic review where the abstracts are used to filter out the literature for systematic review. We will approach this as a binary classification issue, with citations that meet the inclusion requirements forming the positive class and citations that do not meet the inclusion criteria forming the negative class. We will develop a classifier using supervised machine learning to predict whether each abstract submitted should be included or excluded from the systematic review. Much previous research has been conducted to answer this problem; some of these studies attempted to solve it as a binary classification problem, as we are doing now, while others attempted to solve it as a ranking problem. Previously suggested research included the solution of citation screening utilising a traditional machine learning-based strategy, with algorithms such as SVM and logistic regression employed to solve this. One important disadvantage of such techniques is that handcrafted features are required for this purpose. In the context of this challenge, this entails determining the appropriate collection of phrases for the model to employ in distinguishing between included and excluded citations. There have been few experiments in which deep learning-based models have been trained for the same purpose. However, these studies have not looked at the cutting-edge transformer-based models that have shown to break all existing records on NLP tasks. Another essential factor to consider is the difficulties related with the dataset used in the study. We utilised 23 freely accessible systematic review datasets. One issue with this problem is the highly unbalanced nature of the data. Aside from that, the number of data points in some datasets are insufficient to train a deep learning model. To address these issues, we propose a transformer base model trained on the PubMed dataset. This suggests that the model is already familiar with the vocabulary in this area, and we just need to change its weights to make it grasp the task of citation filtering. Another benefit of deep learning models is that we do not need to identify and offer significant features to the deep learning models since the deep learning models can detect important features on their own.

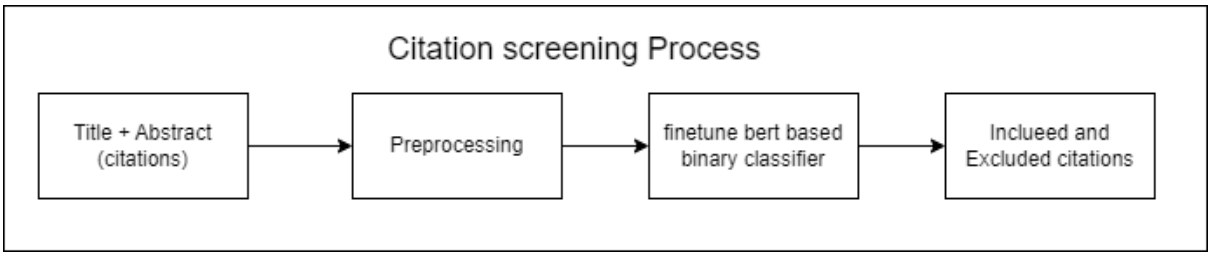


*Figure 2: Citation screening process*

### 1.2. Motivation

Due to the limits of current methodologies, dataset restrictions, and advancements in the field of natural language processing, our solution might tackle the following three aspects.

- The most significant feature of using the transformer-based model for this study is to use the transformer-based models' pretrained knowledge on medical papers to improve the outcomes in the field of citation screening.
- Another key part of the project is to experiment with data augmentation through language translation to improve the outcomes.
- In this investigation, we will also employ PICO phrases as input features to the citation screening classifier. The goal of this experiment is to deliver significant information to the model while reducing noise from the training data. Previous research has demonstrated that PICOS are key information identifiers in the medical literature, and it is believed that they will also generate positive results for citation screening.

### 1.3. Research Questions

To conclude the study's findings, we require answers to the following questions.

- One of the study's most critical questions is whether the transformer's best pre-trained model can aid in automatic citation filtering and assist researchers
- Aside from that, we'd like to see if language translation-based augmentation will help us enhance our outcomes.
- One other question is whether PICO sentences will be able to offer positive results for citation screening.
- To test if adding a convolution layer at the conclusion of a transformer-based model and finetuning for citation filtering

## 2. Literature review

Cohen's study (Cohen et al., 2006) using automated citation classification to reduce workload in systematic preparation is the first of its kind. The authors attempted to answer a critical question by conducting this study: whether machine learning can be used to automate the citation screening process. It is the study that published 15 drug-related datasets that are still used to benchmark solutions for citation screening today. The author also formalised the evaluation metric for the citation screaming task, which is used in all subsequent studies. This metric is defined by the author as work saved over sampling at 95% recall. The author's solution was a voting perceptron-based automated citation classification system that classifies each citation as high- or low-quality drug class

specific evidence. Because the datasets were small and unbalanced, the author used cross validation experiments to improve performance. The author was able to conclude from the study that citation classification could be a useful tool in maintaining systematic reviews of the efficacy of drug reviews. The author also highlighted important features for all datasets in his study, such as medical subject headings and publication type. The author obtained positive results for 11 out of the 15 datasets, with three datasets achieving greater than 50% reduction.

The author (Kontonatsios et al., 2020) used a neural network-based feature extraction method for citation screening in this study. The author created a denoising auto encoder that learned representation from corrupted text. The author created an architecture in which three parallel auto encoders were used, and the result was concatenated and passed through several dense layers. The author also experimented with various combinations of auto encoders, ranging from a single auto encoder to multiple auto encoders of varying sizes. The author's best result was found to be the result model trained on features of three different auto encoders followed by a dense neural network. Along with features from auto encoders, the author also used features from LDA, MeSH, and BOW, which were inspired by Cohen's studies (Cohen et al., 2006); however, the author demonstrated that the features generated by the denoising autoencoder produce the best results. These features were fed into an SVM model for classification. Apart from epochs for fine tuning to find the right number, the author stated that he did not use any dataset specific hyper tuning for all datasets. The author used two different datasets for finetuning: stains and the BPA dataset. The epochs for those are fixed based on the relative size of the other datasets.

The author has not done any hyper tuning with respect to the dataset, and since there is no explicit validation data, it is difficult for further studies to compare the result with the study as taking out a small portion of data as validation split will change the distribution of train and test. Even though the author used two datasets for hyper tuning the epochs, there is no mention of the data used for validation for hyper tuning the number of epochs. During the feature extraction phase using autoencoders, the author used the entire dataset regardless of whether it belongs to the train or test set, resulting in contamination. Even if the auto encoder does not train the model for classification, the model learns a better representation of the text data, which leads to contamination in the wrong way. Recent studies have highlighted this issue in their research, stating that even when people use language models, they remove the test dataset that was used to train the language model.

In the study (van Dinter et al., 2021), input text utilised for the task is a mix of the publication's title and abstract, which the authors of the study employed for the multi-channel convolutional neural network that was used to classify the supplied collection of citations. To select the most effective model, the author also tried out other convolutional combinations. As features for the neural network, pre-trained glove-based embeddings were used. The major motivation for utilising a convolutional neural network for text input is to be able to capture the relationships between words that occur in a short proximity that is covered by the kernel during the convolution process. As this study has highlighted that convolution neural network are influential in producing results for citations screening, this has inspired us to use CNN with a transformer model.

Utilizing a denoising auto encoder to discover a feedforward neural network to generate the features for an SVM-based classifier and training a multi-channel convolution neural network using globe-based embeddings, the authors (van Dinter et al., 2021) of the study attempted to duplicate the findings of the previous two investigations. The author has made an effort to contrast the

findings of the two separate research and to highlight the difficulties in reproducing the findings. Along with this, the authors outlined the shortcomings of the earlier investigations, such as the absence of hyper-tuning in those earlier trials. Along with this, the author has also offered a fresh approach using a fasttext-based neural network that can provide results with comparable accuracy to earlier research but requiring very little training effort. While replicating previous studies, the author has stated that the availability of the code base does not guarantee their application of results. In some cases, even when the code base was available, the author was unable to produce a similar result. The author also emphasised the importance of the environment, which must be maintained in order to replicate the results. In some cases, random seeds and a different split of the train and test split set can result in different results; however, if the same information was provided in the codebase, it would have been easier to replicate the studies. This replicability study has highlighted the shortcomings of the previous 2 studies and given valuable guidelines which will help us in designing and maintaining our experiments better.

The authors of this study (Winata et al., 2021) emphasised the use of language models as a few short learners. The authors demonstrated the use of language models for all NLP tasks and compared it to fully trained and fine-tuned models. Few short learning has been divided into three types: zero shot learning, one shot learning, and a few short learning. In zero shot learning, no example is provided to the model to learn the task; in one shot learning, only one example is provided to the model to learn the task; and in few short learning, examples ranging from 10 to 100 are provided for the model to understand the task. The important thing to understand in this study is that the actual weight of the models is not changed in any way, and the task examples are only provided to identify the task. In the case of tasks such as zero shot learning, the task is provided as part of the query. The authors demonstrated that in some cases, a few short learning models can outperform fine-tuned models. The author has also mentioned contamination in some of the NLP tasks. Contamination is the process by which data used by language models becomes part of the test dataset. The goal of reading this paper is to understand how to properly use a language model for the task of citation screening, as well as to become aware of the contamination process. Another important takeaway from this paper is to become acquainted with all of the NLP tasks as well as the datasets used to benchmark results for the tasks. It also presents the current state of the art results for the respective NLP tasks.

The authors of this study (Wang et al., 2021) developed a solution to generate elements such as population intervention and outcomes from medical literature. The concept behind PICO extraction is that these quantities are part of the inclusion and exclusion criteria for some of the systematic review questions, and automated extraction of PICO elements can aid in clinical studies. For the extraction of PICO entities, the author used 400 abstracts from the PubMed dataset. The solution is divided into two parts: the first part uses a BERT-based classifier to extract PICO sentences from the text, and the second part extracts this entity from the sentences selected in the first part. The study is relevant to our study because it emphasises the use of PICO elements as baseline contributors for answers to the including and excluding criteria in systematic reviews. The authors obtained an approximate 85% F1 score for sentence classification using PICO entities and a 70% F1 score for PICO entity extraction. This study inspired us to determine whether PICO sentences or entities can be used as one of the features for citation screening.

The authors of this (Howard et al., 2016) study evaluated a tool called swift review for citation screening. The application's concept is to treat citation screening as a ranking problem, with each abstract ranked based on its likelihood of inclusion in the systematic study. The author used

frequency best features as well as Latent Dirichlet Allocation LDA features. The authors also evaluated the study using features such as MeSH terms, n-grams, and topic model membership. The authors also discussed how to use the swift review tool for annotation and how it can be used for document prioritisation for citation screening. Similar to this (van de Schoot et al., 2021) provides detailed information about the ASReview tool, which could be used for systematic reviews and meta-analysis. The authors compared it to existing market tools. The authors also discuss how active learning can be used for citation screening. The ASReview tool includes a wide range of algorithms, from basic machine learning to advanced deep learning-based methods. Algorithm to use is determined based on the type of data and the user's preference. Other features include balancing strategy by dynamic sampling under sampling. The approach's best feature is that it can use a human in the loop for active learning, where after each set of iterations, the human provides feedback on whether the predictions are correct or not, and once a desirable accuracy is achieved, the model can be used for further tasks without any human intervention. The study highlights the positive aspects of ASReview tool and compares it to the market's existing tools. Similarly in the study (D'Ambrosio et al., 2022), the authors demonstrated an open-source integrated framework for full citation collection and screening. It is one of a kind in that not only the citation screening but also the citation collection is handled by the same tool. It is linked to various databases and searches for relevant documents using unified query syntax. The author describes the main components as an integrated query-based search and management engine that aids in locating. A Bayesian active machine learning-based citation classifier and a data-driven search query generation algorithm. These three studies help in understanding the citation screening process from the end user’s perspective and help us in designing a better solution.

The authors of the study (Ioannidis, 2021) created solutions for document street screening, which is an important part of evidence-based medicine. The author proposes a deep learning mate solution in the study that uses a transformer-based model PubMed BERT and sentence BERT to generate features in the form of embedding. The author compares the results of information retrieval methods such as BM 25 and RM3 to modern deep learning approaches. One solution proposed by the author was to represent each word in a sentence using BERT based embedding and then classify all these sentences to get an overall score for the document. The similarity score between the query and the input document is represented using sentence BERT embeddings in the second approach. The author also attempted to compare the outcomes of all the preceding experiments. Finally, the author was able to conclude that by combining deep learning methods with the baseline IR approach, the author was able to demonstrate the efficacy of deep learning methods.

The studies show that transformer based fine tuning is a viable option to get a good result in the field of medical NLP when it has been used in certain tasks however not on this dataset. Also, PICO entity extraction is one of the important features as authors have described it as a criterion that helps in distinguishing whether publication should be included or excluded from a systematic review. The preceding works provide inspiration for using a pre-trained model from the biomedical area to get good outcomes. In the realm of NLP, the augmentation has shown to be quite useful for unbalanced datasets. Based on the preceding research, we hypothesised that utilising just the essential relevant information in PICO phrases and deleting all extraneous information from the abstract might offer the model superior features and learning. Some of the research exposed the shortcomings of previous studies, which prompted us to discover a remedy to those shortcomings. The review tool studies assisted us in understanding the end-user needs for citation filtering.

# 3. Methodology

The creation of a solution begins with the selection of the appropriate dataset required to conduct the underlying experiment. Once the correct data has been located, we require scripts and tools to download it. One critical point to remember here is that we must determine whether we have the necessary rights to access user data for the work at hand. This is accomplished by issuing an ethical certificate demonstrating that this data is suitable for the intended purpose and that there are no constraints. The solution to this problem began with the collection of needed datasets for a systematic review and the generation of an ethics certificate simultaneously.

Previously, 23 datasets from systematic research were employed in previous studies. This data set was acquired using the PubMed IDs of the studies and the scripts supplied in the prior research (Kusa et al., 2022). In general, training a deep learning model entails pre-processing the input data to reduce noise and make it appropriate for acceptance by the model. The second stage is to build a deep learning model in order to hyper-tune it. In the third phase, we analyse the model using multiple metrics and personally examine the error data points to see which cases the model is making incorrect conclusions and how they might be improved further.

Following the examination of data points, any improvements that may be made are implemented in the form of modifications in the training data and appropriate adjustments in the deep learning model architecture. After making these adjustments, a new model trend emerges, and the entire cycle begins. As a result, these studies involve a cyclic process in which the model improves from one cycle to the next. And once we've completed all of the experiments, we'll have a better ultimate answer. The figure below shows this approach and helps users understand how a deep learning model is trained over time through an iterative method.

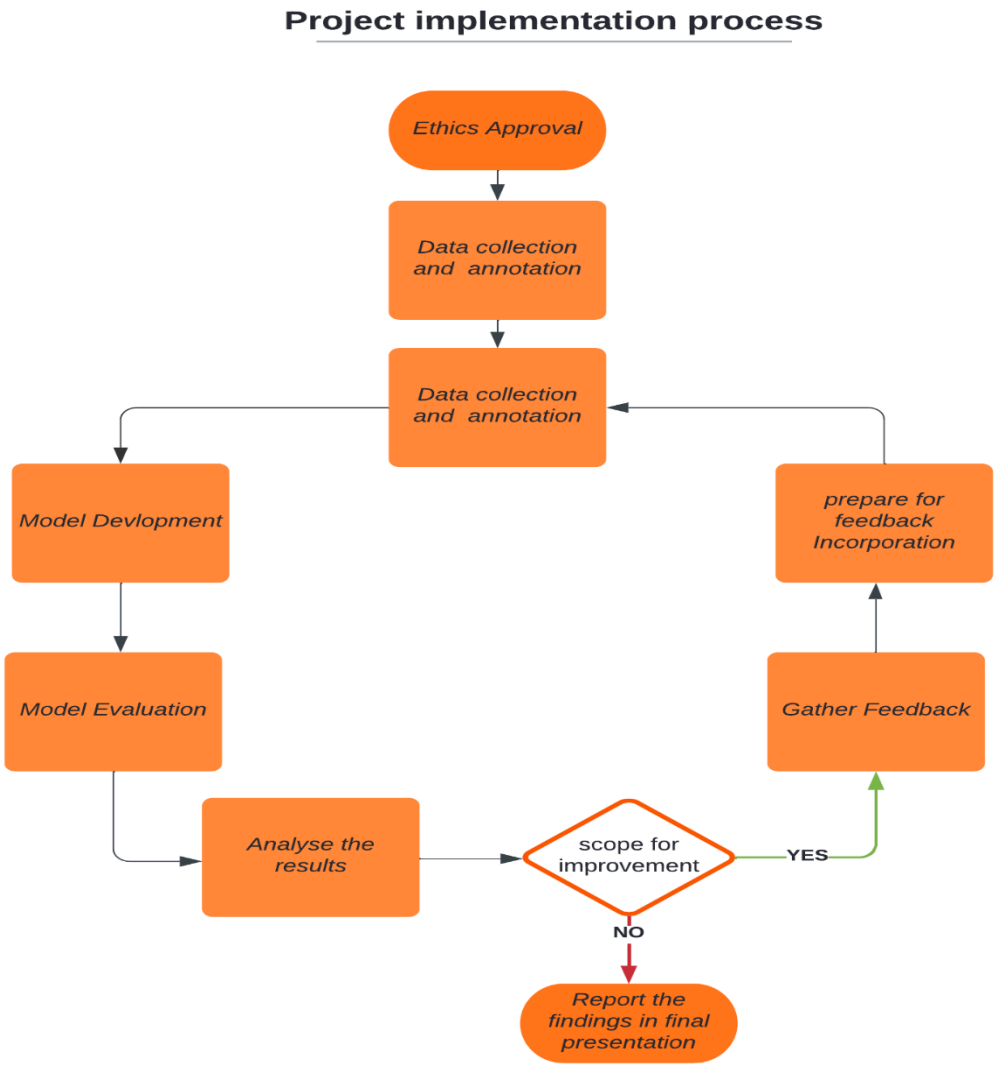


*Figure 3: Iterative model improvement process*

## 3.1. Dataset

In this study, we used 23 publicly accessible datasets from systematic studies, as described in (Kusa et al., 2022). 15 of the 23 datasets are a mash-up of 15 unique systematic review topics given by the Oregon Evidence-based Practice Centre (EPC) and connected to therapeutic effectiveness in various pharmacological classes. Three more datasets from systematic reviews concerning the clinical results of various therapies. The third dataset collection consists of five substantially larger reviews that were used to assess the performance of the SWIFT-review tool. They were created using broader search strategies, which explains why they have more citations. These datasets contain the title of the publication, along with abstract for 18 out of the 23 datasets we also have the bibliographic metadata available for the publication. One of the dataset skeletal muscle relaxants have only 0.5% of the data points included citations, in absolute number 9 positive data points are very less to train a

machine learning model or a deep learning model. It shows that there is a high imbalance in the data for the studies. the range of included citations vary from 0.5% 27%, phenomena of class imbalance is throughout all the 23 datasets. previous studies have used the subset of this dataset for citation screening, we have tried to generate results on most of them to compare with the previous studies. For the purpose of the study, we will concatenate the title and abstract as one text and use it for the training model.

| | Dataset name | Introduced in | # Citations | Included citations | Excluded citations | Maximum WSS@95% | Bibliographic metadata |
|---|---|---|---|---|---|---|---|
| 1 | ACEInhibitors | Drug (Cohen et al., 2006 ) | 2544 | 41 (1.6%) | 2503 (98.4%) | 93.47% | Yes |
| 2 | ADHD | | 851 | 20 (2.4%) | 831 (97.6%) | 92.77% | Yes |
| 3 | Antihistamines | | 310 | 16 (5.2%) | 294 (94.8%) | 89.84% | Yes |
| 4 | Atypical Antipsychotics | | 1120 | 146 (13.0%) | 974 (87.0%) | 82.59% | Yes |
| 5 | Beta Blockers | | 2072 | 42 (2.0%) | 2030 (98.0%) | 93.07% | Yes |
| 6 | Calcium Channel Blockers | | 1218 | 100 (8.2%) | 1118 (91.8%) | 87.20% | Yes |
| 7 | Estrogens | | 368 | 80 (21.7%) | 288 (78.3%) | 74.35% | Yes |
| 8 | NSAIDs | | 393 | 41 (10.4%) | 352 (89.6%) | 85.08% | Yes |
| 9 | Opioids | | 1915 | 15 (0.8%) | 1900 (99.2%) | 94.22% | Yes |
| 10 | Oral Hypoglycemics | | 503 | 136 (27.0%) | 367 (73.0%) | 69.16% | Yes |
| 11 | Proton PumpInhibitors | | 1333 | 51 (3.8%) | 1282 (96.2%) | 91.32% | Yes |
| 12 | Skeletal Muscle Relaxants | | 1643 | 9 (0.5%) | 1634 (99.5%) | 94.45% | Yes |
| 13 | Statins | | 3465 | 85 (2.5%) | 3380 (97.5%) | 92.66% | Yes |
| 14 | Triptans | | 671 | 24 (3.6%) | 647 (96.4%) | 91.57% | Yes |
| 15 | Urinary Incontinence | | 327 | 40 (12.2%) | 287 (87.8%) | 83.38% | Yes |
| | Average Drug | | 1249 | 56 (7.7%) | 1192 (92.3%) | 87.67% | 15/15 |
| 16 | COPD | Clinical (Wallace et al., 2010) | 1606 | 196 (12.2%) | 1410 (87.8%) | 83.36% | No |
| 17 | Proton Beam | | 4751 | 243 (5.1%) | 4508 (94.9%) | 90.14% | No |
| 18 | Micro Nutrients | | 4010 | 258 (6.4%) | 3752 (93.6%) | 88.87% | No |
| | Average Clinical | | 3456 | 232 (7.9%) | 3223 (92.1%) | 87.45% | 0/3 |
| 19 | PFOA/PFOS | SWIFT (Howard et al., 2016) | 6331 | 95 (1.5%) | 6236 (98.5%) | 93.56% | Yes |
| 20 | Bisphenol A (BPA) | | 7700 | 111 (1.4%) | 7589 (98.6%) | 93.62% | Yes |
| 21 | Transgenerational | | 48638 | 765 (1.6%) | 47873 (98.4%) | 93.51% | Yes |
| 22 | Fluoride and neurotoxicity | | 4479 | 51 (1.1%) | 4428 (98.9%) | 93.91% | No |
| 23 | Neuropathic pain — CAMRADES | | 29207 | 5011 (17.2%) | 24196 (82.8%) | 78.70% | No |
| | Average SWIFT | | 19271 | 1206 (4.6%) | 18064 (95.4%) | 90.66% | 3/5 |
| | Average (All datasets) | | 5454 | 329 (7.0%) | 5125 (93.0%) | 88.29% | 18/23 |

*Figure 4: Datasets metadata (Kusa et al., 2022)*

Since the dataset is highly imbalanced and there are very few samples for the positive class in most of the datasets, we will use K cross validation. Since the fraction of positive samples is very small the result of the experiment depends upon the type of data samples included in testing and training phase and these results can vary a lot from one testing set to another testing set. To standardise the result, we will perform the experiment 10 times on different fractions of train and test set and take average, the same has been done by the previous studies which makes the result compatible with them as well. Along with this, we will also perform data augmentation to improve the count of positive samples during the training phase, it is described in detail in the field in further sections.

## 3.2. Transfer learning

In the study we have used transformer-based model which have shown to break all the records in the field of NLP and hold state of the art results for most of the NLP task. The main idea behind using a transformer base model is that these models are too big having parameters ranging from millions to billions it becomes very difficult Train such a big model from scratch every time. These models are provided as pre trained models which means they are already trained to understand the underlying language and domain and we have to tweak the weights/parameters to adjust according to the task

in hand. To explain this in a bit more detail, pre-training is usually done on a very large amount of data and it takes very high compute power to train the model for several days. Finetuning on the other hand is done on top of the pre trained model. To finetune the model first needs to train pretrained/language model on an additional training data specific to the task we want to use it for. This way we don't need a lot of compute power we can finetune the model on a small dataset and still get good result. This is quite appropriate for this experiment as we have very less training data available for some of the datasets. The transformer model consists of encoder decoder architecture where the encoders and decoders are stacked. Before the invention of the Transformers, RNN were the common deep learning architecture used for such tasks. There are 2 limitations of RNN models. First, they were not able to handle the long-term dependency because of the vanishing gradient problem. second, due to the recurrent nature of RNN they cannot process data in parallel which makes it very difficult to train on a very large dataset. Transformer base models have overcome this problem and hence its possible to train very huge models and use it by transfer learning to the downstream tasks

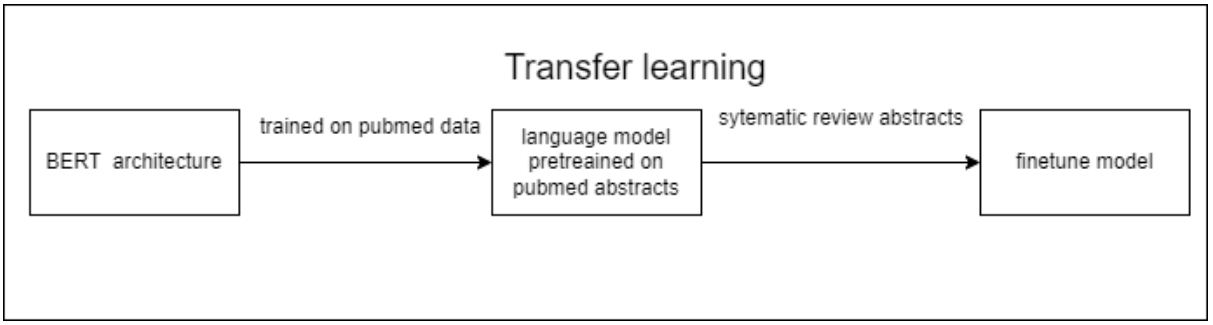


*Figure 5:Finetuning Process*

Transformer library provides a model compatible with the two most common deep learning frameworks tensor flow and the Pytorch. Based on the task in hand the pre-trained model, which is more suitable for the task, boilinkBERT is available only for the pytorch. So, we have done all experiments using the pytorch. Also, the pretrained models used in this study are trained on PubMed abstracts which makes them more suitable for this task and these models have not been explored in the past making this study first of its kind to use these models for citation screening on this specific systematic review datasets. One of the pretrained models used in the study is BioLinkBERT which is trained on PubMed extracts along with the citation information. This model achieves state of the art performance on several biomedical NLP benchmarks such as BLURB and MedQA-USMLE. This model has 110 million parameters and it's trained in the domain of biomedicine which makes it suitable for this task. The other model that we have used is microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract which is trained by Microsoft and it also provides very good results on the above-mentioned benchmark as well.

| **Model** | **BLURB Score** | **PubMedQA** |
|---|---|---|
| BioLinkBERT-base | 84.39 | 70.2 |
| BioLinkBERT-large | 84.30 | 72.2 |
| PubMedBERT | 82.75 | 55.84 |

*Figure 6: Model benchmarks on different tasks*

Here PubMedQA is a Dataset for Biomedical Research Question Answering and BLURB is the Biomedical Language Understanding and Reasoning Benchmark.

Since the models are pre trained we only need to finetune them for the underlying tasks; however these models have a lot of parameters. It's not possible for us to even fine tune them on CPU. to train our models we will be using GPUs.

## 3.3. Evaluation criteria:

To compare the results of different models we need a criterion to evaluate the models. Usually in machine learning and deep learning this evaluation criteria is chosen based on the problem in hand for example in most of the cases for classification we use either the F1 score accuracy or precision based on the description of the problem and the statistics of the data. In the study we will use WSS@95% recall (work saved over sampling): The evaluation criteria (WSS@r%) used in the study measures the reduction in the human screening workload for citation screening due to automation tools. Cohen et al. defined it as "the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out by the classifier)." In simple terms what evaluation criteria measures is out of all the negative samples how many of them have been identified correctly. Here R refers to the recall required for the given task. For example, if it's a very critical problem and we want to recall 1 then the WSS score would show out of all the negative samples how many of them we have already identified. This matrix also makes sense because the goal of the study is to reduce the time taken to filter out the publications which should not be needed in the second part of the systematic review. So, more and more excluded publications we identify in this step more and more time is saved for the researchers which day can utilise in the second part of systematic review where they have to completely read the publications.

$$WSS@r\% = \frac{TN+FN}{N} - (1 - r)$$

For the purpose of the study, we have chosen recall as 95%. That means at least 95% of the included citations should be correctly identified. The reason why we have chosen recall as 95% is because all the previous studies have also fixed the recall 95%, choosing the same recall makes it easier for us to have comparable results. To reach this recall in the test dataset results we sort the outputs and descending order of the score and choose a confidence score threshold above which the recall is 95%.

## 3.4. Baseline

The simplest way to use a transformer base model is to utilize the embedding of the downstream task. In this experiment we concatenated the abstract and the title and generated the embeddings from distil BERT model. These embeddings provided as features to the dense layers of the neural network. Finally, the neural network was trained on the downstream task however the embeddings were not modified in this learning. This approach was not giving good results. The way to implement it in the code is utilise a transformer-based model and then add few dense layers to it, where the value of CLS token is passed to the dense layer. The pretrain transformer models weights were frozen in this case only the following dense layer weights were modified. However, the results were very bad, so we chose to use the results from the previous studies to compare our results instead of a simple baseline.

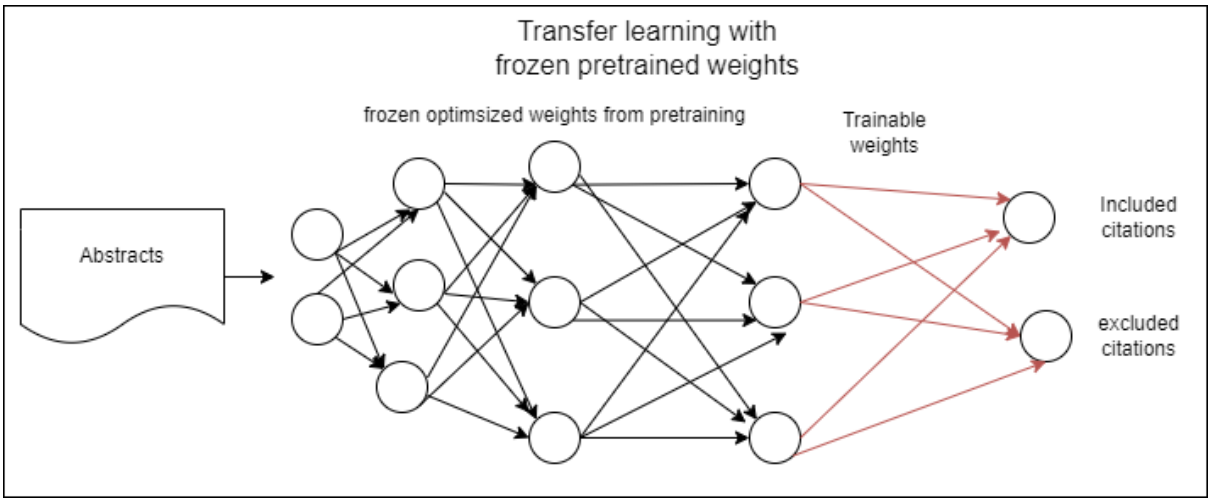


*Figure 7:transfer learning*

# 4. Experiments

## 4.1. Experiment 1: Finetune distilBERT, BioLinkBERT and PubMedBERT

In the fine-tuning phase, we take a pre-trained model, such as DistilBERT, and add a few dense layers, followed by an output layer with the optimal number of outputs based on the task at hand and the appropriate activation function. Because we are dealing with a binary classification problem, we have two neurons in the last layer and later we use the SoftMax function on it. The model is then trained using the training dataset. The weights of the transformer model are also modified according to the training data throughout the fine-tuning procedure. We normally maintain the learning rate modest as we modify the prior pre-trained model's learnings to the train data.

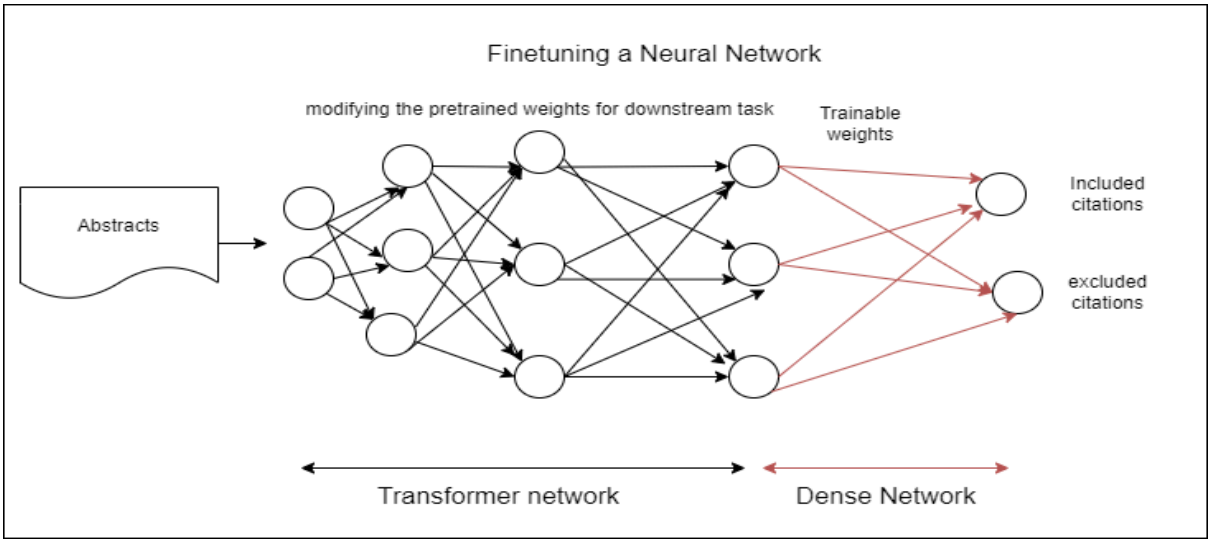


*Figure 8: Finetuning transformer model*

The extent of pre-processing necessary for the experiment is determined by the transformer model selected. We experimented with uncased distilBERT, BioLinkBERT, and PubMedBERT. These models are all trained on uncase data. As part of the preparation, we eliminated any special characters, web URLs, and changed everything to lowercase. The goal of pre-processing is to transform train data into a format that is comparable to the data used by the authors when the pretrained language model

was first trained. In this task, we do not need to remove any stop words or do any lemmatization or stemming since the transformer base model that we are employing learns a context from the surrounding words, thus eliminating these words would modify the context. Furthermore, stemming and lemmatization change the word structure, and certain words may not match the vocabulary of the original model after this, defeating the goal of pre-processing. Also, the authors did not remove stop words during the first training of the pre-trained model in order to preserve the training data as similar to that utilised in the original pre training, we kept the stop words as it is. Once the data has been preprocessed, we should split it into train and test sets. As in prior research, we repeat each experiment ten times with a new seed set to obtain a distinct train and test split. The goal of doing this is to generalise the results because the test set is tiny, and each split of tests yields a distinct result. Normally, when training a deep learning model, we divide the data set into three sets, with the validation set being used to fine-tune the model's parameters. However, because our training data is so small and all previous studies have only divided data into train and test splits, generating a new validation set would harm the models in two ways: first, it would reduce our training dataset, and second, the results would not be comparable to previous studies.

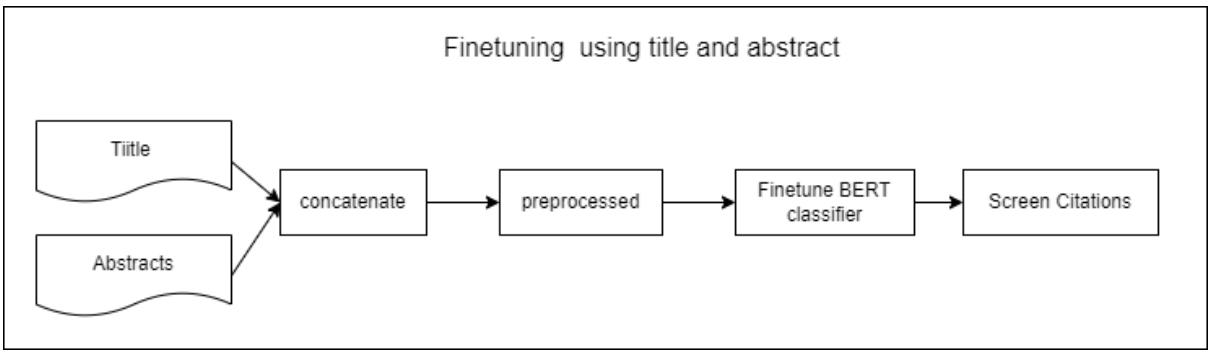


*Figure 9: Finetuning BERT based models using abstract and title*

## 4.2. Experiment 2: Augmentation using language translation

Some of the datasets included in the study are quite tiny and severely imbalanced. One approach to resolving this issue is to enhance the training data. Augmentation is the process of producing extra data from available training data under the constraint that the newly created data is useful and as near to the train set as possible without being distinguishable from it. Language translation is one kind of augmentation in the case of NLP. In this case, we translated the abstracts from English to German and then back again. This double translation alters portions of the content while keeping it useful for human comprehension and usable for training. The study's goal is to expand the sample of included citations during training to give the model with a better balance of data so that it can learn to distinguish between included and excluded citations.

One thing to keep in mind is that we just modified the training data and left the testing dataset untouched. The purpose for not including any data in the test set is to ensure that the findings on the test set are as accurate as possible and free of any contamination.

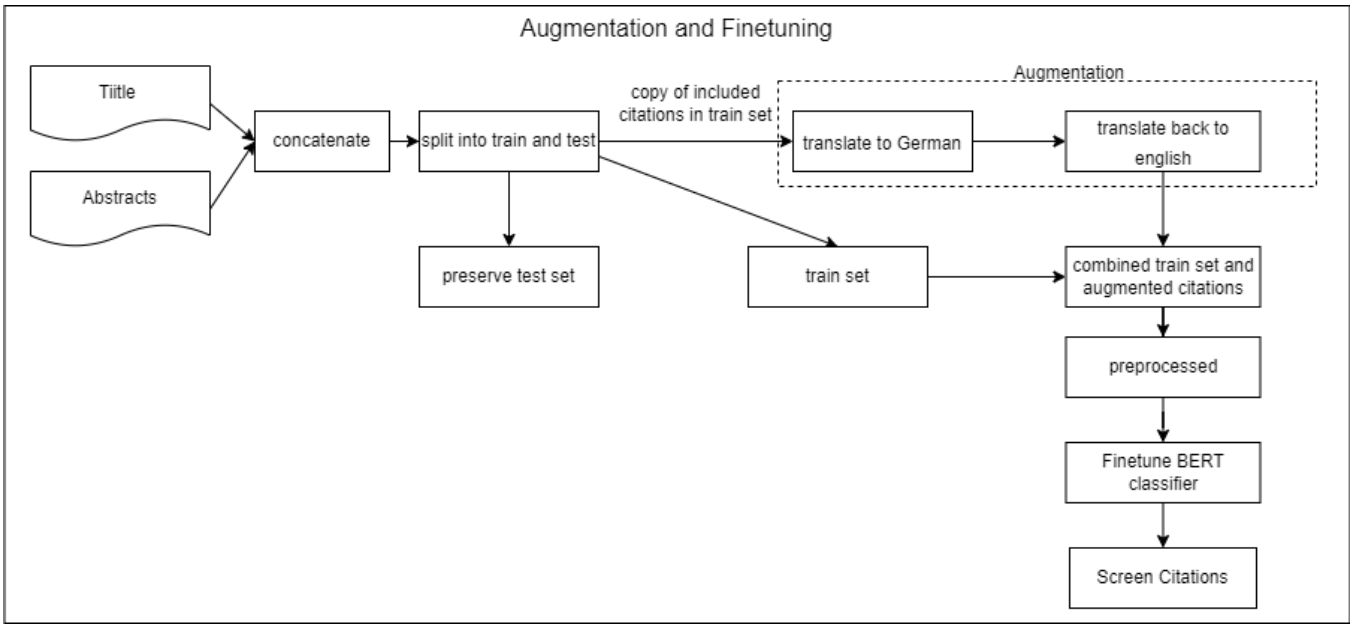


*Figure 10:Fine tuning with train data Augmentation*

## 4.3. Experiment 3: utilization of PICOs

PICO is an acronym that stands for population intervention comparison and outcomes. Before the researchers start any research, they first define a well-built question. PICO is a way to construct a well-built question for research.

| **P**atient/Population | **I**ntervention | **C**omparison | **O**utcomes |
|---|---|---|---|
| Who is your patient?<br>• Age, sex, race or patient<br>• Primary problem<br>• Health status | What do you plan on doing for the patient?<br>• Diagnostic test<br>• Medication<br>• Procedure | What alternative are you considering?<br>• Another test, medication or procedure<br>• Watchful waiting | What do wish to accomplish?<br>• Accurate diagnosis<br>• Relieve or improve symptoms<br>• Maintain function |

*Figure 11: PICO method (anon 2022)*

**Is adherence to the Mediterranean Diet associated with reduced risk of heart attack?**

| **P**atient/Population | **I**ntervention | **C**omparison | **O**utcomes |
|---|---|---|---|
| • Adult<br>• History of heart disease | • Mediterranean diet | • Typical diet<br>• No comparison | • Reduction in heart attacks |

*Figure 12:PICO example (anon 2022)*

Identifying PICO elements serves as the foundation for retrieving and selecting published publications for a systematic review. The goal behind utilising PICO identifiers is to supply the model with the necessary information while removing the clutter and noise from the text. This study's premise is that sentences with PICO entities are more significant than other phrases in abstract. (Wang et al., 2021) Prior research of PICO entity extraction for preclinical animal literature illustrates the results of identifying PICO sentences and PICO entities from the literature. In this investigation, we will apply the PICO sentence classification model (PubMedBERT-abs) from this study(Wang et al., 2021) to detect all sentences containing PICO entities. We will only utilise the PICO sentences and will not filter out the specific entities since we want to offer the entire phrase to the model rather than simply the entities. This is done to ensure that the model does not lose out on any additional information from the context of the statement.

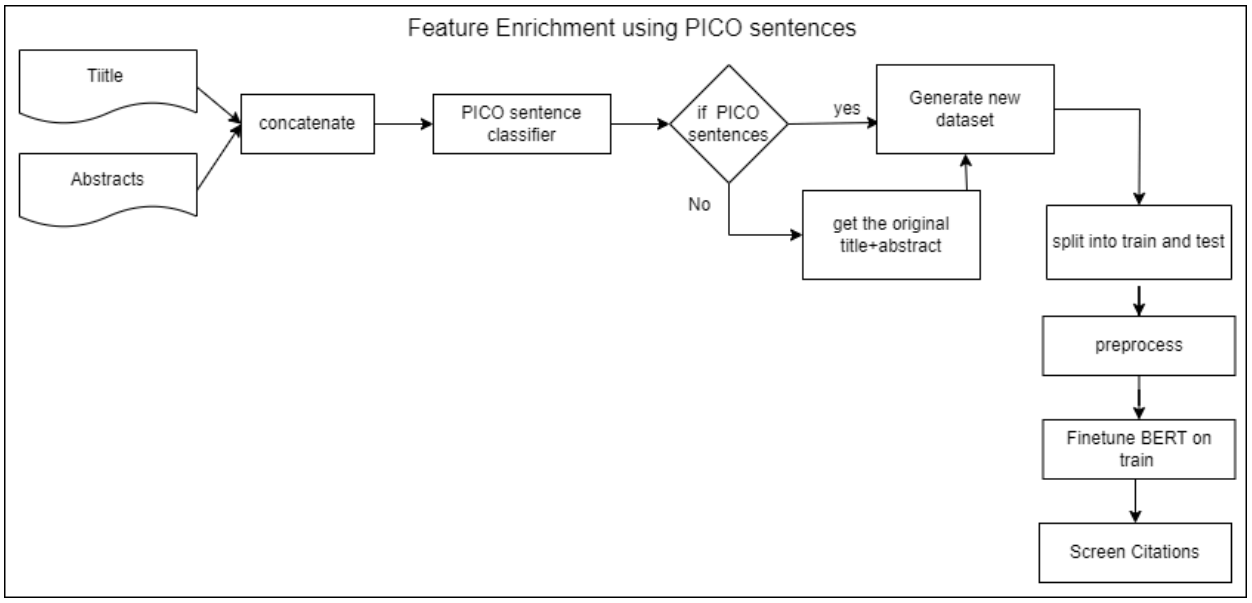


*Figure 13: Feature Enrichment using PICO sentences in citation screening*

One of the challenges of using the PICO sentences framework is that there are some abstracts that do not contain a PICO sentence; in those circumstances, we must utilize the real abstract in the case of PICO sentences for model learning. Using a PICO phrase instead of a full extract is analogous to identifying relevant features for a traditional machine learning system. On the PICO collected text, we do the same preprocessing as in the previous scenario, removing special characters and turning everything to lowercase.

## 4.4. Experiment 4: transformer + CNN

To train models for NLP applications, 1D convolution neural networks are often used. The goal behind using a convolution neural network is to provide the model context from the surrounding world. For example, if a phrase has 5 words and the kernel window has a size of 3, the model will examine 3 words at a time. In most cases, we turn the text into vectors or embeddings before employing a 1D convolutional neural network. It is similar to image convolution except that one dimension of the kernel is fixed according to the dimension of the embedding, hence the kernel only travels in one direction from left to right, hence the name 1D convolution neural network. In this experiment, we added a convolution layer to the output of hidden states from the transformer base BERT model. We employed BERT models' hidden states as input to the convolutional layer, which was then followed by a dense layer and a final output layer in this experiment. There are alternative ways to employ

CNN for text categorization, such as a multi-channel convolutional neural network (Wang et al., 2021), which was used for citation filtering in one of the deep learning publications. Aside from this, there is another way to use convolution neural networks in conjunction with the transformer model in which we take input from the last four or more layers rather than the last hidden state, and this is KIM's convolutional neural network. For this study, however, we will just use the BioLinkBERT model, followed by a 1D convolutional neural network.

## 5. Results:

Our experiments and hypotheses have resulted in mixed but interesting results.

- We will first compare the results of citation screening utilising augmentation and PICO sentences to the results obtained using the real title and abstract. The findings for augmentation and PICO sentences do not compare favourably to the usual content derived from the title and abstract. The table below displays the wss@95R scores for the three models. To provide a fair comparison, we utilised the identical train and test datasets with the same seed for all three scenarios.

| **Dataset** | **Augmentation** | **PICO** | **Abstract + title** |
|---|---|---|---|
| Triptans.tsv | -0.03 | **0.030** | 0.0065 |
| OralHypoglycemics.tsv | 0.283 | -0.018 | 0.0055 |
| ACEInhibitors.tsv | 0.806 | 0.683 | 0.836 |

*Table 1:Model comparison for augmentation, PICO and abs+title, wss@95R scores*

The results reveal that in most situations, the abstract and title produce the best results; nevertheless, there is a case where the PICO sentences produce better results than the abstract and title. Language translation augmentation has not yielded satisfactory outcomes in any of the datasets used here.

- Since the top 2 contenders were PICO sentences and abstract and title combination so further, we have used these to get results on all the possible datasets. The result shows that the combination of abstract and title is still better overall as compared to the PICO sentence with an average WSS@95 score of 41 and 31 % respectively. However, there are certain cases where the PICO sentences perform better than the other, one such example is ProtonPumpInhibitors.tsv dataset, where the score from PICO sentences is 6% higher than the PubMedBERT model . Also in some cases the results of the PICO and abstract model are the same for the BioLinkBERT model.

| **datasets** | **PubMedBERT (abs+title)** | **PICO BioLinkBERT** | **BIOLinkBert** |
|---|---|---|---|
| | | | |
| Opiods.tsv | 0.31 | 0.20 | 0.23 |
| Fluoride.tsv | 0.84 | 0.71 | 0.86 |
| Statins.tsv | 0.41 | 0.23 | 0.24 |
| ACEInhibitors.tsv | 0.69 | 0.49 | 0.68 |
| OralHypoglycemics.tsv | 0.06 | 0.04 | 0.06 |
| AtypicalAntipsychotics.tsv | 0.22 | 0.17 | 0.24 |
| proton_beam.tsv | 0.79 | 0.66 | 0.77 |

| | | | |
|---|---|---|---|
| Triptans.tsv | 0.21 | **0.22** | **0.22** |
| Antihistamines.tsv | 0.16 | 0.10 | 0.17 |
| SkeletalMuscleRelaxants.tsv | 0.31 | 0.18 | 0.18 |
| Estrogens.tsv | 0.29 | 0.25 | 0.29 |
| ProtonPumpInhibitors.tsv | 0.21 | **0.27** | **0.33** |
| ADHD.tsv | 0.54 | 0.22 | 0.32 |
| BPA.tsv | 0.73 | 0.66 | 0.78 |
| **Average** | **0.41** | **0.31** | **0.38** |

*Table 2:comparison of PubMedBERT (abs+title) and PICO BioLinkBERT, wss@95R scores*

- In the experiment to compare finetuning of transformer-based model followed by a dense layer and convolution layers. The results of the transformer followed by a dense layer are slightly better than the ones with the convolution layer. There are a few cases where CNN performs better however overall, the results of CNN are not as good as the dense layers following the Transformers model.

| **datasets** | **PubMedBERT** | **BioLinkBERT CNN** |
|---|---|---|
| Opiods.tsv | 0.31 | **0.39** |
| Fluoride.tsv | 0.84 | 0.83 |
| UrinaryIncontinence.tsv | 0.49 | 0.44 |
| Transgenerational.tsv | 0.46 | 0.31 |
| copd.tsv | 0.64 | **0.67** |
| Statins.tsv | 0.41 | 0.33 |
| ACEInhibitors.tsv | 0.69 | 0.67 |
| micro_nutrients.tsv | 0.70 | 0.65 |
| OralHypoglycemics.tsv | 0.06 | 0.06 |
| AtypicalAntipsychotics.tsv | 0.22 | 0.22 |
| NeuropathicPain.tsv | 0.70 | 0.66 |
| PFOS-PFOA.tsv | 0.85 | 0.80 |
| NSAIDS.tsv | 0.69 | 0.66 |
| proton_beam.tsv | 0.79 | 0.74 |
| Triptans.tsv | 0.21 | 0.20 |
| BetaBlockers.tsv | 0.46 | 0.44 |
| Antihistamines.tsv | 0.16 | 0.16 |
| SkeletalMuscleRelaxants.tsv | 0.31 | 0.23 |
| Estrogens.tsv | 0.29 | 0.25 |
| ProtonPumpInhibitors.tsv | 0.21 | **0.28** |
| ADHD.tsv | 0.54 | 0.37 |
| BPA.tsv | 0.73 | **0.76** |
| CalciumChannelBlockers.tsv | 0.29 | 0.23 |
| **AVERAGE** | **0.48** | **0.45** |

*Table 3:Comparison of PubMed BERT and BioLinkBERT CNN, wss@95R scores*

- The best result that we have got in our study is using finetuned PubMedBERT model. We have compared these results with the existing deep learning studies on the same datasets. There are 2 studies that have been performed previously using deep learning model on the same datasets and a third study that has replicated the results of these 2 studies. We have

compared our results to the original results here and added the full results for comparison including the replicated results in the Appendix. Our model has an average WSS@95 of 48% compared to the best result of 56% with DAFF (Kontonatsios et al., 2020) and 41% of multi-channel CNN (van Dinter et al., 2021). Our finetuned model performs better on bigger datasets, in some cases even better than the DAFF model, for example NeuropathicPain.tsv. When it comes to the transformers models the results of PubMedBERT are better than those of BioLInkBert by 3% on average.

| datasets | DAFF | CNN | PubMedBERT | BioLinkBERT |
|---|---|---|---|---|
| Opiods.tsv | 0.55 | 0.30 | 0.31 | 0.23 |
| Fluoride.tsv | 0.80 | **0.88** | 0.84 | 0.86 |
| UrinaryIncontinence.tsv | 0.53 | 0.27 | 0.49 | **0.55** |
| Transgenerational.tsv | 0.71 | 0.71 | 0.46 | 0.47 |
| copd.tsv | **0.67** | | 0.64 | 0.62 |
| Statins.tsv | **0.57** | 0.44 | 0.41 | 0.24 |
| ACEInhibitors.tsv | **0.79** | 0.78 | 0.69 | 0.68 |
| micro_nutrients.tsv | 0.66 | | **0.70** | 0.64 |
| OralHypoglycemics.tsv | 0.10 | 0.07 | 0.06 | 0.06 |
| AtypicalAntipsychotics.tsv | **0.33** | 0.21 | 0.22 | 0.24 |
| NeuropathicPain.tsv | 0.61 | 0.62 | **0.70** | **0.70** |
| PFOS-PFOA.tsv | 0.85 | 0.07 | **0.85** | 0.76 |
| NSAIDS.tsv | 0.72 | 0.57 | 0.69 | 0.69 |
| proton_beam.tsv | **0.82** | | 0.79 | 0.77 |
| Triptans.tsv | 0.43 | 0.27 | 0.21 | 0.22 |
| BetaBlockers.tsv | **0.59** | 0.50 | 0.46 | 0.39 |
| Antihistamines.tsv | **0.31** | 0.17 | 0.16 | 0.17 |
| SkeletalMuscleRelaxants.tsv | 0.29 | 0.23 | **0.31** | 0.18 |
| Estrogens.tsv | **0.40** | 0.12 | 0.29 | 0.29 |
| ProtonPumpInhibitors.tsv | **0.40** | 0.24 | 0.21 | 0.33 |
| ADHD.tsv | 0.67 | 0.70 | 0.54 | 0.32 |
| BPA.tsv | **0.79** | 0.79 | 0.73 | 0.78 |
| CalciumChannelBlockers.tsv | **0.42** | 0.16 | 0.29 | 0.22 |
| Average | 0.56 | 0.41 | 0.48 | 0.45 |

*Table 4:Comparison with the previous deep learning studies, wss@95R scores*

# 6. Conclusion

- We used augmentation using language translation to balance the dataset for training, however the results reveal that the augmentation is not very effective. We have only attempted augmentation by translating it to one language, German, but there is still a chance that we can test the findings with a new language and see if it performs better. Aside from that, we have only tested it on a few data datasets due to time constraint; it is possible that it will work for certain datasets in the same way that PICO phrases have worked for some datasets.
- The PICO sentence results demonstrate that there is some information in the PICOS that is relevant to the area and may be utilised for the purpose of citation filtering, but PICOs alone are insufficient to make this conclusion according to the results. Another thing to note here is that there are some cases where we couldn't get the PICO sentences at all and we used abstract in those cases, so the right way to do it is to just keep the text that has the PICO

sentences because that way we will be able to compare the actual results of the PICO sentences, right now any positive and negative impact could be due to the original abstracts as well.

- Another problem was that the model we used for PICO sentence categorization was only around 85% accurate. If there was a method to attain 100% accuracy in PICO sentences, the results would be much better. We were unable to realise its full potential.
- We observed that the outcomes of our work are midway between those of a multi-channel convolution neural network and those of denoising autoencoder followed by feedforward network employing SVM classifier in this study. We pointed out that the authors utilised the entire dataset for denoising autoencoder work, including the test set, which causes some contamination from the test set this is the reason why whenever an experiment is performed to learn embeddings or vocabulary in case of classical machine learning its only performed on the train set and the test set is left untouched. But the same could be stated for our study as well because we don't know which abstracts the pretrained models are trained on. Furthermore, the two other researchers used hyper tuning with two datasets to establish the parameters for all the models, which is not a common method across domains. The easiest approach to achieve this is to establish a validation site for all datasets and then hyper tune the models based on it.
- The results of PubMedBert are slightly better than that of BIoLinkBert the possible reason could be that the PubMedBert is explicitly trained on the abstracts and the BioLinkBert is trained on abstracts as well as other parts of the publications as well, which makes it better in general on biomedical NLP tasks. However, PubMedBERT is better for this task of citation screening using abstracts.

## 7. Next steps:

Based on the results and learnings from this study following are the next steps we think should be tried to further improve the citation screening process.

- Because the data is significantly imbalanced, augmentation is still an option. Language translation from languages other than German might be used to supplement the training data. There are other context-based augmentation approaches in which only a few words in a text are changed while maintaining the meaning of the whole phrase intact, we can use nlpaug library for this. Using this for data augmentation might still provide positive outcomes.
- PICOS have demonstrated that there is some meaningful information in the PICO phrases, but it is insufficient to make decisions on its own. One way to use this is to use actual PICO entities in combination with the abstracts.
- In this experiment of PICO sentences we replaced the input text with the abstracts where the model did not identify any PICO sentences. It would be interesting to train the model with only the PICO sentences and no abstracts. This way we would have less training data so this could be tried with a dataset which has enough data points.
- Last but not the least, the researchers filter out the publications using the abstracts because they don’t have enough time to read the entire publication. However, a machine learning model is fast and can go through more than just the abstract and the title. This has not been tried so far according to our knowledge. We can train a model which uses the abstract, results and some more parts of the publications.

# 8. Project Management

Project management is a vital aspect of the study since without it, we might plan a lot but still not accomplish it on time. A realistic strategy with defined objectives in mind can assist an individual in achieving the desired results on time. It is also necessary to have good management in place to keep all the experimental solutions so that we can compare all of the methods and draw conclusions from them.

## 8.1. Project schedule

The first two weeks of the study were spent collecting relevant studies in the area and comprehending the methodologies and obstacles indicated in those. The purpose was to get a better understanding of the problem, which would help with experiment design. The next week was devoted to collecting the 23 publicly available public data sets for systematic investigation, as well as analysing and addressing data-related issues. The next 4 weeks were utilised and developing and assisting the deep learning code base for this experiment. This includes developing the script which can be used for multiple models having Lords and results returned into multiple files to maintain the important information required for the experiment. Following that, it was time to collect the first round of feedback on the completed experiments to improve the findings. The next two weeks were spent implementing the comments from the previous round and designing experiments for PICO sentences and language translation augmentation. These two jobs were requested on the fly and would not necessitate any changes to the code; they are simply an extra step that must be completed prior to training the model. The last 2 weeks were used to evaluate the result of all the previous experiments as well as to design architecture of the transformer model followed by the convolution layers. The code was written in such a manner that by altering the config file, a user may see a new collection of models and also select the argumentation type, whether language or no augmentation. And you can choose whether to utilise PICO phrases or the real abstract by setting a few parameters in the configuration file. They had weekly meetings with the project manager to display the progress report and gain important input. Several of the concepts presented in the research are based on the supervisor's instructions. The plan was to perform the experiment on only five datasets, however because of the project supervisor's large GPUs, I was able to run numerous tests on all datasets.

| | Weeks | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stages | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Data collection | | | | | | | | | | | | |
| Literature Review | | | | | | | | | | | | |
| Model development (finetune) | | | | | | | | | | | | |
| Model evaluation (finetune) | | | | | | | | | | | | |
| Feedback gathering(finetune) | | | | | | | | | | | | |
| Feedback implementation (finetune) | | | | | | | | | | | | |
| Augmentation experiment | | | | | | | | | | | | |
| PICO experiment | | | | | | | | | | | | |
| BERT+CNN experiment | | | | | | | | | | | | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feedback gathering | | | | | | | | | | | | |
| Feedback implementation | | | | | | | | | | | | |

*Table 5: project schedule chart*

### 8.2. Risk Management

There are a few risks involved with the project, one of which is that the model we used for training or fine tuning is not suitable for fine tuning on a laptop, even with the GPU. since this model does not fit in the RAM of standard laptops The professor gave me with a larger GPU, which allowed me to do the tests parallely, which would not have been possible without his assistance. I tried utilising Google collab for some of my tests, however the GPU time there is not guaranteed and can be removed at any time. There were instances where the GPU was available, but it did not last until the completion of the experiment, and I was unable to obtain data from the Google collab or Kaggle GPUs.

Another problem with the study is that the dataset is too small to run all the methods, therefore we chose the best feasible model utilising the transformer architecture, which includes pre-trained information for the biomedical NLP domain.

### 8.3. Quality Management

We choose WSS@95% as the assessment criterion to verify how much of the workload is being cut by this technique to test the quality of the project and to check how much you will it is to the researchers in the field of systematic review. On average, this strategy may save 45% of the time

## 9. Social, Legal, Ethical and Professional Consideration

All of the data utilised in this study is freely available for systematic research and does not include any personal information. The dataset has already been utilised in other research, and it is the appropriate dataset to serve as a benchmark for citation screening. To verify the right use of the data, an ethical certificate has been created. All experiments that result in the study adhere to social, legal, and professional considerations without infringing any legislation in any way.

## 10. Critical Appraisal

In the study, we employed several transformer-based models with pretrained domain knowledge to demonstrate how they may be used for citation filtering. Along with this, we attempted to solve the issues of data imbalance and a lack of data in the training set by employing the data augmentation strategy. We also assessed the influence of PICO sentences, which have shown to be highly helpful criterion for determining the quality of abstracts and are also utilised as one of the popular methods by researchers to filter out articles for citation screening. The effort saved over sample that is relevant to the job at hand and has been utilised in past research makes it appropriate for comparison with the previous study findings. We do not offer any data sets from testing samples or directly provide any contamination during the training process; however, we are unaware of the dataset used to train the pretrained models.

## 11. Achievements

We have obtained results that are halfway in between the previous two deep learning techniques that have been attempted on the same dataset by using the knowledge of Transformers pretrained

models. We investigated two distinct NLP models that are recognised to produce the best outcomes in biological and NLP activities. We may infer that the PICO phrases are very important, but they are not sufficient for the purpose of citation filtering. We attempted optimization through language translation, but there are other methods for augmentation that may be studied, such as utilising a different language for using a context-based model to modify specific terms in the text. However, this analysis shows the problem that is present in the dataset and one possible solution. We tried more sophisticated models with convolutions after the transformer model, but the results were not as excellent, indicating that the transformer model alone is not sufficient to manage the dependencies, and just a few dense layers following that could be sufficient for any experiment. During the experimentation phase, we went through some systematic studies and discovered that it is difficult to generalise systematic studies citation screening. The reason for this is that the questions answered during a systematic study differ from one systematic study to the next, so a model from one systematic study cannot be used in a different systematic study due to the different requirements.

## 12.Student Reflections

This project taught me that a project based on real-world data has quite different obstacles than a toy data collection. It assisted me in understanding alternative approaches to dealing with lesser datasets and what kind of deep learning method may be used when there is minimal data for a domain.

This project also assisted me in developing a comprehensive project plan from start to finish, as well as how each step of the project is performed in accordance with a timetable. It also taught me that when it comes to deep learning frameworks, we need to be adaptable, as one of the frameworks does not support the models we were seeking for. Finally, while simply having a record and a thorough knowledge would not get you good results for an experiment, we do require a certain amount of hardware to actually execute these algorithms. This project has helped me learn how Transformers may be utilised for any underlying problem, as well as how to select the best pre-trained model for every task. technically I learnt the usage of libraries which are relevant todays data science work like the machine learning framework usage of which is very important to this field.

# References

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using Automated Citation Classification. Journal of the American Medical Informatics Association, 13(2), 206–219. https://doi.org/10.1197/jamia.m1929

Kontonatsios, G., Spencer, S., Matthew, P., & Korkontzelos, I. (2020). Using a neural network-based feature extraction method to facilitate citation screening for Systematic Reviews. Expert Systems with Applications: X, 6, 100030. https://doi.org/10.1016/j.eswax.2020.100030

van Dinter, R., Catal, C., & Tekinerdogan, B. (2021). A multi-channel convolutional neural network approach to automate the citation screening process. Applied Soft Computing, 112, 107765. https://doi.org/10.1016/j.asoc.2021.107765

Kusa, W., Hanbury, A., & Knoth, P. (2022). Automation of citation screening for systematic literature reviews using neural networks: A replicability study. Lecture Notes in Computer Science, 584–598. https://doi.org/10.1007/978-3-030-99736-6_39

Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., & Fung, P. (2021). Language models are few-shot multilingual learners. Proceedings of the 1st Workshop on Multilingual Representation Learning. https://doi.org/10.18653/v1/2021.mrl-1.1

Wang, Q., Liao, J., Lapata, M., & Macleod, M. (2021). PICO entity extraction for preclinical animal literature. https://doi.org/10.21203/rs.3.rs-1008099/v1

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., Macleod, M., Shah, R. R., & Thayer, K. (2016). Swift-review: A text-mining workbench for systematic review. Systematic Reviews, 5(1). https://doi.org/10.1186/s13643-016-0263-z

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, 3(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

D'Ambrosio, A., Grundmann, H., & Donker, T. (2022, February 22). An open-source integrated framework for the automation of Citation Collection and screening in Systematic Reviews. arXiv.org. Retrieved December 7, 2022, from https://arxiv.org/abs/2202.10033

Ioannidis, A. (2021, April 16). An analysis of a BERT deep learning strategy on a technology assisted review task. arXiv.org. Retrieved December 7, 2022, from https://arxiv.org/abs/2104.08340

Medical College of Wisconsin. (2022, August 4). Evidence based medicine: Pico. LibGuides. Retrieved October 28, 2022, from https://mcw.libguides.com/EBM/PICO

McCormick, C. (2019, July 22). Bert fine-tuning tutorial with pytorch. BERT Fine-Tuning Tutorial with PyTorch · Chris McCormick. Retrieved December 8, 2022, from https://mccormickml.com/2019/07/22/BERT-fine-tuning/

Wang, Q. (2021, May 10). Preclinical pico extraction. OSF. Retrieved November 8, 2022, from https://osf.io/2dqcg/

# APPENDIX:

1. Comparison of PICO and (abstract+ title)

| datasets | PubMedBERT (abs+title) | pico biolikBERT |
|---|---|---|
| Opiods.tsv | 0.31 | 0.20 |
| Fluoride.tsv | 0.84 | 0.71 |
| Statins.tsv | 0.41 | 0.23 |
| ACEInhibitors.tsv | 0.69 | 0.49 |
| OralHypoglycemics.tsv | 0.06 | 0.04 |
| AtypicalAntipsychotics.tsv | 0.22 | 0.17 |
| proton_beam.tsv | 0.79 | 0.66 |
| Triptans.tsv | 0.21 | 0.22 |
| Antihistamines.tsv | 0.16 | 0.10 |
| SkeletalMuscleRelaxants.tsv | 0.31 | 0.18 |
| Estrogens.tsv | 0.29 | 0.25 |
| ProtonPumpInhibitors.tsv | 0.21 | 0.27 |
| ADHD.tsv | 0.54 | 0.22 |
| BPA.tsv | 0.73 | 0.66 |
| **Average** | **0.41** | **0.31** |

2. Comparison of PUBMEDBERT and BIOLINKBERT+CNN

| datasets | PubMedBERT | BIOLINKBERT+CNN |
|---|---|---|
| Opiods.tsv | 0.31 | 0.39 |
| Fluoride.tsv | 0.84 | 0.83 |
| UrinaryIncontinence.tsv | 0.49 | 0.44 |
| Transgenerational.tsv | 0.46 | 0.31 |

| | | |
|---|---|---|
| copd.tsv | 0.64 | 0.67 |
| Statins.tsv | 0.41 | 0.33 |
| ACEInhibitors.tsv | 0.69 | 0.67 |
| micro_nutrients.tsv | 0.70 | 0.65 |
| OralHypoglycemics.tsv | 0.06 | 0.06 |
| AtypicalAntipsychotics.tsv | 0.22 | 0.22 |
| NeuropathicPain.tsv | 0.70 | 0.66 |
| PFOS-PFOA.tsv | 0.85 | 0.80 |
| NSAIDS.tsv | 0.69 | 0.66 |
| proton_beam.tsv | 0.79 | 0.74 |
| Triptans.tsv | 0.21 | 0.20 |
| BetaBlockers.tsv | 0.46 | 0.44 |
| Antihistamines.tsv | 0.16 | 0.16 |
| SkeletalMuscleRelaxants.tsv | 0.31 | 0.23 |
| Estrogens.tsv | 0.29 | 0.25 |
| ProtonPumpInhibitors.tsv | 0.21 | 0.28 |
| ADHD.tsv | 0.54 | 0.37 |
| BPA.tsv | 0.73 | 0.76 |
| CalciumChannelBlockers.tsv | 0.29 | 0.23 |
| **AVERAGE** | **0.48** | **0.45** |

3. Full result comparison with previous studies

| datasets | DAFF | DAFF replicated | cnn | cnn repl | fast text class | PubMed BERT | Biolink bert |
|---|---|---|---|---|---|---|---|
| Opiods.tsv | 0.553 | **0.580** | 0.295 | 0.249 | 0.559 | 0.309 | 0.227 |
| Fluoride.tsv | 0.799 | 0.806 | **0.883** | 0.808 | 0.390 | 0.837 | 0.861 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| UrinaryIncontinence.tsv | 0.531 | 0.483 | 0.272 | 0.180 | 0.439 | 0.487 | **0.552** |
| Transgenerational.tsv | 0.707 | **0.718** | 0.708 | 0.000 | 0.368 | 0.461 | 0.474 |
| copd.tsv | **0.666** | 0.665 | | 0.128 | 0.312 | 0.640 | 0.624 |
| Statins.tsv | **0.566** | 0.487 | 0.443 | 0.283 | 0.409 | 0.409 | 0.239 |
| ACEInhibitors.tsv | **0.787** | 0.785 | 0.783 | 0.367 | 0.783 | 0.691 | 0.681 |
| micro_nutrients.tsv | 0.662 | 0.663 | | 0.199 | 0.608 | **0.703** | 0.642 |
| OralHypoglycemics.tsv | 0.095 | **0.123** | 0.065 | 0.013 | 0.098 | 0.061 | 0.058 |
| AtypicalAntipsychotics.tsv | **0.329** | 0.190 | 0.212 | 0.081 | 0.218 | 0.218 | 0.241 |
| NeuropathicPain.tsv | 0.608 | 0.598 | 0.620 | 0.091 | 0.613 | 0.698 | **0.703** |
| PFOS-PFOA.tsv | 0.848 | 0.838 | 0.071 | 0.305 | 0.779 | **0.854** | 0.763 |
| NSAIDS.tsv | 0.723 | **0.735** | 0.571 | 0.601 | 0.620 | 0.690 | 0.689 |
| proton_beam.tsv | **0.816** | 0.812 | | 0.357 | 0.733 | 0.794 | 0.765 |
| Triptans.tsv | 0.434 | 0.412 | 0.266 | **0.440** | 0.210 | 0.210 | 0.218 |
| BetaBlockers.tsv | **0.587** | 0.462 | 0.504 | 0.339 | 0.419 | 0.460 | 0.388 |
| Antihistamines.tsv | **0.310** | 0.275 | 0.168 | 0.135 | 0.047 | 0.162 | 0.171 |
| SkeletalMuscleRelaxants.tsv | 0.286 | 0.286 | 0.229 | 0.330 | 0.090 | **0.310** | 0.185 |
| Estrogens.tsv | 0.397 | **0.369** | 0.119 | 0.083 | 0.306 | 0.290 | 0.286 |
| ProtonPumpInhibitors.tsv | **0.400** | 0.229 | 0.243 | 0.129 | 0.283 | 0.208 | 0.329 |
| ADHD.tsv | 0.665 | 0.639 | 0.698 | **0.704** | 0.424 | 0.541 | 0.325 |
| BPA.tsv | **0.793** | 0.780 | **0.792** | 0.369 | 0.637 | 0.734 | 0.778 |
| CalciumChannelBlockers.tsv | **0.424** | 0.347 | 0.159 | 0.069 | 0.178 | 0.286 | 0.225 |
| Average | 0.565 | 0.534 | 0.405 | 0.272 | 0.414 | 0.480 | 0.453 |

4. Our best model with the previous studies

| datasets | DAFF | DAFF replicated | cnn | cnn replicated | our result (biomednlp) | fasttext |
|---|---|---|---|---|---|---|
| Opiods.tsv | 0.55 | **0.58** | 0.30 | 0.25 | 0.31 | 0.56 |
| Fluoride.tsv | 0.80 | 0.81 | **0.88** | 0.81 | 0.84 | 0.39 |
| UrinaryIncontinence.tsv | **0.53** | 0.48 | 0.27 | 0.18 | 0.49 | 0.44 |
| Transgenerational.tsv | 0.71 | **0.72** | 0.71 | 0.00 | 0.46 | 0.37 |
| copd.tsv | **0.67** | 0.67 | | 0.13 | 0.64 | 0.31 |
| Statins.tsv | **0.57** | 0.49 | 0.44 | 0.28 | 0.41 | 0.41 |
| ACEInhibitors.tsv | **0.79** | 0.79 | 0.78 | 0.37 | 0.69 | 0.78 |
| micro_nutrients.tsv | 0.66 | 0.66 | | 0.20 | **0.70** | 0.61 |
| OralHypoglycemics.tsv | 0.10 | **0.12** | 0.07 | 0.01 | 0.06 | 0.10 |
| AtypicalAntipsychotics.tsv | **0.33** | 0.19 | 0.21 | 0.08 | 0.22 | 0.22 |
| NeuropathicPain.tsv | 0.61 | 0.60 | 0.62 | 0.09 | **0.70** | 0.61 |
| PFOS-PFOA.tsv | 0.85 | 0.84 | 0.07 | 0.31 | **0.85** | 0.78 |
| NSAIDS.tsv | 0.72 | **0.74** | 0.57 | 0.60 | 0.69 | 0.62 |
| proton_beam.tsv | **0.82** | 0.81 | | 0.36 | 0.79 | 0.73 |
| Triptans.tsv | 0.43 | 0.41 | 0.27 | **0.44** | 0.21 | 0.21 |
| BetaBlockers.tsv | **0.59** | 0.46 | 0.50 | 0.34 | 0.46 | 0.42 |
| Antihistamines.tsv | **0.31** | 0.28 | 0.17 | 0.14 | 0.16 | 0.05 |
| SkeletalMuscleRelaxants.tsv | 0.29 | 0.29 | 0.23 | **0.33** | 0.31 | 0.09 |
| Estrogens.tsv | **0.40** | 0.37 | 0.12 | 0.08 | 0.29 | 0.31 |
| ProtonPumpInhibitors.tsv | **0.40** | 0.23 | 0.24 | 0.13 | 0.21 | 0.28 |
| ADHD.tsv | 0.67 | 0.64 | 0.70 | **0.70** | 0.54 | 0.42 |
| BPA.tsv | **0.79** | 0.78 | 0.79 | 0.37 | 0.73 | 0.64 |
| CalciumChannelBlockers.tsv | **0.42** | 0.35 | 0.16 | 0.07 | 0.29 | 0.18 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Average | 0.56 | 0.53 | 0.41 | 0.27 | 0.48 | 0.41 |