Continuous Active Learning with Systematic Reviews in Medicine

Aaron HA Fletcher
School of Computer Science
Sheffield
ahafletcher1@sheffield.ac.uk

Acronym	Full Form
SR	Systematic Review
CAL	Continuous Active Learning
AL	Active Learning
TAR	Technology-Assisted Review
EBM	Evidence-Based Medicine
DTA	Diagnostic Test Accuracy
WSS	Work Saved over Sampling
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine
BMI	Base Model Implementation
BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
RCT	Randomised Controlled Trial
OCEBM	Oxford Centre for Evidence-Based Medicine
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
MeSH	Medical Subject Headings
SAL	Simple Active Learning
SPL	Simple Passive Learning

Table 1: List of Acronyms in Systematic Review Research

Abstract − Continuous Active Learning (CAL) has emerged as a promising technique to enhance the efficiency and accuracy of systematic reviews in medicine. This PhD proposal investigates the application of CAL, specifically focusing on the title and abstract screening substage of systematic reviews. The primary goal is to minimize expert intervention while maintaining high accuracy in document classification, thereby addressing the increasing volume of research and limited resources in healthcare. The PhD research will focus exclusively on a pool-based sampling approach within active learning. The PhD will use datasets such as CLEF-TAR and the Synergy Dataset, which provide real-world scenarios and imbalances typical of systematic reviews. Research gaps to addressed are the lack of long context model integration, such as longformer and big bert into CAL, the limited use of metadata in the title and abstract screening process and utilising encoders as document representations.

Keywords - Systematic Reviews, Continous Active Learning, Technology-Assisted Review, Medical Literature Screening, Evidence-based medicine, BERT, Metadata analysis

Contents

I INTRODUCTION

This document presents a literature review and planned research questions for the author's Ph.D. proposal. The Ph.D. proposal focusses on enhancing the performance of title and abstract selection through the application of continuous active learning in systematic reviews.

The proposal starts by motivating the need for research in this area, highlighting key stages of the systematic review process and the challenges faced. Then it outlines data sets and existing previous work with commentary on what the author feels are limitations of this research body.

The core research questions centre around how documents are represented within the CAL process and are as follows:

1. Introduce novel stopping algorithms for TAR
2. Investigate whether additional metadata, such as citation networks, can enhance model performance.
3. Evaluating the potential benefits of using decoders as embedders rather than encoders in CAL performance.

It also outlines a timeline, potential risks and mitigation strategies, ethical considerations and how the requirements for a professional development plan have been met. The author also mentions other publishable research projects with which he is involved and their potential impact on the Ph.D.

II BACKGROUND LITERATURE

A Systematic Reviews

A systematic review (SR) is one approach, among many, to provide evidence that can be used to support clinical decisions [1]. Evidence-based medicine attempts to incorporate this into clinical practice, by recommending the preferential use of the strongest available evidence in guiding decision making. EBM ranks each approach to support decision making, sometimes referred to as a hierarchy of evidence, where SRs are deemed to provide the strongest evidence to support any clinical decision - for relative rankings of evidence strength, see Table 2.

Level	Type of Evidence
1a	Systematic reviews of randomized controlled trials
1b	Individual randomized controlled trials
2a	Systematic reviews of cohort studies
2b	Individual cohort studies
3a	Systematic reviews of case-control studies
3b	Individual case-control studies
4	Case series
5	Expert opinion

Table 2: A summarised form of the 2009 OCEBM Levels of Evidence [2] There are conflicting thoughts on the absolute ranking of strength for all evidence sources [3, 4]; however, a key commonality is that systematic reviews (of RCTs) are considered the strongest evidence type.

SRs use reproducible systematic methodology to collect existing research, critically assess each study, and synthesise the findings into new research, and aim to provide a complete exhaustive summary of current evidence related to the research question [5].

The need to improve efficiency in SRs is based on two main areas: the increasing volume of research and the resources available within healthcare care. It is known that the amount of research available to be included in these SRs is increasing; with an estimated number of peer-reviewed journals in 2020 being 46,739 (from 14, 694 in 2001), and the total number of articles published increasing threefold, and the total number of clinical research trials increasing twofold - see Figure 2 [6]. However, the merits and disadvantages of the SR process are outside the scope of this Ph.D., but it has been succinctly conveyed in the existing literature [7], and, notwithstanding, SRs represent the best approach available to providing EBM.

A.1 The systematic review process

To understand how we might aim to improve SRs, we first need to outline the stages through which a SR progresses. Typically, the process is broken down into five distinct phases, as outlined in Table 3. My PhD will focus on stage 2: Identifying relevant work, which can be further granularised into several substages:

- Inclusion/exclusion criteria generation

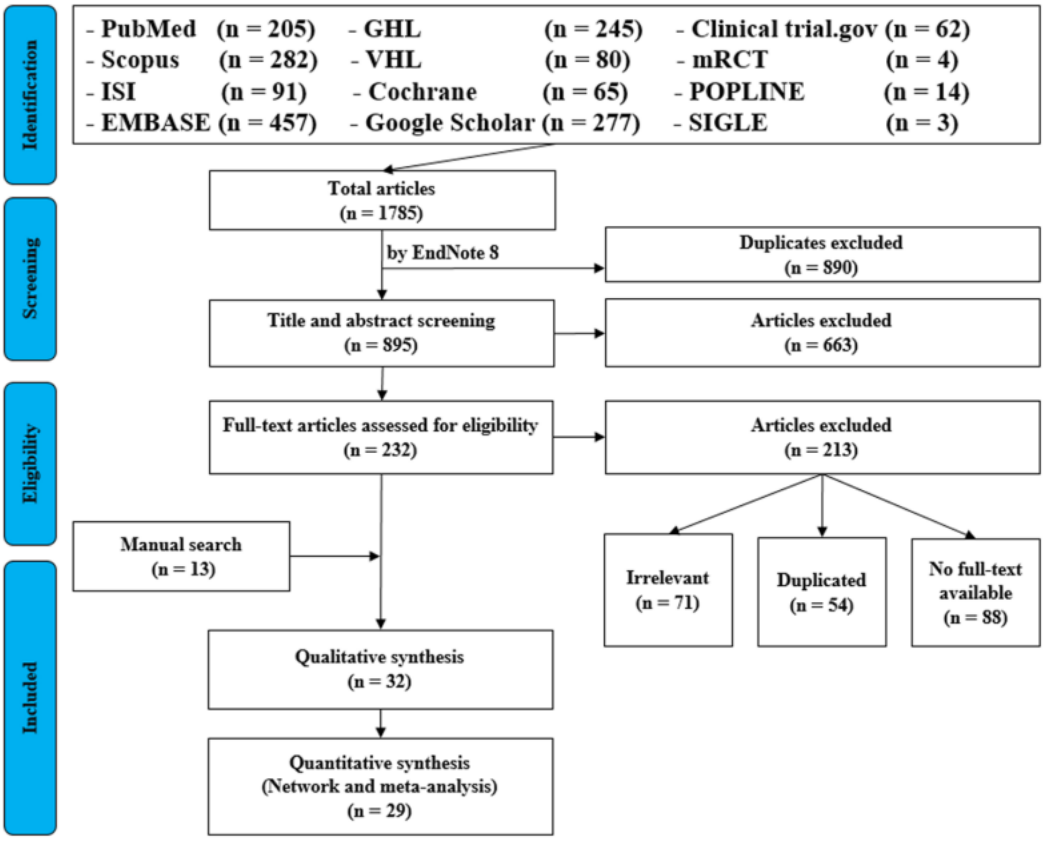


Figure 1: PRISMA flow diagram of studies' selection and screening process: Copied from [8]

- Search strategy development
- Database searching
- Protocol writing
- Title and abstract screening
- Full-text download and screening
- Manual search

Figure 1 illustrates these substages. Specifically, this PhD will concentrate solely on the title and abstract screening substage of the "identifying relevant work" phase. At this point in the identification of works process, preliminary work has been identified through a Boolean search, providing a large list of potentially included research. Traditionally, the titles and abstracts of these works are then manually evaluated by 2-3 reviewers to decide whether they should be included or excluded based on predetermined criteria, reducing the additional work that occurs within the full text download and screening substage. This process is similar to information retrieval.

A.2 Efficiency within the title and abstract Screening process

Abstract screening averages 0.13-2.88 abstracts per minute [**felizardo˙visual˙2013**, 9, 10]. Conflict resolution, which is often necessary when multiple reviewers are used (which is preferred), takes on average 5 minutes. Screening

Stage	Purpose
1	**Framing questions for a review:** The research question is structured and explicitly formulated.
2	**Identifying relevant work:** A wide range of databases are searched to identify research to be included. Potential research is first identified, screened, eligibility checked, and then a decision is made on the inclusion of that research [8].
3	**Assessing the quality of studies:** Research is tested for quality, such as minimum research design, and subjected to higher quality assessment checks, including tests for research heterogeneity.
4	**Summarizing the evidence:** Data synthesis occurs with tabulation of study characteristics and quality. Statistical testing is performed at this stage.
5	**Interpreting the findings:** Any issues highlighted in the previous steps should be addressed. Generate recommendations guided by reference to the strength of the evidence.

Table 3: Stages of a Systematic Review

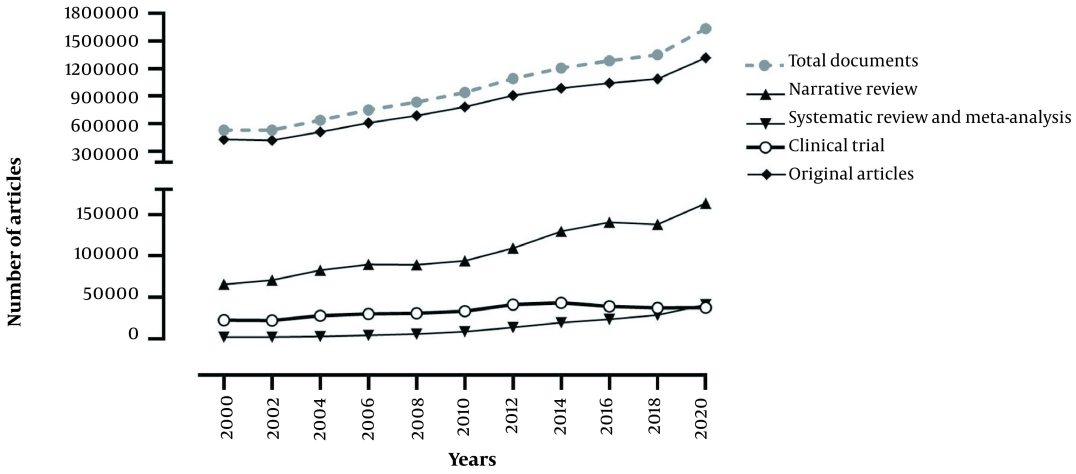


Figure 2: Increasing publications over that past two decades [6]

a full-text article takes 4 minutes on average [9]. Given a recently released Cochrane SR on the use of preoperative statin therapy in adults undergoing cardiac surgery, we can estimate that the total time to review all the text and abstracts for this would take 7.43 hours at best and 164.64 hours at worst [11]. Factored into an inflation adjusted average research cost per minute (£1.598), expected costs **for just this substage alone** could be expected to be £721.97 to £15,785.68 [12].

B Active Learning

Active Learning (AL) presents a promising approach to address one of the most resource-intensive stages of the SR: title and abstract screening. In the context of SRs, where the volume of potentially relevant literature is vast and growing, AL offers a method to significantly reduce the manual workload while maintaining high accuracy in document selection. Traditional systematic review methods require domain experts to manually screen all titles and abstracts identified in the initial search phase. This process is time-consuming and costly, especially given the increasing volume of published research. AL aims to optimise this process by intelligently selecting which documents should be reviewed by human experts, potentially saving significant time and resources. By applying AL techniques to the screening process, we can:

- Prioritise potentially relevant documents for expert review
- Reduce the overall number of documents that require manual screening
- Potentially identify relevant documents that might be missed in manual screening due to human fatigue or error
- Accelerate the overall SR process without compromising on quality

Deep-learning models have traditionally relied on large-labelled datasets for training. However, this approach contrasts sharply with real-world scenarios, particularly in specialised domains such as medicine. Although data collection is relatively straightforward in these fields, labelling is often time-consuming and requires expert knowledge [13, 14]. This disparity presents a significant challenge to optimise model performance with a limited number of labelled examples. This challenge is particularly relevant to the screening process in SRs, where we currently ask experts to screen all titles and abstracts returned from the identification phase. However, we would want to move to a scenario where minimal expert screening is sought from research returned from the identification phase, whose screening can then be safely extrapolated to a larger pool.

Active learning (AL) studies how to do just this. Through AL terms, it attempts to use a sampling policy π to select samples $\mathbf{T}C, i$ from an unlabelled dataset $\mathbf{T}U, i$ and pass them to an oracle for labelling and added to a known dataset $\mathbf{T}_{K,i}$. Technology-Assisted Review (TAR), also known as Computer-Assisted Review or Predictive Coding, is a process that uses machine learning to assist in document review tasks. In the context of SRs, the TAR applies AL principles to streamline the selection process of titles and abstracts to screen. By iteratively training a machine learning model on human-labelled examples, TAR can prioritise potentially relevant documents for expert review, significantly reducing manual workload and, in some cases, exceeding human ability [15]. This approach aligns closely with the goals of AL in SRs, as it aims to maximise the efficiency of expert input while maintaining high accuracy in document classification.

Different approaches can be taken to AL, such as membership query synthesis [16], stream-based selective [17] and pool-based sampling [18]. These approaches are divided on how much of the unlabelled dataset ($\mathbf{T}U, i$) that a model has access to when utilising a policy π for the selection of data points to be labelled $\mathbf{T}C, i$. In pool-based sampling, the entire unlabelled dataset, $\mathbf{T}U, i$, is evaluated in the selection of $\mathbf{T}C, i$, in stream-based selective sampling, datapoints are evaluated one at a time, and in membership query synthesis, synthetic data are generated from an underlying natural distribution. As we are concerned solely with the subprocess of title and abstract screening, it aligns strongly with pool-based sampling as the entire unlabelled dataset is known ahead of time and will be the approach used within this PhD. Figure 3 outlines the AL cycle with pool-based querying.

The oracle (O) within this PhD will denote a human-verified label resulting from screening potential research for inclusion within the SR within the title and abstract screening stage. Domain experts typically perform this. More

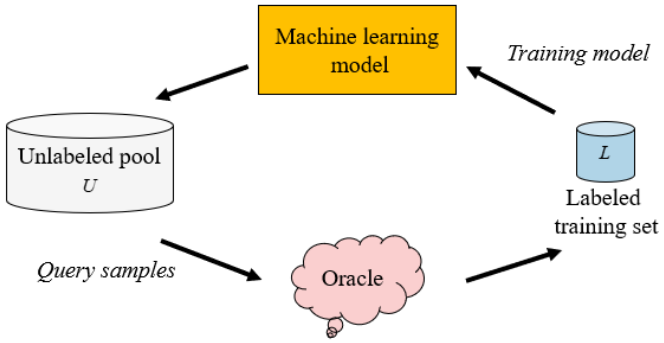


Figure 3: Overview of a pool-based query strategy for AL, replicated[19]

.

Notation	Explanation	Notes
$\mathbf{T}$	Total dataset	e.g. Research gathered after Identification phase of the selection process.
i	Iteration	A single cycle within the active learning process.
$\mathbf{T}_{K,i}$	Known datapoints per iteration	e.g. research that has been screened by a reviewer
$\mathbf{T}_{U,i}$	Unknown datapoints per iteration	e.g. research that has not been screened by a reviewer
$\mathbf{T}_{C,i}$	A subset of $\mathbf{T}_{U,i}$ to be labelled	chosen by a policy, datapoints to be screened by a reviewer.
π	Policy	How $\mathbf{T}_{U,i}$ is selected, e.g. uncertainty, random, certainty, diversity sampling
O	Oracle	Often a domain expert, who assigns labels to unscreened research.
T_R	Total Relevant Documents	All research that should be included in a systematic review.
T_{IR}	Total Irrelevant Documents	All research that should not be included in a systematic review.

Table 4: Notation used for active learning within this review

concretely, O can be considered a function $O(x) = y$ where X is a representation, such as embedding the research title and abstract, and y is the assigned category (included or excluded). We assumed that for each datapoint (x), O provides a single judgement (y), which is always correct and do not concern ourselves with any potential intercoder agreement or bias within that decision process [20].

The broader literature has evaluated the effects of the choice of π. Traditionally, π used the sigmoid response of the final layer of a model as a proxy of confidence, which is not a reliable measure, as these responses tend to be overly confident [21]. The use of the softmax response has been shown, in some cases, to be worse than random sampling [22]. However, the effect of the choice of π in combination with the final output layer in this specific domain (i.e. the SR process) is unknown, and this remains an active area of research which will not be covered during the PhD. The author uses techniques such as temperature scaling to effectively combat overly confident soft-max responses [23].

In the author's mind, it is unclear the exact difference between continuous/online active learning (CAL) and AL, and indeed, it seems that much of the current literature refers to CAL when, in fact, it means AL. Some authors refer to eliminating models between iterations and the process occurring in discrete rounds as AL [24]. Cormack differentiates the two based on objectives, with the aim of CAL being to find and review as many of the responsive documents as possible, as quickly as possible, and AL is to produce the best classier possible, considering the level of training effort (which are subtlely different objectives) [25]. For this research, we will use continuous to denote the incremental streaming of newly available information to any model.

B.1 Active Learning Key Literature

There exists a plethora of research within the AL area; however, due to the specific focus on medicine within this PhD, I will focus on existing literature within the medical domain and some key ones from others. It is valid to delineate between research within differing domains within AL (e.g., e-discovery in the Legal Domain, sentiment

analysis on social media, or image classification in computer vision) as each domain presents unique challenges and characteristics that influence the application of AL techniques. In the medical domain, particularly in SRs, AL must contend with highly specialised vocabulary, complex interrelationships between concepts, and the critical importance of high recall to ensure that relevant studies are not missed. The emphasis of the medical domain on evidence-based practice and the potential impact on patient care requires a more stringent approach to AL. Unlike other domains, where missing a small percentage of relevant items might be acceptable in systematic medicine reviews, overlooking a crucial study could have significant consequences. This requirement for near-perfect recall and the need to process large volumes of literature efficiently create a unique set of demands for medical TAR AL algorithms. This literature review will assess each work for their respective contributions, the datasets used, the evaluation metrics (and scores achieved), the models used, and the representation of data points used in each approach.

An early contribution to this field demonstrated that automated classification of document citations can be used to reduce the time reviewers spend screening evidence for inclusion in SRS of drug class efficacy [26]. The researchers used a novel data set created from annotated reference files from 15 SRs of drug classes. The features were extracted from the research articles using the "bag-of-words" approach for the title and abstract, the MeSH terms, and the MEDLINE publication type. The features were one-hot encoded and selected using the chi-square test to drop insignificant features. Finally, this input was used to train a perceptron model and was evaluated using precision, recall, and the F measure in a range of sample weighting and the WSS@95%. This work's significant contribution was using machine learning approaches to address this screening issue.

A significant contribution to the field came from a simulation study that mimics the process of a human reviewer screening records while interacting with an active learning model [27]. This work used six previously labelled SRs. It looked at four classification techniques (naive Bayes, logistic regression, support vector machines, and random forest). Two feature extraction strategies (TF-IDF and doc2vec) were evaluated based on the work saved on sampling and recall. The title and abstracts were used to generate these inputs. It showed that in a simulated approach, the models reduced the number of publications screened from 91.7 to 63. 9% (WSS@95). The naive Bayes and TF-IDF models yielded the best overall results in this study. This study is limited due to the smaller datasets used and the feature extraction approaches used (which, for the year of publication, other potential superior choices, such as contextual embeddings, could have been explored). It introduced some new evaluation metrics, such as TTD and ATD. This research only superficially evaluates the variability of these approaches in SRs, reporting the range of WSS@95 and not attempting to consider factors within SRs that may have led to this variability.

More recent work reported a protocol denoted "CAL" (Continuous Active Learning), initially performed on legal datasets [28]. The process involves selecting an initial set of seed documents, typically using keyword search, which are then reviewed and coded. This training set is used to train a model that scores each document based on its response (relevant) likelihood. The top-scoring documents that have not been coded are then reviewed and coded. The extended training set is then used to retrain the model, and this process continues until "enough" of the responsive documents are found. The key difference between this approach and previous ones, such as SAL (Simple Active Learning) and SPL (Simple Passive Learning), is their selection strategy: CAL uses relevance feedback (selecting highest-scoring documents), SAL uses uncertainty sampling, and SPL uses random selection. CAL achieved better results on the recall@75 The study primarily used SVM (Sofia-ML implementation of Pegasos SVM) for all protocols. It mentions briefly that it replicated most experiments using logistic regression, achieving similar results to SVM. It also tested with Nave Bayes, which achieved generally inferior results overall but maintained the same relative effectiveness among the protocols. This paper is essential to outline the CAL process and demonstrate its effectiveness. However, it had some limitations: It used a fixed batch size of 1,000 documents for efficiency, though they noted slightly better results with a batch size of 100 for CAL. Although multiple classifiers were tested, detailed

performance comparisons between models were not reported. The study did not explore extensively the effects of feature engineering methods. The human factors in review accuracy were not fully addressed in the simulation. Despite these limitations, the article provided a strong foundation for understanding and further investigating TAR protocols in SR TAR.

The follow-up was to introduce the autonomous TAR process (Auto TAR), which showed better performance than CAL [25]. The algorithm is outlined in Figure 4. The AUTO TAR process differs from CAL through:

- AUTO TAR uses a single seed document, and CAL uses a 1,000 document set.
- AUTO TAR uses word-based features TF-IDF, and CAL uses binary byte-4 grammes.
- AUTO TAR exponentially increases batch sizes, starting with one and increasing by 10% each iteration; CAL's batch size is fixed at 1,000

Finally, this approach was augmented to use BM25 (a saturated form of TF-IDF) + logistic regression and has been considered state-of-the-art for the past eight years. This is often referred to as "base model implementation", or abbreviated to BMI.

Since 2016's AutoTar approach, the transformer architecture has advanced virtually all fields within natural language processing, permeating almost every aspect of the field [30]. However, the TAR process has remained surprisingly resistant to these advances. Although a comprehensive analysis of transformer-based improvements is beyond the scope of this literature review, their primary advantage is their ability to provide a nuanced contextual understanding of written texts. This contextual comprehension differs significantly from traditional feature extraction techniques such as TF-IDF. Transformers employ self-attention mechanisms to learn context-dependent text representations, enabling them to capture subtle semantic relationships and long-range dependencies not present with TF-IDF-like approaches, in the TAR process.

Add in about CALBERT

Subsequent research using decoder architectures within CAL attempted to understand why their performance was underwhelming compared to BMI. Goldilocks [31] took a pre-trained BERT model and fine-tuned it first on the unlabelled corpus (0-10 epochs were tested). They then randomly selected a positive example seed, and on each iteration of the AL process 200 documents were sampled using either relevance feedback or uncertainty sampling. During each iteration, after labelling, all labelled documents were used to fine-tune BERT again for 20 epochs, using the previous model from the previous iteration as the starting point. The input was the concatenated title and body text, truncated to 512 tokens. The model was then used to classify a document's relevance. They found that 5-epoch fine-tuning on the unlabelled dataset was "just right", and achieved similar performance to that of TF-IDF & logistic regression within domain dataset (RCV1-V2 corpus), and statistically significantly worse on out-of-domain dataset(Jeb Bush corpus). Although not discussed in the paper, this could be explained through the phenomenon of "catastrophic forgetting", where too much fine-tuning on the task corpus might cause the model to forget useful general knowledge from pertaining [32]. Additionally, they used the model from the previous iteration as a starting point for new classification fine-tuning iterations, which could have further compounded this.

However, the concept of a "goldilocks" epoch for CAL within medical data sets did not seem to hold, with subsequent research demonstrating that a "just right retaining epoch" was not present in the CLEF data set [33].

Issues with Goldilocks:

1. Contrived experimental design.
2. Sensitive to pre-trained model choice.

Background on stopping algorithms

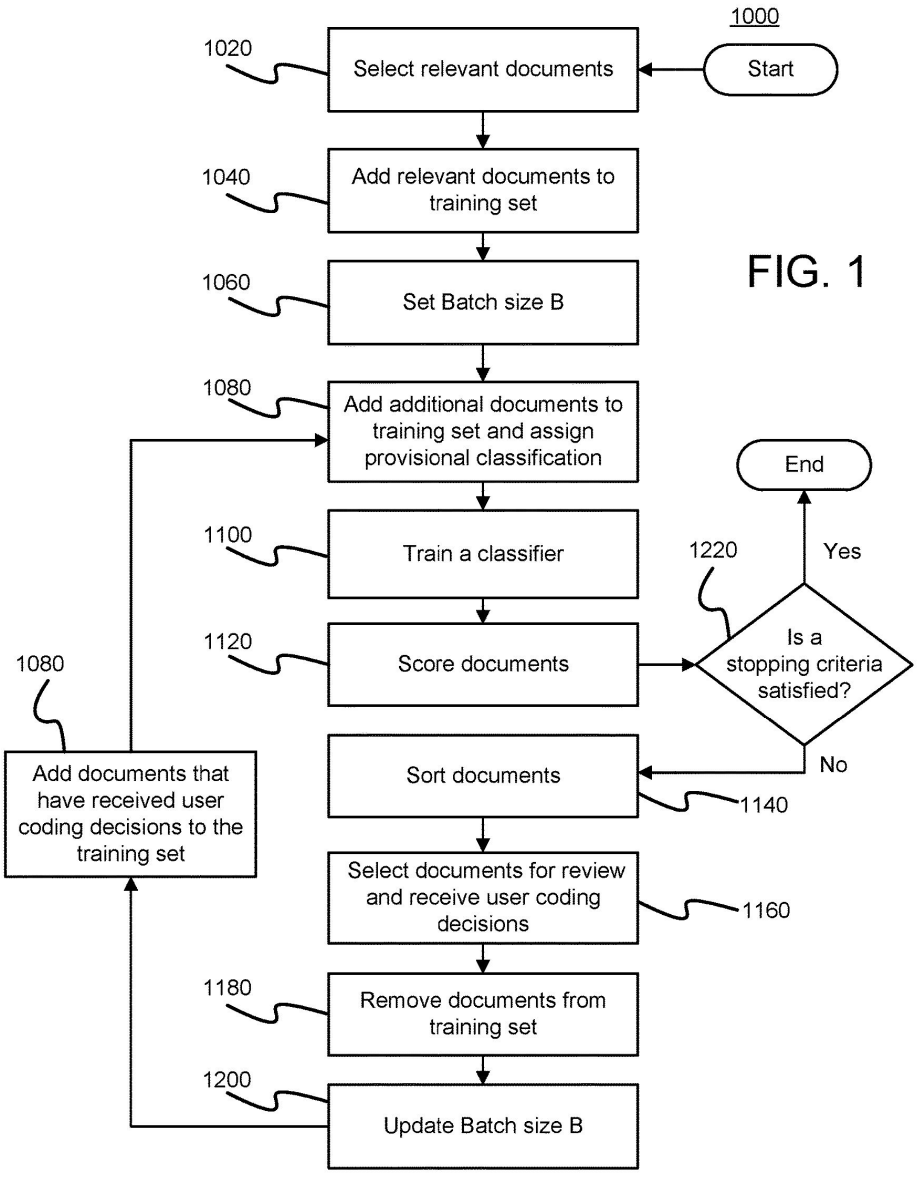


FIG. 1

Figure 4: Auto TAR outline, as documented within the patient filing by Cormack et al. [29]

Dataset	Total SRs	Type(s) of SR	T	TR	TR/T
CLEF 2017	50	DTA	269628	4661	0.017
CLEF 2018	50	DTA	266657	4351	0.016
CLEF 2019	80	DTA	485153	8315	0.017
CLEF 2019	80	Intervention	31644	448	0.014
Synergy	26	Not Applicable	169288	2834	0.017

Table 5: Training Dataset sizes for the TAR datasets

C Datasets

Numerous data sets related to this area have been used in the existing literature.

C.1 CLEF-TAR (2017, 2018, 2019)

CLEF-TAR is a dataset that was released as part of CLEF eTASK 2, and is available on github[1] [44, 45, 46]. Originally designed with document ranking as the primary focus, the information contained within the data set allows for the subprocess simulation of the title and abstract selection of the SR procedure, using published real-world Cochrane SRs. Each year, this data set was incrementally updated and Table 5 outlines the scope of the issue and succinctly highlights the presence of a large imbalance of the TIR class. Diagnostic test accuracy SRs (DTA) summarise a test accuracy, while intervention reviews assess the effectiveness/safety of a treatment, vaccine, device, preventative measure, procedure, or policy. Delineation between the types of SRs is not required for research within this Ph.d.

Of importance, the CLEF dataset did not provide the titles or abstracts for each research found in the Identification Phase, rather relying on the users to download them for experimentation. This is an important oversight of the data set as titles and abstracts can be updated or retracted post-publication, meaning, fair comparison across time might become increasingly challenging. Within this Ph.D. I intend to use this recently collected source 2024 of titles/abstracts that have been collected as part of other work in this area, which has extracted all titles and abstracts for **T** within the CLEF dataset [36][2].

C.2 Synergy Dataset

The Synergy dataset [47], while less frequently used in the literature, offers a more contemporary collection of SRs[3]. This data set comprises 26 SRs that span multiple domains, with a predominant focus on the medical field (20 out of 26 reviews). Reviews included in this data set range from 2002 to 2020, potentially providing more recent information compared to the CLEF data set. The Synergy data set features diverse domains, allowing cross-domain analysis despite its primary focus on medical reviews. It also includes an expanded variable set. In addition to the basic information found in the CLEF dataset, Synergy incorporates authorship details, referenced works, and publication years, all sourced from the OpenAlex API. Due to its more recent compilation and limited use in existing research, this dataset could be used to externally validate pre-trained language models.The inclusion of SRs from nonmedical domains, such as computer science, allows evaluations on the transferability of TAR approaches across different fields. Synergy's TR/T ratio of 0.017 is consistent with the class imbalance observed in the CLEF datasets, making it suitable for comparative studies and model evaluation in the context of title and abstract selection tasks.

[1] https://github.com/CLEF-TAR/tar
[2] https://github.com/ielab/goldilocks-reproduce
[3] https://github.com/asreview/synergy-dataset/tree/master

C.3 TREC Total Recall Track Dataset (2015, 2016)

The TREC Total Recall Track produced data sets specifically designed for high-recall retrieval tasks, similar to those encountered in SRs[4][48, 49]. This data set simulates scenarios where the goal is to find all or nearly all relevant documents in a collection, which aligns closely with the objectives of the title and abstract screening phase in SRs. The data set includes a corpus of documents, topics (which can be seen as analogous to research questions in SRs), and relevance judgments.

C.4 Jeb Bush Emails Dataset

The Jeb Bush Emails dataset is an unconventional choice for TAR research, originally consisting of emails released by former Florida Governor Jeb Bush[5]. This data set is suitable for TAR experiments because of its large size and the presence of both relevant and irrelevant documents. Although not directly related to SRs, it provides a real-world corpus that can be used to simulate document classification tasks inherent in the SR process.

C.5 RCV1-v2 Dataset

The RCV1-v2 (Reuters Corpus Volume 1, Version 2) is a large, manually categorised newswire data set [50] that was published by Reuters between August 20, 1996, and August 19, 1997[6]. The dataset features 804,414 documents with multi-label classification across 103 topic categories, organised in a hierarchy. The documents are provided in XML format with rich metadata and the content is primarily English news stories covering a wide range of topics. Although not originally designed for SRs, RCV1-v2 has been used in various text classification and information retrieval tasks. In the context of SRs and TAR, the use of the RCV1-v2 data set lies in the simulation of approaches on a large-scale dataset to test the scalability and efficiency of screening algorithms and to evaluate any potential transferability of the approaches.

RCV1-v2 dataset is adapted for use in AL by denoting all documents as T, T_R as all documents having a specific label, and those without it, as by treating the entire corpus as T, T_{IR}, we can approximate the binary classification challenge of title and abstract Screening within SRs.

D Evaluation Metrics

Evaluation metrics for SR TAR process can be categorised between assessing how well a classifier minimised the relevant documents excluded by the classifier with a set work budget (i.e. effectiveness) or the reduction in the reviewer's workload by excluding the maximum number of irrelevant documents while maintaining recall (efficiency). The majority of the research produced within this will focus on improving the effectiveness of AL models within the medical TAR domain. The author chooses not to optimise the computational efficiency between approaches, rather to improve the final result achieved. This is for numerous reasons; however, the main two are that as computer processing increases, these practical limitation concerns become less and improvement in effectiveness will have a greater impact on SR usefulness than maximising efficiency. The author aims to report the time taken to run the algorithms, time complexity and the hardware that ran upon, so that comparison to time taken by humans undertaken can occur.

[4]https://trec.nist.gov/data/total-recall/
[5]https://ab21www.s3.amazonaws.com/JebBushEmails-Text.7z
[6]https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/datasets/_rcv1.py

D.1 Recall@k

In the context of SRs, achieving high recall is more critical than high precision. Recall represents the proportion of relevant documents correctly identified among all truly relevant documents [51]. This focus on recall may seem counterintuitive, but it is crucial for two reasons. First, each missed document could potentially contain significant information for the SR. Second, the initial screening is followed by a more precise full-text review (as outlined in the PRISMA workflow, Figure 1), where precision is emphasised. Although maximising recall is important, it is not practical to aim for 100% due to diminishing returns. As recall approaches higher levels, the computational cost of screening additional documents increases substantially, often yielding minimal benefit. To balance effectiveness and efficiency, researchers of TAR for SR commonly use and consider recall @ 95% as useful (that is, k = 0.95). This measure indicates the recovery achieved when 95% relevant documents are recovered, striking a pragmatic balance between comprehensive coverage and resource use. A higher recall@k is considered a more effective approach. Recall is calculated via:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

Recall@95% is calculated via:

$$\text{recall@95\%} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where recall is calculated once an AL classifier achieves 95% TP.

A small point on nomenclature: Historically, recall has been referred to in the medical literature as sensitivity. These are two different terms for the same metric, and the use of either term depends on the domain. Additionally, in the legal domain, there might be references to recall@75% which is not as useful for the medical domain. Legal domains often prioritise based on cost-effectiveness, time-constraints and proportionability, which medical reviews require as close to absolute information as possible [52].

D.2 R-Precision

This effectiveness metric determines, given the T_R, what proportion of documents returned by the approach within the total number of relevant documents were actually relevant [53]. The best score for R-precision is 1 (i.e., all relevant documents were returned in the top T_R position). It allows for an adaptive cutoff for T_R, which adapts to the SR, and also considers precision. Note that this evaluation metric can only be used when the T_R for a query is known. It is calculated via:

$$\text{R-Precision} = \frac{\text{Relevant Documents in top } T_R}{T_R} \tag{3}$$

D.3 Work Saved Over Sampling (WSS@k)

WSS@k is an efficiency metric that would be valuable to report on to enable other researchers in the field to compare their approaches to mine and, if appropriate, improve upon. Again, k (recall) is typically set to 0.95. This metric evaluates the work saved over random sampling, with a higher WSS@k being more efficient and is calculated via:

$$\text{WSS} = \frac{\text{TN} + \text{FN}}{\text{T} - (1 - \text{Recall})} \tag{4}$$

This can also be expressed as:

$$\mathrm{WSS} = \frac{\mathrm{TN} + \mathrm{FN}}{\mathrm{T} - 1 + \frac{\mathrm{TP}}{\mathrm{TP+FN}}} \tag{5}$$

Statistic	Value
Average Percentage of Lost words	21.18%
Average number of words lost	66.94
Maximum percentage of lost words	87.50%

Table 6: Statistics demonstrating limitation of the 512 embedding process on the CLEF dataset, code generated by Author.

[8]

III RESEARCH GAPS

The main "theme" of this PhD is understanding the best approach to embedding information before input into a transformer-based active learning system. This is based on the research gaps identified, such as short context windows given with previous transformer-based TAR research and its strict adherence to faithful replication of how humans screen documents for inclusion into SRs.

Research questions identified through this literature review can be summarised into:

- Does utilising a larger context window, or domain-pretrained models improve BERT performance in the CAL process?
- Does the provision of additional metadata, such as a citation network, improve model performance in CAL?
- Does utilising decoders as embedders rather than encoders result in better CAL performance?

A Enabling fair comparison

So far, the existing literature represents documents as a concatenation of the title and abstract, which is then embedded using models such as 'bert-based-uncased'. A significant limitation is a fixed input of 512 tokens. If we concatenate the title and abstract of the CLEF dataset, this leads to a truncation of the input, as demonstrated in Table 6[7]. If we further consider how abstracts are formatted, with it approximating the logical structure of the content (background, method, results, conclusions), then the likely truncation of the conclusion, which outlines the vital contribution of that paper on the topic, would undoubtedly lead to poorer performance. Compare this to the TF-IDF approaches of Auto TAR, which will still likely contain important missing words from those documents, making the comparison ultimately unfair.

The author will make use of existing approaches that have been developed to deal with a long context, such as LongFormer or Big Bird, which are available with an expanded token input length (up to 4096 tokens).

Longformer is a model that adapts RoBERTa to accept longer token inputs [54]. This was achieved through the introduction of an attention mechanism that grows linearly with the length of the sequence through a sliding window of size w, changing the computational complexity of query generation, the key values from $O(n^2)$ (in original transformer implementation) to $O(nw)$, where n is the length of the sequence and w is the average window size. This resulted in a minor increase in accuracy compared to the original RoBERTAa base. Models have been made available for use, which include a 4096 token window, which can be fine-tuned on datasets (such as all the data returned from the previous screening stage) and optimised for smaller token windows[9]. There are more domain-specific versions of long former available, such as clinical long former [55], which have been further trained in MIMIC 3 clinical notes [10].

[7] https://github.com/afletcher53/bert-clef-diff

[9] https://colab.research.google.com/drive/$1m22nj5A3g-KigoHT1xPgxe_0uYX8HPfA$?usp=sharing

[10] https://huggingface.co/yikuan8/Clinical-Longformer

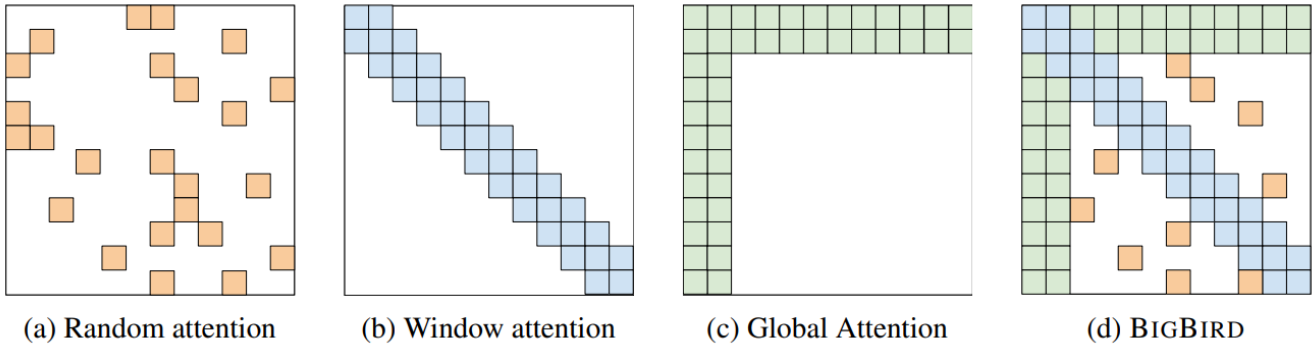


Figure 5: Differing attention mechanisms: Longformer utilises window attention, while Big Bird utilises all 3. Note that the original attention mechanism utilises all squares (including white ones)

Alternatives include the more recently introduced Big Bird [56], which attempts to combine global attention, window attention, and random attention - see Figure 5, and resulted in an improvement of the then state of the art by 5% in document classification.

Additionally, BERT models run so far have used a non-domain-specific pre-trained model, despite them existing (such as BioBERT).

This RQ is deliberately limited in scope. The design of subsequent RQs depends on advances made within RQ1 (as if a longer contextual representation improves the TAR process, it is deemed likely to result in improvements within the other approaches). Additionally, the technical ability gained from undertaking this RQ, such as the utilisation of BERT and replication of experimental approaches, provides a solid basis for further RQs, which will attempt to augment these processes further.

Research needed: To address this research gap, the author proposes replicating previous experiments, including the "Goldilocks reproduce" study, and comparing them with models utilizing larger context windows. This comparison aims to determine whether these extended context approaches yield significant improvements. The research will fine-tune each long-context model using either a) the complete set of document candidates for individual SRs within the datasets or b) the entire corpus of document candidates across all SRs in the datasets. The choice between these approaches will depend on time and computational constraints. This methodology will enable a comprehensive evaluation of the potential benefits of larger context models in SR automation, and hopefully demonstrate a fairer comparison.

B Limited feature useage

The medical TAR AL process aims to emulate human decision-making in SRs. Traditionally, these approaches have mirrored the human workflow: screening titles and abstracts are first, followed by full-text analysis. This sequential method minimises the time-consuming and resource-intensive full-text review phase of human reviewers.

However, computational approaches are not subject to the same constraints as human reviewers. By limiting machine learning algorithms to title and abstract screening, we may inadvertently restrict their potential to perform this task effectively. Consider, for example, the value of analysing author-related information. As demonstrated by the author's previous research, metadata, such as an author's name, can determine if research is valuable. Some authors consistently produce high-quality research within specific domains, making their publications more likely to meet the inclusion criteria for SRs. To further this point, each article will have a list of references. These references

create a citation network where you can trace significant research back to its source, which can be used to judge its significance. By confining AL algorithms to title and abstract data, we exclude potentially crucial information that could inform screening decisions. It is important to note that the intention is not to eliminate the human full-text screening phase of the SR process. Instead, the goal is to use all available information during the initial screening phase to identify papers that warrant full-text review more accurately. This approach could improve the efficiency and precision of the overall SR process, allowing computational methods to use a broader range of data points to make preliminary screening decisions. The TAR process is meant to emulate humans, yet ultimately its goal is to surpass them.

Many tools are available to explore the metadata for medical papers, such as Open Alex [57], which provides granular access to all available metadata, such as citations, references, author metadata, etc.

Different approaches exist to augment information for inclusion within language models such as BERT. BERT was used to process text features from book blurbs and titles, which were then concatenated with metadata features (such as number of authors, academic titles, word counts, and more)[58]. Crucially, the authors also included pre-trained graph embeddings derived from the Wikidata knowledge graph, representing authors and their relationships. This combined representation (text features, metadata, and graph embeddings) was used for classification tasks.

The results demonstrated that the incorporation of metadata features and author embeddings led to better performance for both classification tasks compared to a text-only approach. For a coarse-grained classification task with eight labels, the enriched BERT model achieved an F1-score of 87.20. For a more detailed classification task with 343 labels, the model achieved an F1-score of 64.70. These results outperformed other configurations, including a baseline model using Logistic Regression and TF-IDF vectors.

Additionally, the concept of TwinBERT, which employs two parallel BERT models to separately process text and metadata, has been explored to address the token limit constraints of BERT. This approach showed promising results, particularly in improving recall and precision using textual and contextual information effectively [59]. This RQ is very closely tied to the first one (and could correctly been seen as enabling a fair comparison); however, because it augments the existing title and abstract screening phase, and can potentially build upon any advances found in RQ1, it has been separated.

Research needed: To address this research gap, the author proposes a series of experiments that append metadata from Openalex to the existing data (title/abstract) and then use either TwinBERT, or concatenation of the BERT representation approaches within Active learning. The exact architecture for this is currently undetermined.

C Decoders as text encoder

To date, all approaches utilising LLMs within the TAR process employ encoder architectures, such as BERT, to represent documents. This is logical because encoder architectures are trained in masked language modelling (MLM) and, in some cases, next-sentence prediction, which facilitates effective document representation. In contrast, decoder architectures are trained on next token prediction, enabling them to learn from both preceding and following tokens, thus leveraging information from the entire context rather than a masked subset. Recent studies, such as LLM2Vec[60], have demonstrated that decoder-derived embeddings outperform those from encoder-only models across various tasks, including classification. However, no research in active learning has yet explored the potential of decoder-derived embeddings.

Research Needed: To address this gap, the author proposes selecting a pre-trained decoder-only LLM, such as LLaMA-2-7B, applying the LLM2Vec transformation to it, and using the transformed model to represent documents. Subsequently, the traditional active learning loop can be performed using classifiers such as logistic regression.

D Stretch RQs

As discussed in the subsequent timeline section, time has been allocated to pursue further developments that might arise while undertaking this Ph.D. The stretch RQs have yet to be fully formulated; however, currently the author is exploring:

Using Mixtures of Experts (MOE) in SR TAR process [61]. MOE attempts to use multiple models, each of which specialises in a specific portion of the data (which in information recall we could consider models optimised for true precision, true recall, etc.) and then using a router to decide which expert should handle each input data. Both experts and router are trained simultaneously. This idea could be further developed into a policy selection MOE, which would allow policy selection that adapts to the already known data pool.

Using Query by Committee (QBC) in SR TAR process[62]. In this approach, a committee of models is used to answer the question on if a document should be marked as included within SR, with a measure, such as vote entropy or kullback-eibler divergence, used to determine the aggregated committee result. QBC is different from MOE, as it does not necessarily use specialised models or a router, rather, an ensemble of diverse models to make decisions, with the more disagreed about input being sent to humans for classification.

Elimination/minimisation of oracle within the SR TAR process. In the TAR process for SR, we employ sampling policies (π) to select documents for human relevance judgement. These policies, such as uncertainty sampling, aim to present the most challenging examples for the LR model to classify. We currently use the sigmoid output of the LR model as a proxy for classification certainty, with the 0.5 mark on the sigmoid curve representing maximal uncertainty for the LR model. However, this approach has limitations. The sigmoid output more accurately reflects the LR model's uncertainty, not the true difficulty for human classification. As a result, we may inadvertently present "easy-to-classify" documents to the human oracle, leading to inefficient use of human expertise. The author proposes incorporating more sophisticated models into the document selection process. This improvement involves introducing a pre-screening phase using advanced language models (e.g., GPT-4, Gemini, Claude). These models would analyse candidate documents before human review. Measures such as Shannon entropy of the model responses can be used to gauge true classification difficulty, only presenting documents to the human oracle when deemed necessary based on this analysis.

This approach offers several advantages. It balances computational resources with human expertise, potentially reduces the number of "easy" documents sent for human review, and reserves human judgment for truly ambiguous or complex cases. By implementing this enhanced process, the author aims to optimise the efficiency and effectiveness of the SR TAR workflow, making better use of both computational and human resources.

E Citations in medical research

Systematic reviews utilise research evidence to provide clinical practice recommendations. The communication of medical research follows standardised formatting conventions and primarily occurs through peer-reviewed publication [34]. When authors compose research papers, they must reference related works to substantiate their claims and situate their findings within the existing body of knowledge. These citations follow standardised formatting guidelines and are documented in the paper's reference section. This rigorous documentation of citations enables analysis of the relationships between research papers, operating under the assumption that studies that cite or are cited by a research article are relevant to that research.

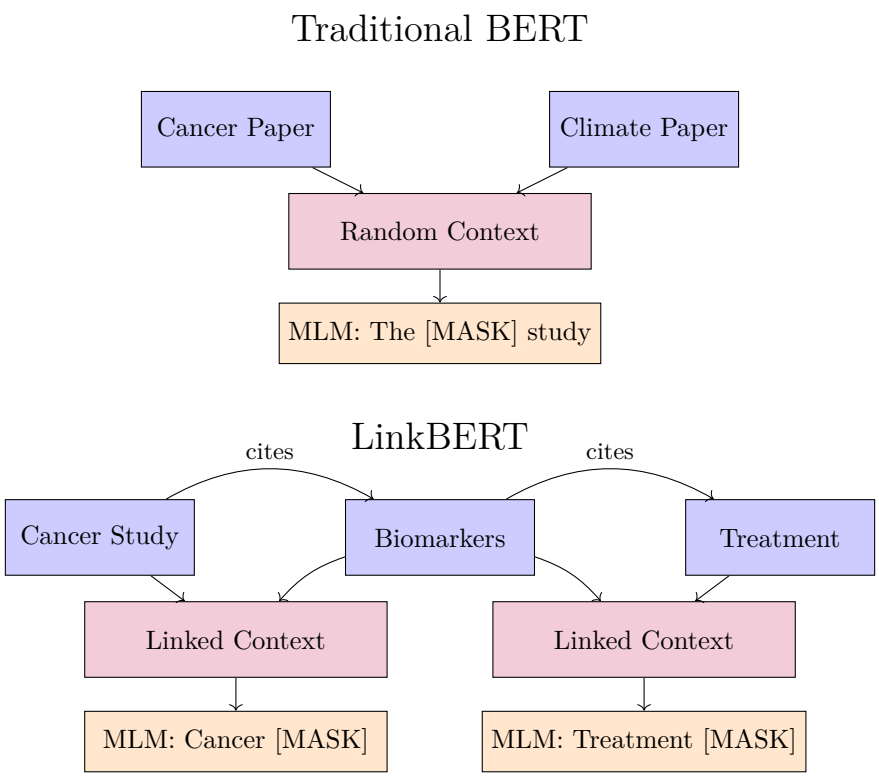


Figure 6: Comparison of document processing in traditional BERT versus LinkBERT. Traditional BERT (top) randomly groups documents into context windows, while LinkBERT (bottom) uses citation relationships to create meaningful document groupings for pretraining. The citation-based grouping ensures that semantically related documents are processed together during masked language modeling tasks.

E.1 Relation analysis improves CAL TAR performance

Recent advances in medical CAL TAR have indirectly demonstrated the benefit of relationship analysis for citations. The current leading encoder model, $BioLinkBERT_{base}$ achieved state-of-the-art performance on the CLEF dataset in a CAL setting by leveraging citations networks between research papers [35, 36].

The *LinkBERT* approach was to view a pertaining corpus as a graph of documents, with each document being a vertex and hyperlinks forming edges between documents. These related documents were then placed within the same context window. The approach differs from traditional *BERT* architectures, which randomly allocate documents to context windows without considering their relationships. While this might appear similar to curriculum learning approachings, *LinkBERT* is distinct in that it does not organise context windows by difficulty level.

BioLinkBERT, a domain-specific adaption of *LinkBERT*, was developed specifically for biomedical aplications and pretrained exclusively on PubMed articles, using citation relationships to estimate document relationships [11]. The model trianing process incorporated standard masked language modelling and next-sentence prediction techniques. Analysis of both the base model (100M parameters) and large model (340M parameters) against *PubMedBERT* across multiple benchmarks: BLURB[37], MedQA-USMLE[38], and MMLU-professional medicine[39]. The results demonstrated $BioLinkBERT_{large}$'s superior performance across all evaluated benchmarks, notably achieving a 3.2% improvement over PubMedBERT in the BLURB score.

Current research on document relationship-based encoders in the CAL process has not definitively established that

[11] https://huggingface.co/michiyasunaga/BioLinkBERT-base

document relations are the primary driver of performance improvements. Furthermore, the assumption that larger models consistently yield better results is not always the case. The author replicated the previously reported Goldilock Reproduce study, where $BioLinkBERT_{base}$ formed the classifier, except changing the model to the $BioBERT_{large}$ variant[12] as a classifier model, yet only achieved higher performance in R-Precision in 7 of 12 datasets/policy combinations. The empirical results, detailed in Table 8, show peak R-precision values of 0.847 for the relevancy selection policy (at FPT epoch 2) and 0.832 for uncertainty selection (at FPT epoch 1).Statistical analysis using the Friedman test revealed significant differences between Further Pre-Training (FPT) epochs in only 4 of 12 datasets when examined individually. More importantly, when analyzing all datasets collectively, no statistically significant differences emerged in R-precision values across FPT epochs for either relevancy selection or uncertainty selection policies. This findiing challenges the previously documented "Goldilocks problem" observed in non-medical domains. Specifically demonstrating that FPT does not yield statistically significant improvemnts in R-Precision.

This replication study has generated valuable insights for this PhD investigation. A significant finding indicates that seeking an optimal pretraining epoch within the CLEF dataset is unlikely to be productive for future research endeavors. The experimental design revealed several methodological considerations, particularly regarding the implementation of hyperparameters without robust empirical justification. These include the selection of a batch size of 25, the decision to fine-tune for 20 epochs, and the termination criterion of 501 labeled documents. These parameter choices, while functional, may impose limitations on potential improvements to encoder CAL process performance within the experimental framework.

The significance of these limitations becomes particularly evident when considering that observed R-Precision values approach the theoretical maximum of 1.0, with some instances achieving values as high as 0.945. In the context of the Goldilocks reproduce paper, datasets showing lower performance metrics, such as the CLEF 2019 dataset (with R-precision values of 0.82 for relevancy and 0.791 for uncertainty), present additional analytical challenges. The utilization of Large Language Models (LLMs) introduces complexity in interpreting the underlying causes of reduced performance in these cases.

While exploring larger, more sophisticated models presents a potential avenue for improvement, this approach faces practical constraints. Given the limitations of High-Performance Computing resources and PhD time constraints, pursuing research dependent on the development and availability of superior LLMs may not be the most pragmatic direction.

A crucial observation emerged from this research regarding the relationship between early document classification and overall performance. In iterations where strong performance was ultimately achieved at iteration 20, a notably higher number of relevant documents were classified earlier in the CAL process. This finding aligns with theoretical expectations: a larger corpus of correctly classified documents early in the process provides a more robust foundation for subsequent classification decisions. This insight carries significant implications for the next phase of this PhD research, suggesting that enhancing document availability in the early stages of the active learning process could substantially improve overall performance outcomes.

While *BioLinkBERT* represents a sophisticated approach that combines citation networks with contextual language understanding, this integration presents both advantages and limitations. The model's ability to capture complex semantic relationships between documents is valuable, but the contextual processing introduces potential inefficiencies. During pretraining, when linked documents are placed in the same context, the model must process all content within those documents—including sections that may be tangential or unrelated to the citing paper's specific reference. This contextual noise could potentially dilute the precision of the more direct relationships that citations inherently represent. In contrast, pure citation links directly capture intentional scholarly connections made

[12]https://huggingface.co/michiyasunaga/BioLinkBERT-large

by domain experts, providing a cleaner signal without the additional complexity of processing potentially irrelevant contextual information.

A fundamental question emerges from this research: Is contextual understanding of references truly necessary for effective CAL? Several factors suggest that citation networks alone might be sufficient and potentially superior. First, citations themselves represent a form of knowledge distillation, where domain experts have already identified meaningful relationships between documents. Second, analysing reference networks is computationally more efficient than processing full textual contexts. Third, citation network models tend to be more stable when updated, compared to contextual models. Fourth, the contextualization of citation networks may actually introduce noise into what would otherwise be clear citation signals.

Collection	**Dataset size**	**Model**	**R-Precision (↑)**		**Friedman (p)**	
			Rel.	**Unc.**	**Rel.**	**Unc.**
Clef 2019 dta test	8	BiolinkBert-Base-ep0	**0.909**	**0.857**	—	
		BiolinkBert-Large-ep0	0.897	0.803		
		BiolinkBert-Large-ep1	0.827	0.832		
		BiolinkBert-Large-ep2	0.812	0.774	0.914	0.632
		BiolinkBert-Large-ep5	0.841	0.814		
		BiolinkBert-Large-ep10	0.881	0.846		
Clef 2017 test	30	BiolinkBert-Base-ep0	0.812	0.794	—	
		BiolinkBert-Large-ep0	0.828	0.797		
		BiolinkBert-Large-ep1	0.826	**0.827**		
		BiolinkBert-Large-ep2	**0.858**	0.804	**<0.05**	**<0.05**
		BiolinkBert-Large-ep5	0.827	0.777		
		BiolinkBert-Large-ep10	0.799	0.757		
Clef 2017 train	20	BiolinkBert-Base-ep0	**0.838**	0.761	—	
		BiolinkBert-Large-ep0	0.778	0.765		
		BiolinkBert-Large-ep1	0.808	0.789		
		BiolinkBert-Large-ep2	0.767	0.701	**<0.05**	0.28
		BiolinkBert-Large-ep5	0.816	0.786		
		BiolinkBert-Large-ep10	0.827	**0.796**		
Clef 2018 test	30	BiolinkBert-Base-ep0	0.794	0.780	—	
		BiolinkBert-Large-ep0	0.789	0.774		
		BiolinkBert-Large-ep1	**0.812**	0.790		
		BiolinkBert-Large-ep2	0.797	**0.791**	0.52	0.50
		BiolinkBert-Large-ep5	0.763	0.773		
		BiolinkBert-Large-ep10	0.763	0.769		
Clef 2019 DTA int. train	20	BiolinkBert-Base-ep0	0.939	0.923	—	
		BiolinkBert-Large-ep0	0.939	0.902		
		BiolinkBert-Large-ep1	0.941	0.935		
		BiolinkBert-Large-ep2	0.948	0.921	0.78	0.50
		BiolinkBert-Large-ep5	0.952	0.945		
		BiolinkBert-Large-ep10	**0.945**	**0.947**		
Clef 2019 DTA int. test	20	BiolinkBert-Base-ep0	**0.934**	**0.900**	—	
		BiolinkBert-Large-ep0	0.899	0.856		
		BiolinkBert-Large-ep1	0.904	0.840		
		BiolinkBert-Large-ep2	0.909	0.878	0.87	**<0.05**
		BiolinkBert-Large-ep5	0.882	0.835		
		BiolinkBert-Large-ep10	0.865	0.841		

Table 7: Performance comparison across different collections and models

Table 8: Average R-precision of each FPT epoch for CLEF dataset

Policy	**ep0**	**ep1**	**ep2**	**ep5**	**ep10**
Uncertainty	0.813	0.832	0.813	0.815	0.814
Relevancy	0.840	0.845	0.847	0.842	0.835

E.2 Direct citation network mining within medicine research

Performant, simple, and robust approaches to citation network mining already exist within medicial research. Let G be a citation graph where:

- D_i represents a research article of interest as a vertex in G
- D_{ip} represents the set of articles referenced by D_i
- D_{if} represents the set of articles that reference D_i
- Both sets are subsets of G: $D_{ip}, D_{if} \subset G$
- $D_{ip} \cap D_{if} = \emptyset$, so searching both sets will provide different relevant articles

Relevancy is defined as a function $R : D \to [0, 1]$, where:

- 0 denotes no relevance
- 1 denotes maximum relevance
- For any set of documents D_{set}, relevancy is defined as $R(D_{set}) = R(d)|d \in D_{set}$

Two primary citation network mining approaches are defined:

- Backward citation searching (BCS): examining all articles in D_{ip}[40, 41]
- Forward citation searching (FCS): examining all articles in D_{if}*[13]

Backward and forward citation searching (BCS and FCS) are both straightforward and effective approaches that inherently respect the chronological relationships between research articles, as papers can only cite previously published work. The significance of these methods is demonstrated by their recommended use in Cochrane systematic reviews, particularly during the identification phase. A study of Cochrane reviews conducted between November 2016 and January 2017 found that 87% reported using BCS, while 9% utilized FCS [42].. The Cochrane Handbook explicitly mandates the use of BCS (criterion C30) in the search stage, though it makes no mention of FCS [43]. However, neither the use of BCS nor FCS is addressed in the Handbook's guidelines for the screening phase.

The application of Backward and Forward Citation Searching (BCS and FCS) within an active learning process represents an understudied area of research (see Figure 7 for search strategy details). To establish the novelty of this augmentation, several key distinctions must be clarified. While this PhD research focuses on the title and abstract screening phase of systematic review generation, BCS and FCS have traditionally been confined to the identification phase (as illustrated in Figure 1). Conventionally, title and abstract screening serves to reduce the workload for the more resource-intensive full-text screening phase. However, from a computational perspective, restricting the screening process to titles and abstracts is unnecessary, as the computational cost remains manageable when including full texts.

[13]FCS involves using a citation index to identify studies that cite a source study. A citation index is a database of scholarly articles and their citations, such as PubMed, Google Scholar, Scopus or OpenAlex

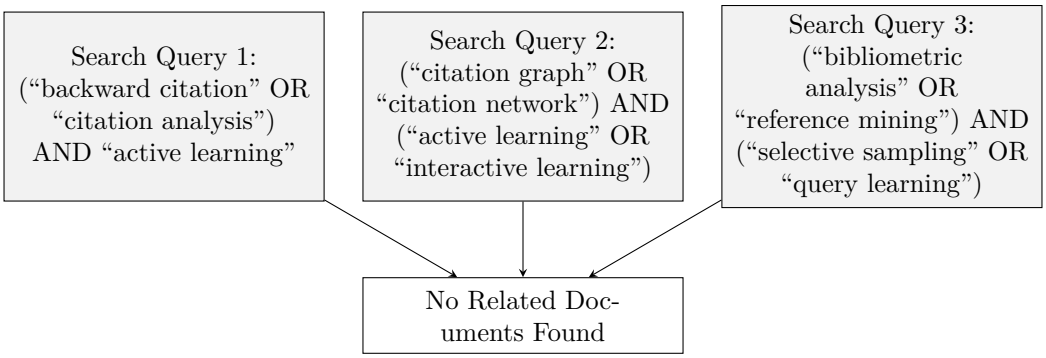
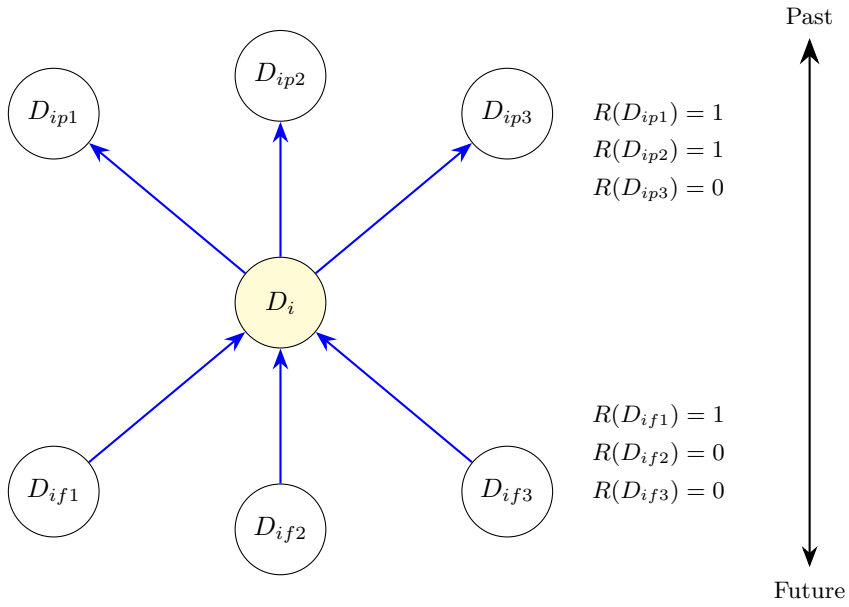


Figure 7: Results from literature search on citation index arxiv and pubmed demonstrating absence of related works, ran on 13th November 2024



These citation networks are rich in relevant documents, much more so than that of the document collection, which is demonstrated by the author comparing precision of pools using BCS and FCS against that of the entire document collection in Figure **??**. The logical, and simple augmentation of the encoder CAL approach would be to exhaust both BCS and FCS networks of a seed document prior to initiating the encoder CAL process.

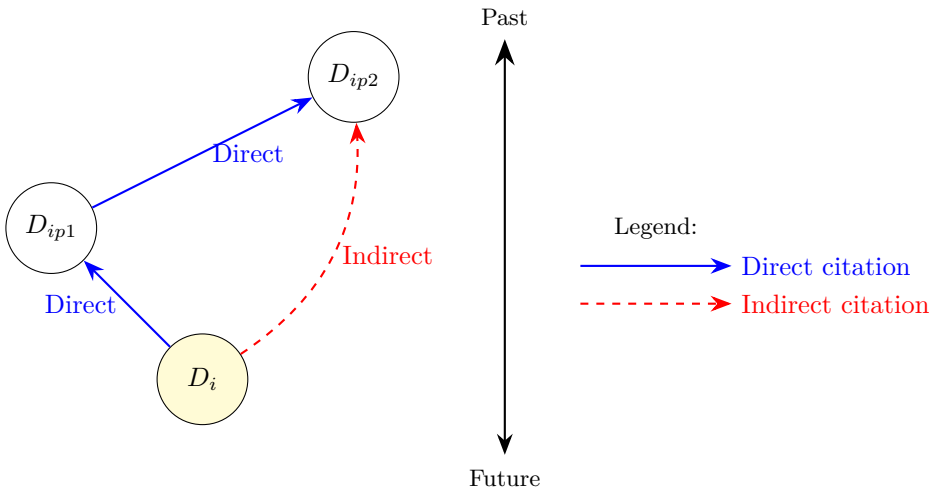
The theoretical benefits of citation network mining are that it can be used to augment the CAL process in ways that overcome some of the limitations of this process. Firstly, CAL requires labelled data to train a classifier model, which is assumed to perform better with more data points. Encoder CAL approaches suffer disproportionately to that of feature-based CAL approaches due to their need for larger amounts of training data to effectively learn meaningful representations. This is because encoder models like BERT need to learn complex contextual relationships between words and concepts, whereas feature-based models can rely on simpler statistical patterns. When working with limited labeled data in the early stages of screening, encoder models may struggle to generalise well, potentially leading to suboptimal performance in identifying relevant documents. As discussed in the Encoder CAL process, often a single sample seed document is used during the first epoch for fine-tuning. A better approach would be to exhaust the citation network of that seed document first for labelling, before using revealed relevant documents to fine-tune the model, potentially resulting in a more performant model at the earlier stages of screening with less

oracle cost.

E.3 Extending current citation network mining approaches

BCS and FCS citation network mining faces a significant limitation in its inability to identify indirect citation relationships. An indirect citation occurs when research papers are connected through intermediate references, forming a chain of citations rather than a direct reference. For instance, when document D_i cites document D_{ip1}, which in turn cites document D_{ip2}, a relationship exists between D_i and D_{ip2} despite the absence of a direct citation. This relationship represents an indirect citation, which is shown in Figure E.3. This causes issues if D_{ip1} is not included in the document pool, as D_i and D_{ip2} will no longer have an edge.

This constraint makes it unsuitable as a complete solution for document relationship discovery for the encoder CAL process. However, researchers have proposed several modifications to the citation network mining process to address this limitation:



- **Matching isolated nodes based on similarity metric of their embeddings**: If N is all the documents in the total pool, and $N_{isolated}$ is the set of documents that are not cited by any other document in N, then for each document $D_{ip} \in N_{isolated}$, find the document $D_i \in N$ with the highest similarity metric (i.e. cosine similarity) to D_{ip}. Add a artificial edge between D_i and D_{ip}.
- **Matching isolated compoments on similarity metric of their embeddings**: When analyzing document clusters, some small groups of documents (called isolated components) may be disconnected from the main cluster. These isolated components have fewer connections to other documents, which can reduce classification accuracy. To fix this:
  - Identify isolated components $C_{isolated}$ that have fewer or equal nodes than the main cluster
  - For each node in these isolated components
  - Calculate a similarity metric (i.e. cosine similarity) to nodes in larger clusters C_i
  - Connect it to the most similar large cluster by adding a artificial edge

This constraint however doesn't make it unsuitable as a partial improvement to the encoder CAL process for identifying relevant documents based on the initial seed document. Even without considering indirect citations, assessing the citation network of the seed document is potentially more relevant than that of the entire document

collection. In table 9 it is unequivacle that the precision of relevant documents within pools using BCS, FCS and both together against that of the entire document collection is much higher.

Make this data!

Table 9: Precision of relevant documents within pools using BCS, FCS and both together against that of the entire document collection

E.4 Research Question 1

Proposal: Using BCS and FCS pools before encoder CAL process can improve the precision of the encoder CAL process.

E.5 Graph Neural Networks

A research paper is a rich source of information, and contains multiple features that can be used to represent that document, however in the title and abstract screening phase, it is limited to only using the title and abstract features. As the previous research area aims to demonstrate, utilising other features could improve the precision of the encoder CAL process, so, logically utilising more features could improve the precision of the encoder CAL process further.

Previous work by this author has demonstated that features about authors, primary topic and publication date all impact classification accuracy. link this to the retraction watch paper

In-keeping with the above research theme, graph neural networks offer a natural extension to considering additional document features, and still being able to utilise the structural information about relationships between documents.

E.6 LLMs and citation network mining

The motivation for using LLMs and graph networks is to combine the structurality of graph citation networks with the ability of the LLMs to comprehend the semantic meaning of documents. As outlined, citation network graph analysis occurs above the document level through utilisation of extracted features about documents. LLMs are a natural replacement for extraction of features, as they possess the ability to understand semantic meaning of documents. The ultimate goal to use LLMs and graph networks is to complment and enhance isuses with the other.

Research has been conducted into the use of LLMs within the graph neural networks, and has developed a robust taxonomy for categorising the use of LLMs within graph networks [**llm4g**].

The first application of LLms within graph networks is to use LLM as an enchancer. Typically graph neural networks encode text into nodes using simple bag-of-words, skip-gram or TF-IDF. LLMs are able to encode text into nodes using more complex features, such as semantic meaning, which can be used within the graph neural networks. This can be further subdivided into explanation based and embedding based enchancers.

Explanation based enhancers query an LLM using prompting to capture higher level features about documents, which is used to enrich node representations prior to processing with a graph neural network, with the process being abstracted in in Figure 8. The approach used by https://arxiv.org/pdf/2305.19523 was to prompt GPT 3-5 with the abstract and text of a document along with a questions about that document using a zero-shot approach. The LLM reponse then forms features which are amended to the original node representations. Issues with this approach is that this requires domain specific knowledge, as features which are deemed important (and hence prompt used) are dependent on the domain of the research. It was performant on the pubmed domain, scoring greater node

classification accuracy using this approach (0.9618 ± 0.0053) than utilising an LLM alone on pubmed data (0.9494 ± 0.0046).

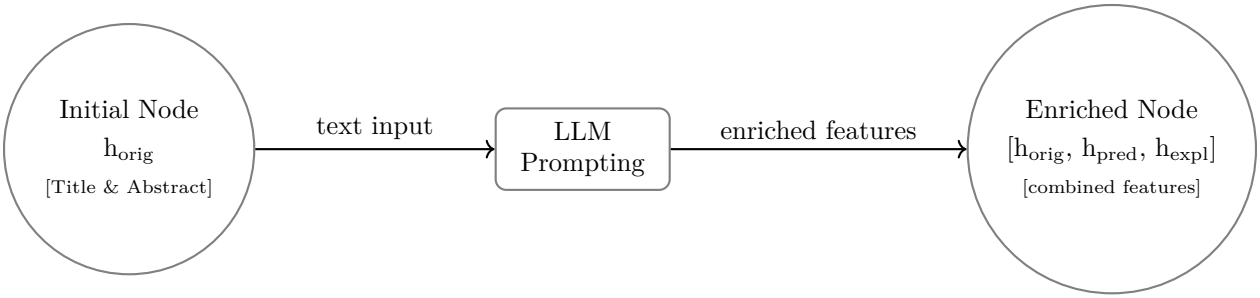


Figure 8: Node feature enrichment process using LLM and LM

E.7 Research Question 2

Proposal: Utilising more features in the encoder CAL process can improve the precision of the encoder CAL process.

F Notes on Graph Neural Networks

A node is represented by a feature matrix, which contains information about the document. This **Node feature matrix**, X, which has the dimensions of m (the number of nodes) and n (the number of features). $X \in \mathbb{R}^{m \times n}$. X does not have to be a square matrix, and does not encode any information about the structure of the graph.

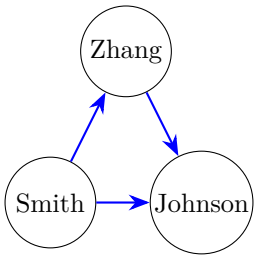
Consider 3 research papers as nodes, with features: [Author, Title Length, Abstract Length, Citation Count]

$X = \begin{bmatrix} \text{"Smith"} & 82 & 500 & 45 \\ \text{"Johnson"} & 95 & 475 & 23 \\ \text{"Zhang"} & 67 & 612 & 89 \end{bmatrix}$ Where $X \in \mathbb{R}^{3 \times 4}$ represents:

3 papers (rows) 4 features per paper (columns) Mixed data types (categorical and numerical)

Structural information is encoded in the **adjacency matrix**, A, which has the dimensions of m (the number of nodes) and m (the number of nodes). $A \in \mathbb{R}^{m \times m}$. A encodes information about the structure of the graph, and is used to determine relationships between nodes. Conventionally the source nodes are the rows, and the destination nodes the columns of the matrix. 1 indicates an edge between the source node u and destination node v. Note that there is a choice to make here, with the diagonal of the matrix being 0 or 1. This choice is based on wheter you consider the source node to be connected to itself. In cases where the representation of the node is dependent on itself and adjacent nodes, the diagonal should be set to 1. In the scenario of citation networks, the diagonal should be set to 1, as a paper is likely to reference and build upon its own findings throughout. By setting the diagonal to 0, it is akin to attempting to predict the representation of the node base only on its adjacent nodes, which is not the case in citation networks. If an adjaceny metrix is symmetric around it's diagonal, then the graph is undirected, otherwise it is directed (i.e. U is connected to V and V is conencted to U). In citation networks, this is not the case, as because paper A cites paper B, it does not mean the reverse is true.

Consider the same 3 research papers, with the following adjacency matrix: $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ Which represents the following graph:



With both X and A defined, we can numerically represent the graph. The node feature matrix X is the initial/input node features, with our goal for learning on graphs to learn node embeddings $H \in \mathbb{R}^{N \times D}$ where D is a chosen hidden dimension size.

G Message Passing Neural Networks

We need an approach that can work with the graph structure, which has variable number of nodes and edge conenctions between nodes. Historically with the CNN architecture, the input size was fixed, and the network was able to learn spatial invariance through the use of convolutional filters that were invariant to the location of the feature in the input. With graph structured data, the number of nodes and connections between nodes can vary for each graph, and spatial invariance is not invariant to the location of the feature in the graph.

Message Passing Neural Networks (MPNNs) are a type of graph neural network that can learn spatial invariance through the use of message passing between nodes. The basic idea of MPNNs is to iteratively update node representations by passing messages between connected nodes. This process is repeated for a fixed number of iterations, or until convergence.

The process is defined as follows:

- Message: every node decides how to send information to neighboring nodes it is conencted to by edges
- Aggregate: nodes recieve messages from all their neighbors, who also passed messages and decides how to combine the information from all of its neighbors.
- Update: each node decies how to combine neighbourhood information with its own information and updates it embedding for the next timestep.

By doing this we have nodes pass each other information and disseminate information around the graph, allowing the network to learn spatial invariance. This can be repeated for a fixed number of iterations (K), with the larger the value of K, the more the more diffuse the information around the graph becomes.

Each section of the MPNN process in more detail:

G.1 Message

The source node U will pass a message m_{uv} to the destination node V. The message depends on the GNN architecture with the easiest example message being passed being U node's feature h_u vector to V.

G.2 Aggregate

The destination node V will recieve messages from all its neighbouring nodes, and needs to decide how to combine the information from all of its neighbours. This is typically done using a sum, average or max pooling of the messages from all neighbouring nodes. It is important that the aggregation function has to be a permutation invariant function, as the order of the messages should not affect the output.

This gives us a combined neighbourhood node embedding, denoted as $h_{N(V)}$, where $N(V)$ is the set of all neighbouring nodes to V, meaning all nodes connected to V by an edge.

$h_{N(v)}^{k+1} = AGGREGATE(h_u^k, \forall u \in N(v))$

G.3 Update

Each node updates its own embedding based on the combined neighbourhood embedding and its own embedding from the previous timestep.

$h_v^{k+1} = \sigma(W \cdot CONCAT(h_v^k, h_{N(v)}^{k+1}))$

Search criteria for Graph Neural Networks and Active Learning ("graph neural network" OR GNN) AND ("active learning" OR "interactive learning") AND (document OR citation OR literature) AND ("relevance feedback" OR "document classification") AND ("semi-supervised" OR "partially labeled") Database-specific versions:

arXiv: search within cs.LG, cs.IR, cs.CL categories PubMed: add "systematic review" OR "literature review" terms

IV TIMELINE

This is a non-binding timeline for this PhD. It is accepted that this will likely change and only represents estimates given the current available information, the timeline is available in Gantt format in Figure 9

In creating this time-line for the PhD several assumptions were made:

- **Holiday Periods:** Two weeks of holiday have been accounted for during the summer and a two-week period over Christmas and New Year. This timeline does not account for additional holidays, which will be determined later.
- Front-Loaded Research: The research plan is front-loaded, with significant emphasis placed on the early stages of the project. This approach allows additional literature searching, coding, and experimental setup to be completed in advance, providing a solid foundation for later stages of research, the research questions of which are currently flexible. This strategy also ensures that any necessary adaptations can be made based on early findings, reducing the risk of major delays later in the project.
- **Research Flexibility:** Significant breaks are scheduled between work on research questions to allow for adaptation or extension if research questions need to be adjusted. In addition, there is a dedicated period for each of the three research themes before beginning coding or experimental setups. This time is intended for further literature review, acknowledging the rapidly evolving nature of this field. Advancement of the literature are expected before the start of each research question period.
- **Undetermined Research Questions:** Two research questions have deliberately been left open. Depending on the findings of the first three research questions, new research opportunities or developments in the field may emerge that require investigation.
- **Publication goals:** The goal is to produce at least one publishable piece of work for each research question. The preferential publication venue is SIGIR, recognised as the highest-rated publication venue for this sub-domain, aside from broader higher impact journals such as ACL. This goal is deliberately ambitious, as until the research is complete, it is impossible to determine the merits of its findings.
- **Conference Scheduling:** SIGIR abstract submissions and conferences occur around the same time each year, which has been considered in the planning. Note that many RQs are concluded prior to the expected submission dates.

A Potential Threats

A risk matrix of all potential threats is outlined in Table 10.

- **Research Delays:** Unforeseen challenges in research or coding could lead to delays. These could arise from the complexity of the research questions, issues with experimental setups, or unexpected results that require additional analysis. This is mitigated through periods at the end of research questions which can be utilised, if necessary.
- **Technological advancement:** The fast-moving nature of the field could result in the emergence of new technologies or methodologies that could make parts of the planned research less relevant or require a significant change of focus. This is mitigated through periods allowing additional literature review before coding on that research question.

- **Publication risks:** There is always the risk that the research might not yield results suitable for publication or that the publication process itself might be more time-consuming than anticipated, especially if revisions are required. The nature of the publication system is that there will be extended gaps between the completion of the research and its publication, making modification of the research to match a reviewer's expectations difficult.
- **Personal and Health Factors:** Extended periods of high-intensity work can lead to burnout or health problems, potentially affecting the planned schedule. The Gantt chart does not and cannot fully account for extended absences due to illness or other personal factors.
- **Open ended research questions:** It could be that at the time of reaching the open-ended research questions, a research area has not been identified. The author believes that this is unlikely, given that this field is limited, and other potential research questions have been dropped to ensure this flexibility.
- **Competing time constraints from PGDip:** Due to the concurrent requirements of the PGDip with this PHd, there will at times be constraints placed on my available time. So far, I believe, I have demonstrated good time keeping skills, and believe this schedule to be reasonable given the other competing time constraints. If there is a clash, PH.d. will take priority over PGDip.

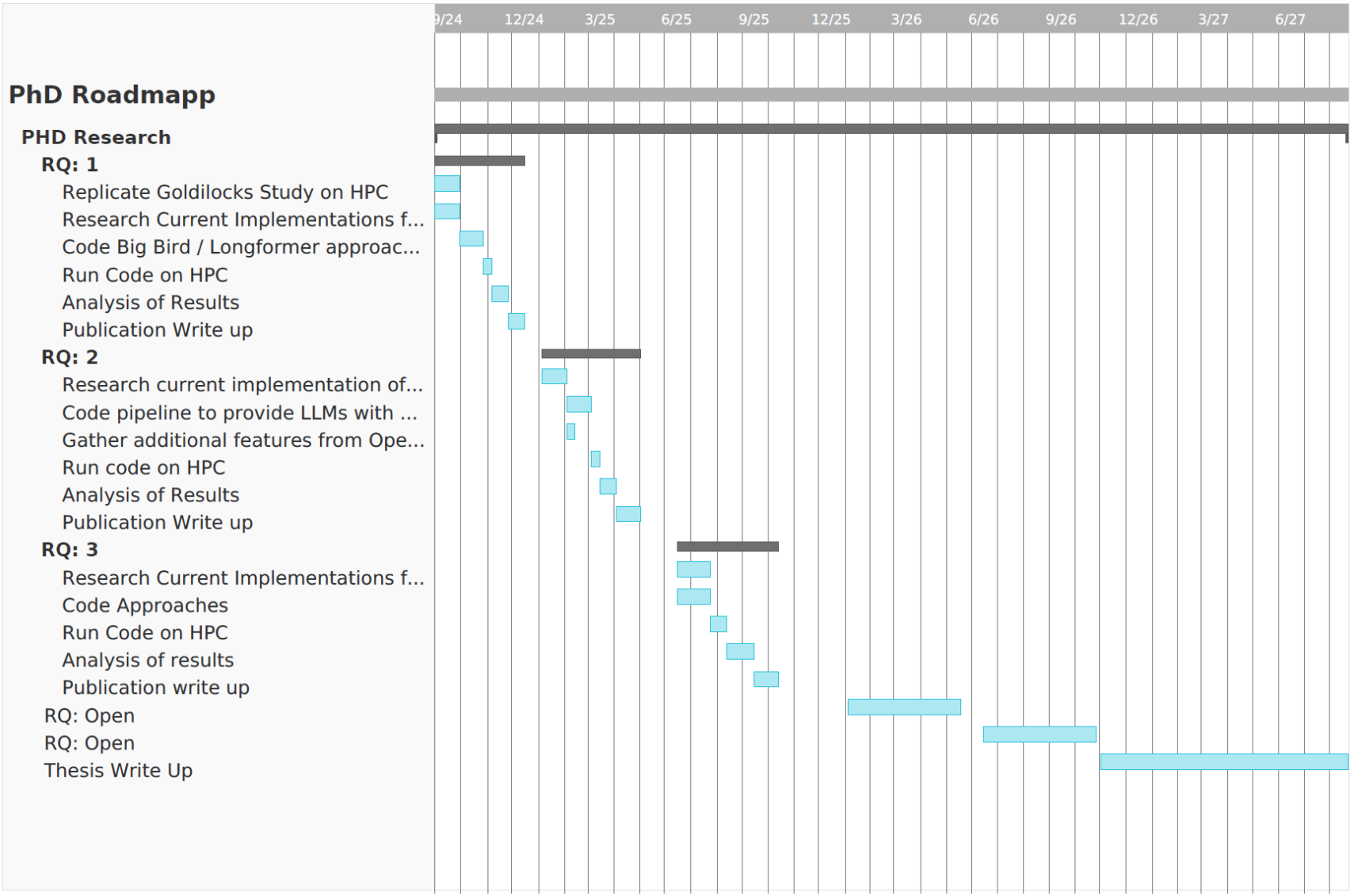


Figure 9: Gantt Chart for OVerview of Timeline for PHd

Threat	Likelihood (L)	Impact (I)	Risk	Response
Research Delays	3	4	12	Reduce the number of research questions to mitigate delays.
Technological Advancement	3	3	9	Increase time allocated for additional literature review to stay updated with advancements.
Publication Risks	3	3	9	If results are not deemed valid during the analysis phase, proceed to the next research question without spending time on write-up.
Personal and Health Factors	2	4	8	Prioritise health by taking regular breaks and managing workload effectively to avoid burnout.
Open-ended Research Questions	2	3	6	Continuously monitor new publications within the domain to ensure relevant and timely research questions.
Competing Time Constraints (PGDip)	3	3	9	Minimize involvement in PGDip activities where possible to focus on PhD work.

Table 10: Risk Scoring Matrix for threats to PhD. Responses are provided for medium impact risks.

V ETHICS

This research will comply with the research ethics norms and rules outlined by the University of Sheffield. Most of the data sets used by this research (CLEF, Synergy, and RCV1-v2 data sets) would fall into the category of publicly available anonymised data/published media; hence, ethical approval is not required.

However, the Jeb Bush email dataset contains unanonymised emails. Although it is publically available and used for research, an application for its use will be sought through the university's ethical approval process in due time.

VI PROFESSIONAL DEVELOPMENT PLAN

The training provided as part of the PhD with integrated PGDip replaces the standard Doctoral Development programme, and further activities outside of this are not required to be undertaken.[14]. The PGDip component has multiple strands associated with it, such as participation in group work, a requirement for technical training, responsible research and innovation, attending discipline-specific lectures, involvement with journal clubs, outreach work from the department itself, and completion of the FCE6100 (Research Ethics and Integrity training) as part of CDT's COM61003. As evidence of active involvement with professional development, the author submits his logs from his first year of engagement with the process, which were submitted in June 2024 - see Appendix B and C. These are due yearly (i.e. every June) as the programme progresses. Due to the large size of this, this is not included in this body of work but rather is included as additional PDFs for the readers.

[14] https://sites.google.com/sheffield.ac.uk/slt-cdt-handbook/the-centre

VII AUTHORS SUPPORTING WORKS

The author is actively involved in 3 other projects with publishable outcomes.

A Predicting Retracted Research

In my study, "Predicting Retracted Research," I, along with my supervisor Mark Stevenson, explored the challenge of identifying flawed scientific publications before their dissemination - see Appendix A. We developed a novel data set by combining information from the Retraction Watch database and the OpenAlex API. This dataset includes metadata, abstracts, and citation metrics for 16,224 articles (8,112 retracted and 8,112 non-retracted) published between 2000 and 2020. Various machine learning models are used to predict retracted articles, with a gradient booster model achieving the highest precision at 0.691. An ablation study highlighted the critical role of the abstract in classification accuracy, recall, and F1 score, whereas the First Author's Country was pivotal for precision in feature-based classifiers. The research demonstrates the feasibility of using machine learning to aid peer review by highlighting potentially problematic research, although further refinement is necessary for practical implementation. The data set and code are publicly available to encourage further research in this area. This work has been written and is intended to be submitted to Informetrics[15]for publication. This fits the overall PhD research theme of assessing research data and ultimately showing that the evidence being assessed, even despite being published, can be flawed. Currently, no further follow-up is intended for this work.

Threat to Ph.D.: Low. Most of the work has been completed for this project.

B The stopping problem

Within the creation of SRs, the overall goal of TAR is to get to as near perfect total recall as possible, or when you have exhausted a resource such as a human reviewer. However, other stopping strategies could be more materially useful, to fulfil an information need, such as stopping when you have returned enough information to make a decision. In this joint research, between the author, Mark Stevenson and Anthony Hughes, we use information provided from cochrane reviews and create algorithms that stop when the acquisition of knowledge (positively included studies) meets a criterion and then evaluate how close these stopping algorithms got to the final outcome.

Threat to Ph.d.: Medium, this work is ongoing and has undergone many revisions. However, it is likely to result in high-quality publishable research.

C CPET analysis and deep neural networks

A cardiopulmonary exercise test (CPET) is performed before certain anaesthetic procedures, the outcome of which is used to determine the suitability of the patient for this procedure. Current approaches use summarised data to generate decision models, whose data are derived from summary values provided by the machine. The machine also records " breath-by-breath" data measurements, which, while they are the basis for the summary values, are not used by these models. This research project attempts to determine whether the use of deep neural networks, with these " 'breath-by-breath"' data, is superior to that of the traditional summary-model approach. This research was devised by an NHS researcher.

Threat to Ph.d.: Low. I am providing coding assistance to this project, and will not be involved in analysis or extensively involved in research write-up, outside of the technical side. This is also likely to result in publishable research, and will likely be published in medical domain venues, promoting interdisciplinary work.

[15]https://www.sciencedirect.com/journal/journal-of-informetrics

VIII APPENDIX

A RETRACTION WATCH RESEARCH

Predicting retracted research

Aaron HA Fletcher[a], Mark Stevenson[a]

[a]*Department of computer science, The University of Sheffield, Sheffield, S1 4DP, United Kingdom*

Abstract

Retracting published research is an important safeguard against the dissemination of flawed or fraudulent scientific information. However, detecting problematic research prior to publication remains a challenge. This paper describes the creation of a novel dataset and machine learning models to predict retracted articles. The data set combines information from the Retraction Watch database and the OpenAlex API, including article metadata, abstracts, and citation metrics. A total of 16,224 articles (8,112 retracted and 8,112 nonretracted) published between 2000-2020 were included. Several machine learning models were trained on this data, with a gradient boosting approach achieving the best precision (0.691). An ablation study revealed that the abstract of the article was the most important feature for classification for the accuracy, recall, and F1 score metric. First Author Countries was the more important feature for feature-based classifers with the Precision Metric. This work demonstrates the potential for using machine learning to assist in identifying problematic research during the peer review process, though further improvements in model performance are needed before practical application. The data set and code are made publicly available to support future work in this area.

Keywords: Retraction prediction, Machine Learning, Scientific publishing

1. Introduction

Retracting journal articles is a crucial safeguard against disseminating inaccurate or unreliable information. Conversely, the number of journal retrac-

Email addresses: ahafletcher1@sheffield.ac.uk (Aaron HA Fletcher), mark.stevenson@sheffield.ac.uk (Mark Stevenson)

tions can act as a proxy for failures within the publishing process, indicating instances where previous safeguards, such as peer review and editorial oversight, have failed to prevent the publication of flawed research. Numerous studies have successfully and repeatedly demonstrated an increasing trend in journal retractions [1, 2]. However, given the constraints of the current peer review system and the growing ability of natural language generation tools, preventing the publication of inaccurate or unreliable information versus providing a venue for the dissemination of research remains a challenging balance [3].

Abstractly, the decision to publish research is a binary classification task. The reviewer(s) act as a function that classifies an input (research) into two classes: to accept or not accept. This process closely resembles a text classification task in natural language processing (NLP), where the text is categorised into classes based on a function (such as sentiment) [4, 5]. Recent advances in NLP, including word embeddings, deep neural networks, and transformer architectures, have demonstrated considerable success in text classification across various domains. However, all supervised machine learning approaches fundamentally rely on labelled datasets, which to date have not been available.

Seeking the automatisation of identification of flawed research is important because of the potential benefits it could bring: warning the peer reviewer of any potential retraction risk before deciding to publish or precluding flawed research circulation. Although retracted articles are not void of academic usefulness, as they can be used to dismiss prior domain knowledge or direct future research areas, the valid use of retracted articles hinges on whether the end user is aware of an article's retraction status, which given the differing approach on how works are retracted, is not always clear. Continued improper research publication can have severe consequences, not just for the authors but also for the journal's reputation and the domain's integrity.

1.1. Existing Literature

Retractions occur for various reasons, broadly categorised into two main groups: (1) honest errors in otherwise ethically conducted research (estimated 73.5% of PubMed retractions between 2000 and 2010), and (2) improper or fraudulent research practices, including data fabrication, plagiarism, or false authorship claims (estimated 26.6% of PubMed retractions in the same period) [2]. However, these proportions can vary significantly

between disciplines, as evidenced by the prevalence of misconduct-related retractions in BioMed Central journals [6]. This conflicting evidence likely stems from the challenges in accurately determining researchers' motivations, resulting in imprecise classification criteria. From the perspective of researcher end-users, the specific reasons for retraction may be less critical than the fact that retracted research is inherently unreliable, following the principle "garbage in, garbage out" [7]. Moreover, retraction reasons are determined retrospectively and would not be available when initially classifying a publication's risk of retraction.

Limited research exists on the demographics of authors producing retracted articles. Studies have found a significant association between first authors from lower-income countries and retractions due to plagiarism, suggesting global variations in retraction reasons [8]. Interestingly, contrary to the global trend, fraudulent research is more prevalent in the United States, with over half (53%) of fraudulent articles authored by "repeat offenders"[1]. These findings suggest that demographic characteristics, such as an author's name or country of origin, could be valuable features in a retraction classification dataset.

A 2022 study that examined retracted medical articles using the Retraction Watch dataset, Web of Science, and journal citation reports highlighted the growing phenomenon of paper mill retractions. The study reported an increase in retractions related to paper mills, predominantly associated with China [9]. The median time to retraction was two years, decreasing as the impact factor of the journal increased. Another study in 2021, also focused on the medical domain, revealed differences in the types of articles retracted, with 83.8% being original research and 8.6% being "meta" research [10]. These findings suggest that characteristics such as the impact factor of a journal, the country of origin, and the type of article may contain valuable information for modelling retractions.

To date, no research has investigated the plausibility of machine learning modelling in the prediction of these retracted articles.

1.2. Paper Contributions

- A novel open-source data set that can be used to model the prediction of retracted articles.
- The creation of classifier models to classify if an article is retracted.

2. Dataset Generation

Retraction watch is a human-validated retraction dataset. Retraction watch is compiled from various sources, including journal databases, institutional reports, social media, and direct tips [11, 12]. Although not exhaustive due to stealth "retractions" [13], it provides partial metadata for some retracted articles, such as title, journal, publisher, and author. OpenAlex, an open online catalogue of works, similar to Scopus and Web of Science [14], aggregates data from multiple sources monthly. The combination of these data sets was used to create a single data set suitable for predicting article retractions.

2.1. Dataset Generation

The Retraction Watch dataset was obtained and queries to the OpenAlex API occurred on 24/07/2024. All data retrieved are available in the GitHub repository data folder[1].

2.1.1. Inclusion/Exclusion Criteria

Unique journals in the retraction watch dataset were calculated after applying the journal and work exclusion criteria listed in Tables 3 and 4. Retractions were limited to a 20 year period from 2000 to 2020 due to the median lag of the works being retracted being 2 years, the lack of retracted works before this date and the increased use of natural language technologies subsequent to this period; see Figures A.6 and A.6. This generated a list of journals with retractions and retracted articles.For each journal, title, works count, citation count and H index features were recorded.

Only articles and review works were used within the dataset, with the type being determined by OpenAlex's *"type"* field. Conference papers were excluded due to the previously reported mass retraction of conference papers undertaken by the Institute of Electrical and Electronic Engineers between 2009 and 2011 (having pulled over 10,000 such papers in the past two decades) [15]. For each retracted article, another article was sampled randomly from the year of the retracted article also did not meet the works exclusion criteria outlined in Table 4, was not included in the retraction watch dataset and who's OpenAlex API flag of *'is_retracted'* was False. Unretracted

[1] `https://github.com/afletcher53/RetractionWatch`

works's who's title contained the keywords *"retraction", "retraction:", "withdrawn", "correction", "erratum", "retracted", "withdrawal", "conclusion", "editorial", "contributions", "commentary", "contributors"* were not eligible for sampling. From all works (retracted and unretracted), any were dropped if the abstract or title contained the words "elsevier", "notice", "editor", "editors", "publisher". For further information on works exclusion criteria, see Table 4.

For each work (retracted and non-retracted), the following features were recorded from OpenAlex:

1. Abstract Inverted Index
2. Publication Date
3. Primary Topic
4. First Author
5. Institution
6. Citation Count
7. First Author Countries
8. Is Retracted Flag
9. Article Type

2.1.2. Preprocessing

All textual features were preprocessed by converting them to lowercase and eliminating non-ASCII characters, HTML tags, numbers, additional symbols, and whitespace. The words listed in Table 2 were removed from the textual characteristics. Each data point was labelled as 0 (unretracted) or 1 (retracted). Finally, all features were concatenated with a descriptor preceding the values, resulting in the data format shown in Table 1. To balance the data set, the nonretracted works were randomly undersampled to match the number of retracted works. The data set (n=16224) was randomly assigned to three groups: test (20%, n=3245), train (64%, n=10383), and validation (16%, n=2596). For all feature-based classification models, all inputs were vectorised using a count vectorizer (max ngrams = 1). The resulting count vectors were then adjusted using saturated term frequency-inverse document frequency (B = 0.3 and K1 = 2). For contextual language understanding models (i.e., BERT), all inputs were tokenised using the *"bert-based-uncased"* model, with a maximum token length of 512 [16].

Table 1: Examples of generated data.

Label	Input String
1	**title** the feasibility of improving impact resistance and strength properties of sustainable concrete composites by adding waste metalized plastic fibres first **first author** hossein mohammadhosseini **first author countries** MY **primary topic** fiber reinforced concrete in civil engineering abstract **abstract** waste plastic results in waste discarding disaster and consequently cause significant harms to the environment the utilisation of industrial wastes production sustainable concrete has attracted much consideration recent years because lowcost materials along with saving a place for landfill purposes also enhance performance concrete in this paper feasibility metalized wmp fibres palm oil fuel ash pofa composites was investigated by assessing impact resistance strength properties six mixes containing wmp varying from length mm were made ordinary portland cement opc a different six mixtures same fibre content made where pofa substituted [Abstract truncated for brevity] **citated by count** 82 **publication date** 2018-04-01
0	**title** synergetic photoluminescence enhancement of monolayer mosvia surface plasmon resonance and defect repair **first author** yi zeng **first author countries** CN **primary topic** twodimensional materials **abstract** A the weak lightabsorption and low quantum yield qy in monolayer mos are great challenges for the applications of this material practical optoelectronic devices here we report on a synergistic strategy to obtain highly enhanced photoluminescence pl by simultaneously improving intensity electromagnetic field around qy mos selfassembled submonolayer au nanops underneath bistrifluoromethanesulfonimide tfsi treatment surface used boost excitation qy respectively an enhancement factor pl as high is achieved mechanisms analyzed inspecting contribution spectra from a excitons a trions under different conditions our study takes further step developing highperformance devices based **citated by count** 10 **publication date** 2018-01-01

Table 2: Banned Words / Phrases removed from corpus.

retraction	retracted	retract
retractionwatch	retraction watch	removed
withdrawn	withdrawal	withdraw
retracted article	article	

Table 3: Journal Exclusion Criteria.

Criteria	**Description**
CrossRef	If journal was not included in CrossRef's journal title list.
Works Count	If work count (*based on OpenAlex API*) - total retraction count (*based on Retraction Watch Dataset*) $<$ Sample Size (*1*)
Retraction Count	If journal total retractions < 5 (*determined by the Retraction Watch dataset*).

Table 4: Work Exclusion Criteria.

Criteria	**Description**
Retracted Works	If Retraction Watch parameter *'ArticleType'* not in {Research Article, Conference Abstract/Paper, Clinical Study, Review Article, Case Report, Meta-Analysis}
Retracted Works/Non-retracted works	If OpenAlex *'source'* not in {Conference, Journal} and *'type'* not in {article, review}
English Language	Work excluded if OpenAlex API *'language'* value not 'en'
ISSN Data	If OpenAlex API *'issn'* value not available
OpenAlex ID	If OpenAlex API *'id'* value not available
Article Type	If OpenAlex API *'type'* value not in {article, review, conference-paper}
Publication Year	If *'publication_year'* OpenAlex API value not available
Publication Year Minimum	If *'publication_year'* OpenAlex API value < 2000
Publication Year Maximum	If *'publication_year'* OpenAlex API value > 2020
Reformulated Abstract Length	If < 5 words

3. Investigations

3.1. Dataset overview and characteristics

The generated data set maintained the trend of increasing retraction counts per year observed in the original Retraction Watch dataset, as illustrated in A.5, A.8 & A.8. When normalised between 0-1, the root mean square error between the generated and original data sets was 0.106, with the similarity shown in Figure 1. Shockingly, within this data set, 7.54% of the articles marked as retracted by the Retraction Watch data set were not marked as retracted by OpenAlex.

For all statistical tests $\alpha = 0.05$. Analysis of correlations between journal features revealed two notable findings:

1. A weak, significant positive correlation between the work count log and the retraction count log (Pearson correlation coefficient 0.065, p-value < 0.05), as shown in Figure A.10.

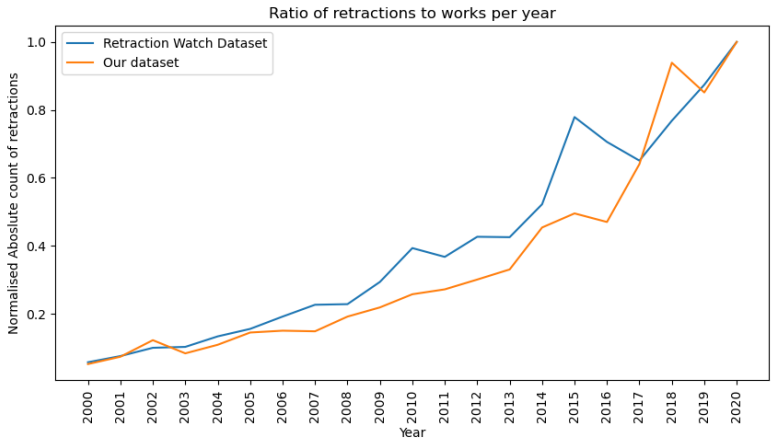


Figure 1: Comparison of our dataset and Retraction Watch Dataset when normalised between 0 and 1

2. A strong negative correlation between the retraction count and log of h-index (Pearson correlation coefficient -0.656, p-value < 0.05), as illustrated in Figure A.11.

3.2. Model Evaluation for Retraction Prediction

Various machine learning models were evaluated to determine if prediction of retraction was possible. The generated data set was used to train the models in the training data set and evaluate them with the test data set. Default settings from the Sklearn package or Huggingface were used unless specified, and random states were set to 42 [17].

The following models were used:

- Multi-layer Perceptron classifier (max_iter = 1000)
- Logistic Regression (max_iter = 1000)
- Decision Tree
- Random Forest
- Support Vector Machine (SVM) (kernel = rbf)
- Gradient Boosting
- XGBoost (n_estimators = 100, learning_rate = 0.1)

- AdaBoost
- BERT (Pretrained model = bert-base-uncased, AdamW optimizer with learning rate 2e-5, fine-tuned for 5 epochs on the training data)
- Llama 3.1 (Pretrained model = unsloth/llama-3-8b-bnb-4bit, AdamW optimizer with learning rate 1e-4, fine-tuned for 2 epoch on training data with early stopping).
- Gemma 2 (Pretrained model = unsloth/gemma-2-9b, AdamW optimizer with learning rate 1e-4, fine-tuned for 2 epoch on training data with early stopping).

Table 5: Model Scores on test dataset: Highest scoring approaches are in bold.

Model	Accuracy	Precision	Recall	f1 score
Gradient Boosting	0.654	**0.691**	0.543	0.608
SVM	**0.668**	0.691	0.595	0.639
XGBoost	0.648	0.669	0.572	0.617
Random Forest	0.644	0.668	0.559	0.609
Llama 3.1*	0.662	0.645	0.708	**0.675**
BERT	0.644	0.639	0.646	0.643
AdaBoost	0.618	0.638	0.529	0.578
Logistic Regression	0.633	0.636	0.605	0.620
MLP	0.620	0.594	**0.729**	0.655
DecisionTree	0.573	0.566	0.590	0.578
Gemma 2*	0.543	0.543	0.477	0.507

*Zero-shot prompting approach used.

Model performance metrics, including precision, recall, and the F1 score, are reported for all models in Table 5 and visualised in Figure 2. The Gradient Boosting model demonstrated the highest precision, with a reported value of 0.691.

3.3. Ablation Study Generation

To assess feature importance in our feature-based classification models, we conducted an ablation study on all input string features (e.g., Title, Date

Published, Abstract). We created data sets for each feature by permuting the data to exclude that feature. We then calculated the average model evaluation metrics (F1 score, precision, recall, accuracy) across all models for each ablation. Lower scoring metrics indicate more contribution to a classifier's performance.

Table 6: Average Ablation Model Scores: lowest scoring ablation are in bold.

Ablation	Accuracy	Precision	Recall	f1 Score
Abstract Inverted Index	**0.629**	0.648	**0.556**	**0.597**
Citated By Count	0.635	0.654	0.565	0.605
First Author	0.633	0.648	0.574	0.607
First Author Countries	0.632	**0.644**	0.590	0.613
Primary Topic	0.633	0.651	0.560	0.601
Publication Date	0.630	0.645	0.570	0.604
Title	0.638	0.652	0.582	0.613

Averaged ablation study results are reported in Table 6 and visualised in Figure 3. Interestingly, within the ablation studies, the averaged lowest-scoring precision ablation was achieved by ablating First Author Countries (Precision 0.644). For all other metrics, the lowest averaged scored metric was achieved through ablating the Abstract (Accuracy 0.629, Recall 0.556, F1 Score 0.597).

3.4. Coefficient Analysis

To assess the importance of individual words in our classification task, we analysed the coefficients of our trained logistic regression model. We loaded the previously trained logistic regression model and the count vectorizer used to preprocess the text data. The feature names (individual words) of the count vectorizer were then extracted. These correspond to the columns in the document-term matrix used to train the model. The coefficients of the logistic regression model were extracted. In logistic regression, these coefficients represent the log-odds impact of each word on the classification decision. A positive coefficient indicates that the presence of the word increases the likelihood of a positive classification, while a negative coefficient decreases

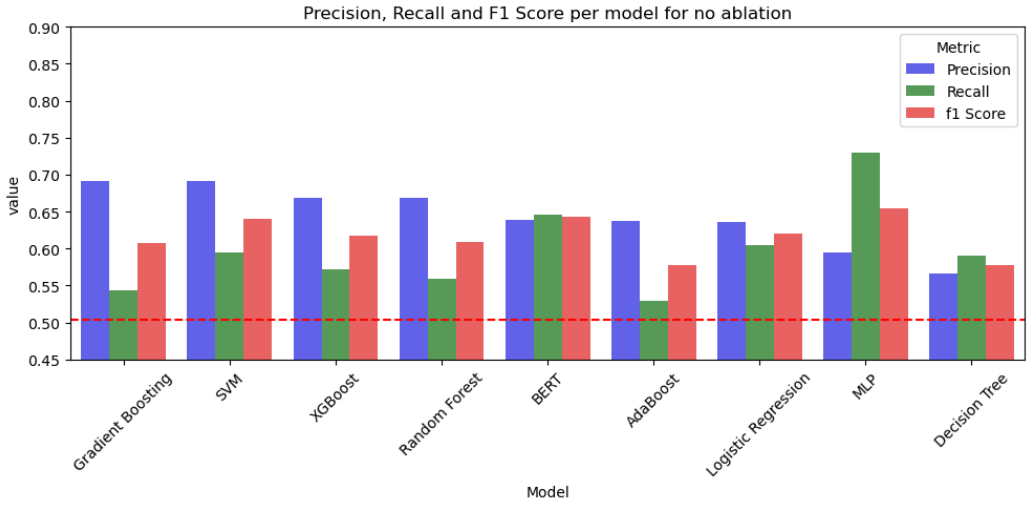


Figure 2: Different evaluation metrics on the test dataset.

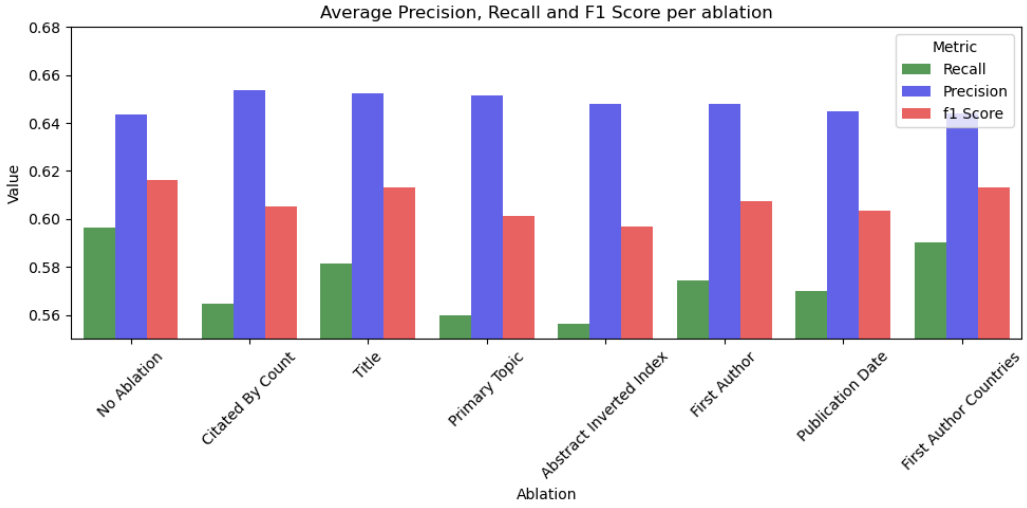


Figure 3: Average model scores per ablation study.

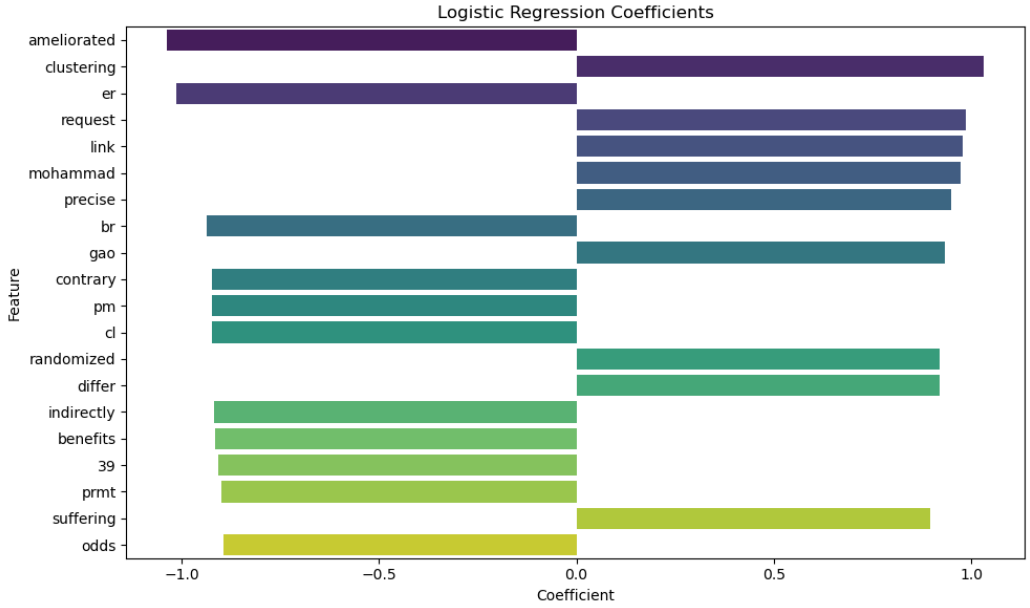


Figure 4: Coefficience visualisation of the linear regression model.

this likelihood. The absolute values of the coefficients were calculated to rank words according to their overall impact, regardless of direction (positive or negative). The features were then sorted in descending order of importance.

The coefficients are visualised in Figure 4.

4. Discussion

4.1. Journal Metadata analysis

Two notable findings were reported from the analysis of the journal metadata: a weak significant positive correlation between a journal's log of work found and the log of retraction count and a strong negative correlation between the journals' retraction count and log of a journal H index. This seems counterintuitive, as more retractions are likely to occur given more publications, and hence, a strong positive correlation would be present. This interesting finding could indicate that journals that publish fewer works are less proactive at detecting potential retractions or that publishing research that will be retracted is more complicated within journals with greater work output, presumably due to increased scrutiny of these works. The strong negative correlation between the retraction count and the log of the h-index

indicates that as a journal's h-index increases, its retraction count tends to decrease significantly, which is expected.

The h-index, a widely used measure of a journal's productivity and impact, is based on its most cited papers. Typically when a work is considered for inclusion in a journal or conference, a peer reviewer is tasked with subjecting that research to the scrutiny of others who are experts in the same field [18]. This reviewer is sourced from academics who review for many, primarily, altruistic reasons, such as keeping up with the latest developments, building associations with journals, and demonstrating a commitment to the scientific field [19]. Importantly, not every researcher is a peer reviewer, which means that available review time is a finite resource [20]. It is likely that more reviwers are available for greater h-index journals. Publishing venues with higher h-index values potentially have a more rigorous peer review process, authors are more diligent when submitting to these journals, or higher-quality journals attract better-quality research.

4.2. Machine Learning For Predicting Retractions

This research demonstrates that machine learning techniques are appropriate for exploring and predicting article retractions. In particular, more traditional feature-based approaches such as gradient boost, SVM, XGBoost, and Random Forest achieved superior precision compared to the more modern contextually aware BERT model. However, this finding needs to be contextualised within the objectives of this investigation itself. The study aimed to establish baseline results for the generated data set rather than to optimise individual model performance. In particular, BERT, as a pre-trained model, typically requires extensive fine-tuning on large, domain-specific datasets to leverage its capabilities, which was not the case here. The limited fine-tuning described in the study (5 epochs on the training data) may have needed to have been sufficient to achieve superior precision performance in this domain. Furthermore, BERT was limited to 512 tokens, which could have truncated important information for this model, given the verbosity of the abstracts. The feature-based classifiers did not have this limitation. Furthermore, while BERT's precision was lower than traditional machine learning models, it demonstrated comparative performance in other metrics, such as recall and the F1 score. This finding could suggest that different model approaches could excel in different evaluation metrics.

Importantly, investigations into appropriate model selection lay the groundwork for future investigations. Optimising the performance of these ap-

proaches, particularly for specific metrics such as precision, remains an open challenge for this data set. Future work could explore more extensive fine-tuning of BERT and more elaborate feature engineering within feature-based classifier models.

4.3. Ablations

Several observations on the ablation of features can be made given the results reported in Table 6 and Figure 3. Unsurprisingly, given the amount of information contained within an abstract, it appears to be the most influential feature among all feature-based models when considering recall or F1 score, as when ablated, it resulted in the lowest average scores for accuracy (0.629), recall (0.556) and F1 score (0.597). This suggests that the abstract contains significant information to identify potentially retracted articles. Interestingly, ablating the "First Author Countries" feature resulted in the lowest precision score (0.644). This indicates that the geographical origin of the first author provides valuable information for precise classification, which supports previous work outlined in the introduction.

4.4. Coefficient Analysis

Certain coefficients (words) were associated with the data set classes (retraction / non-retraction). Although analysis of this is speculative, the author suggests the reasons why certain coefficients were associated with classes as follows:

1. "Randomized" was strongly associated with a paper retraction; this could be due to the increased scrutiny that this type of research is subjected to (such as medical/health domains).
2. "Contrary", which is likely to be seen in papers contradictory to established ideas, is less likely to have been retracted.
3. "Indirectly" and "Benefits" have negative coefficients, which could suggest that more cautious or nuanced claims are less likely to be retracted.
4. The presence of particular names ("Mohammad" and "Gao") could indicate some geographic or cultural factors in retractions, supporting previous research in this area [8].

4.5. Ethics of Automating Retraction Prediction

Using predictive models to identify potential retractions in the scientific literature raises several ethical concerns that warrant careful consideration.

Although these approaches offer promising tools for improving research integrity, they also present significant challenges that current methodologies have not adequately addressed. These concerns also partially explain why precision was the evaluation metric focused on when interpreting the results. A primary concern is that these models rely on correlations rather than causal relationships. This limitation may inadvertently perpetuate existing biases within the publication system. For instance, the coefficient analysis revealed an association between cautious language (e.g., "indirectly," "benefits") and retraction likelihood. However, such correlations do not necessarily imply causation and may lead to misinterpretation of results. Furthermore, the indiscriminate application of these models could potentially hinder the publication of genuinely innovative research that challenges established paradigms. Scientific progress often relies on work that contradicts prevailing doctrines, and we must be cautious not to impede such advancements. An examination of the model coefficients raises concerns about potential unintended consequences of using it's findings. Authors may be incentivised to engage in self-censorship or overly cautious reporting of results to avoid being flagged by these systems. Conversely, bad actors might exploit this knowledge to circumvent detection, potentially facilitating the dissemination of invalid results. Furthermore, implementing these models could introduce inductive bias into investigations, potentially leading to unforeseen consequences in the scientific publishing landscape. This bias can manifest itself in various ways, from shaping research questions to influencing methodological choices.

5. Conclusions

1. This research demonstrates that machine learning approaches can be used with some success to predict if an article is retracted.
2. Feature-based classifiers, such as gradient boosting machines and SVM, outperformed contextual approaches such as BERT.
3. The abstract of a work contains the most important feature for determining if a work is to be retracted (except for precision, where First Author's country is).

Appendix A. Appendix

Table A.7: Ablation Model Scores

Model	Ablation	Precision	Recall	f1 Score
Gradient Boosting	First Author Countries	0.691	0.543	0.608
SVM	First Author Countries	0.691	0.595	0.639
SVM	Citated By Count	0.691	0.595	0.639
SVM	Publication Date	0.689	0.580	0.630
Gradient Boosting	Title	0.688	0.532	0.600
SVM	First Author	0.685	0.591	0.635
Gradient Boosting	First Author	0.685	0.525	0.594
SVM	Primary Topic	0.684	0.597	0.637
Gradient Boosting	Publication Date	0.684	0.532	0.598
XGBoost	Title	0.683	0.570	0.621
Gradient Boosting	Citated By Count	0.682	0.531	0.597
Random Forest	Citated By Count	0.680	0.558	0.613
SVM	Abstract Inverted Index	0.680	0.571	0.621
Random Forest	Abstract Inverted Index	0.679	0.506	0.580
SVM	Title	0.678	0.593	0.632
MLP	Primary Topic	0.675	0.514	0.583
Gradient Boosting	Primary Topic	0.675	0.537	0.598
XGBoost	First Author	0.674	0.562	0.613
Random Forest	Title	0.673	0.567	0.616
Random Forest	Publication Date	0.672	0.559	0.610
XGBoost	Citated By Count	0.671	0.564	0.613
Random Forest	First Author	0.671	0.568	0.615
XGBoost	Primary Topic	0.669	0.562	0.611
Random Forest	Primary Topic	0.669	0.571	0.616
XGBoost	First Author Countries	0.669	0.572	0.617
Random Forest	First Author Countries	0.668	0.559	0.609
XGBoost	Publication Date	0.668	0.549	0.602
XGBoost	Abstract Inverted Index	0.666	0.555	0.605
MLP	Citated By Count	0.665	0.554	0.604
Gradient Boosting	Abstract Inverted Index	0.661	0.509	0.575
AdaBoost	Title	0.653	0.523	0.581
AdaBoost	Abstract Inverted Index	0.645	0.517	0.574
Logistic Regression	Citated By Count	0.639	0.606	0.622
AdaBoost	First Author Countries	0.638	0.529	0.578
Logistic Regression	Abstract Inverted Index	0.637	0.647	0.642
Logistic Regression	First Author Countries	0.636	0.605	0.620
AdaBoost	Citated By Count	0.635	0.528	0.576
AdaBoost	First Author	0.634	0.531	0.578
Logistic Regression	Primary Topic	0.632	0.606	0.619
AdaBoost	Publication Date	0.632	0.516	0.568
MLP	Abstract Inverted Index	0.631	0.597	0.613
Logistic Regression	First Author	0.629	0.606	0.618
AdaBoost	Primary Topic	0.629	0.514	0.566
MLP	Title	0.627	0.633	0.630
Logistic Regression	Title	0.626	0.616	0.621
Logistic Regression	Publication Date	0.625	0.603	0.614
MLP	First Author	0.623	0.625	0.624
MLP	Publication Date	0.618	0.644	0.631
MLP	First Author Countries	0.594	0.729	0.655
Decision Tree	Title	0.592	0.618	0.605
Decision Tree	Abstract Inverted Index	0.585	0.547	0.565
Decision Tree	First Author	0.581	0.587	0.584
Decision Tree	Primary Topic	0.579	0.578	0.578
Decision Tree	Publication Date	0.573	0.578	0.576
Decision Tree	Citated By Count	0.569	0.582	0.575
Decision Tree	First Author Countries	0.566	0.590	0.578

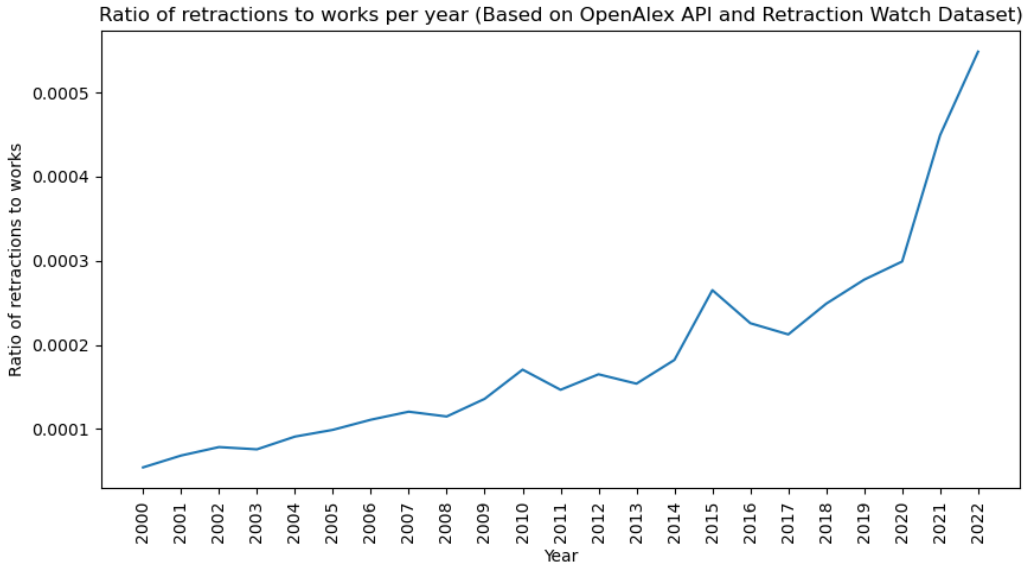


Figure A.5: Ratio of retractions to works: A positive increase is noted over the time period.

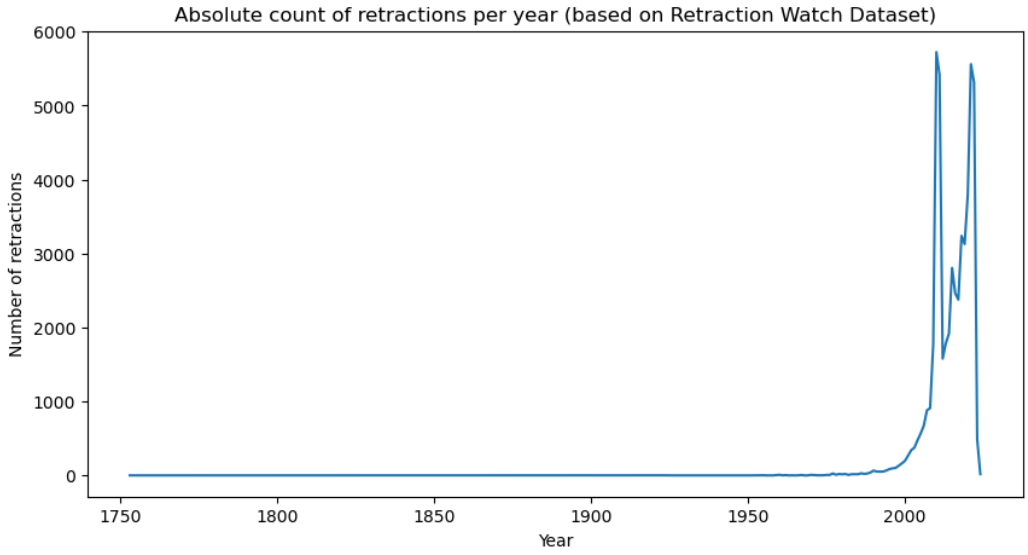


Figure A.6: Absolute count of retractions per year. Demonstrating that the absolute count of retractions has increased over the past 20 years.

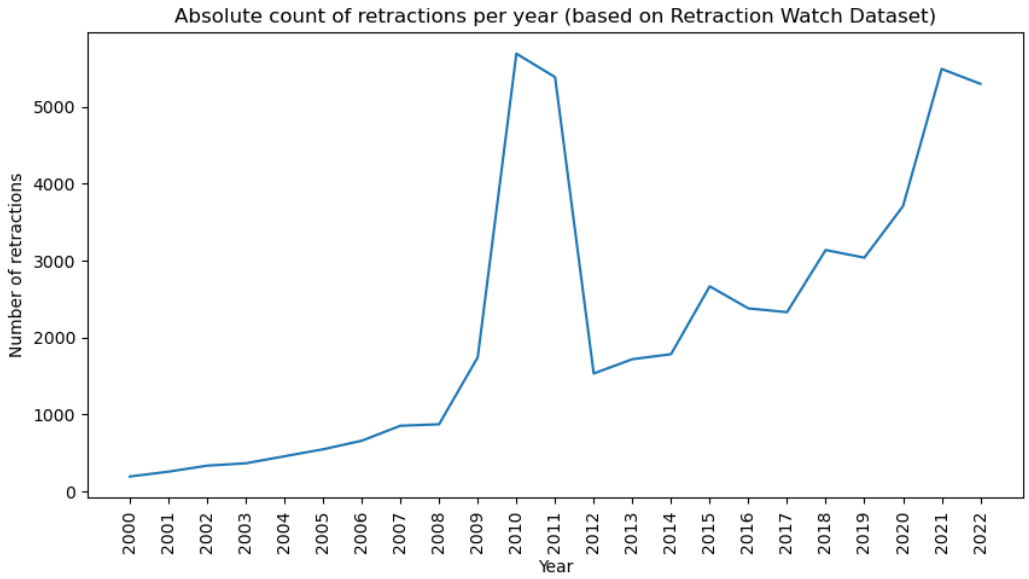


Figure A.7: Absolute count of retractions per year between 2000 and 2020. Note that peak of retractions around the 2010-2012 period, potentially due to the more than 8000 IEEE conference papers that were retracted in 2009-2011

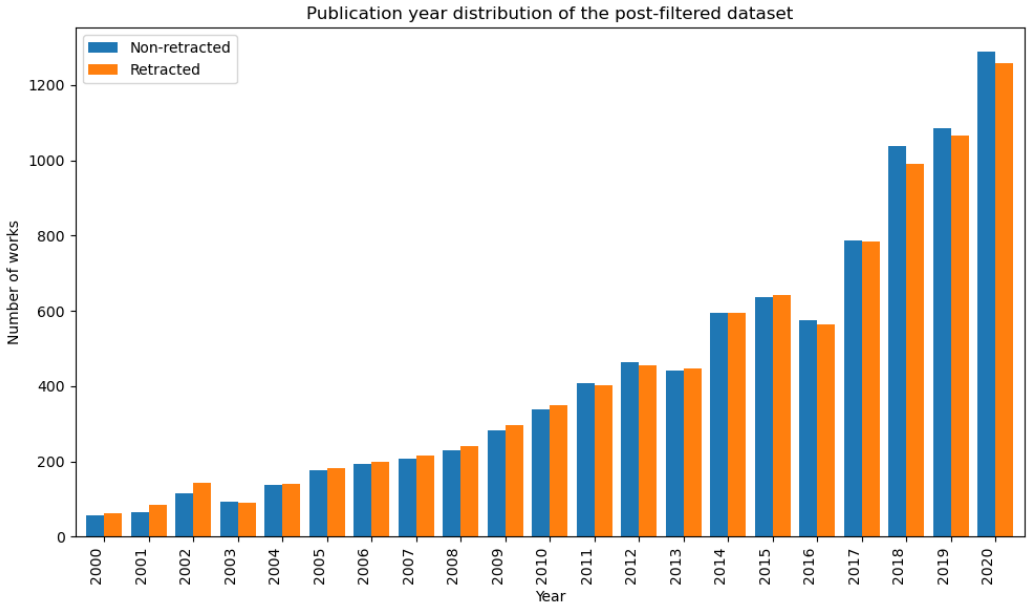


Figure A.8: Dataset publication parity: Publication year distribution for retracted and non-retracted works.

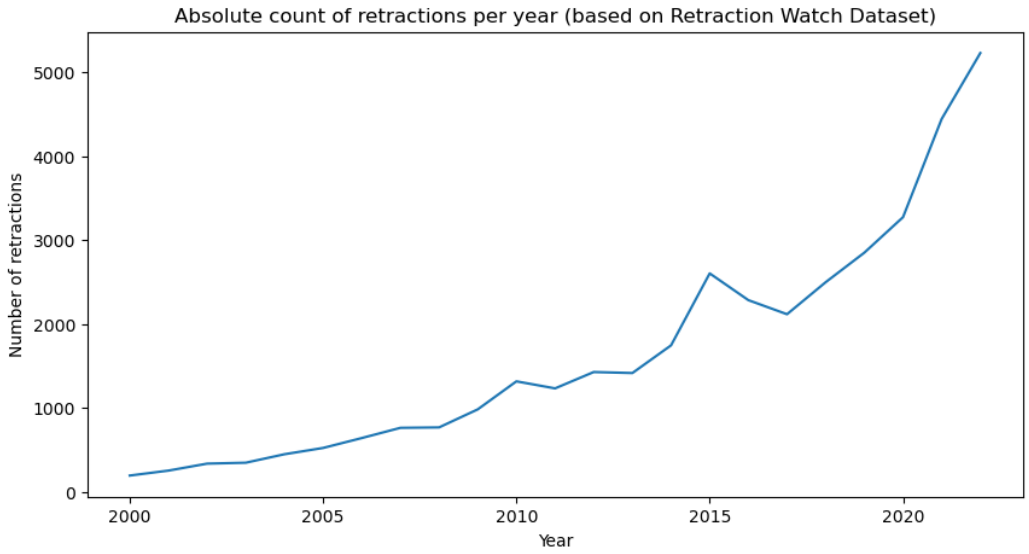


Figure A.9: Ratio of retracted works normalised by total works per year, after removing conference papers: An increasing trend for retractions.

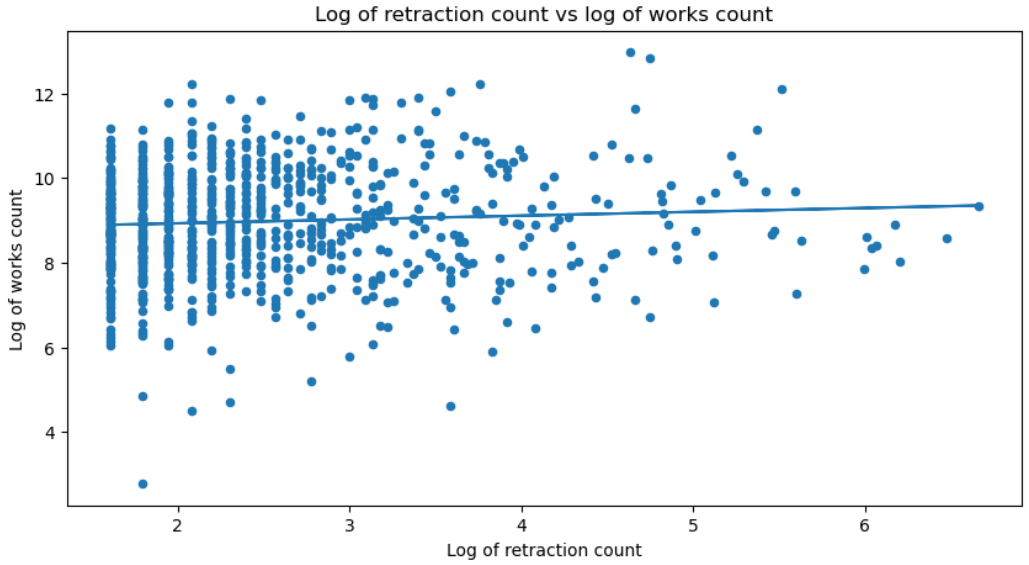


Figure A.10: A weak positive correlation between the log of works counts and log of retraction counts.

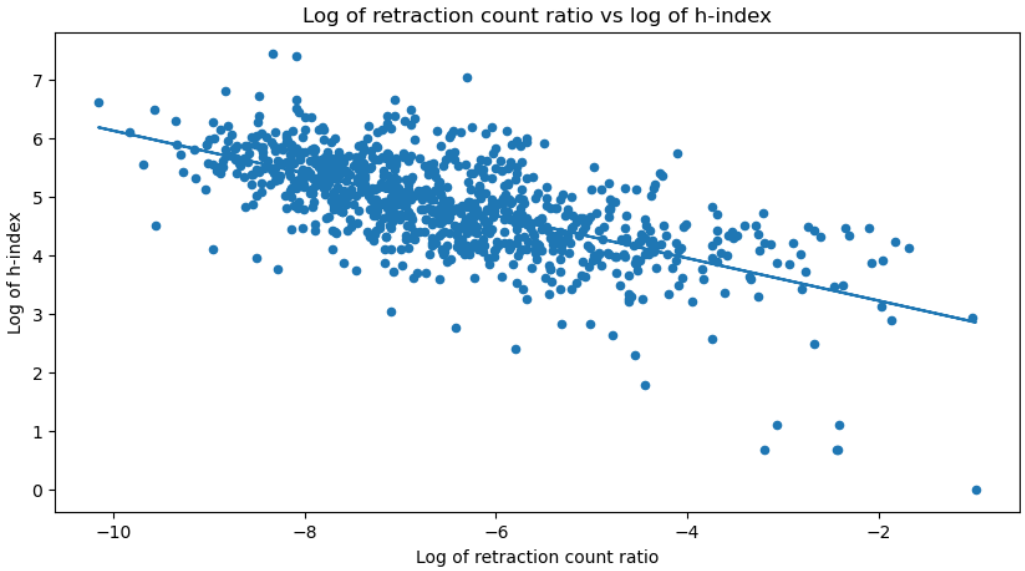


Figure A.11: A strong negative correlation between lof of h index and log of retraction counts.

References

[1] R. G. Steen, Retractions in the scientific literature: is the incidence of research fraud increasing?, Journal of Medical Ethics 37 (4) (2011) 249–253. doi:10.1136/jme.2010.040923.

[2] R. G. Steen, Retractions in the scientific literature: do authors deliberately commit research fraud?, Journal of Medical Ethics 37 (2) (2011) 113–117. doi:10.1136/jme.2010.038125.

[3] R. Perera, P. Nand, Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature, COMPUTING AND INFORMATICS 36 (1) (2017) 1–32, number: 1. doi:https://doi.org/10.4149/cai_2017_1_1.
URL https://www.cai.sk/ojs/index.php/cai/article/view/2017_1_1

[4] J. Dan, M. James H., Speech and Language Processing (3rd ed. draft), in: Speech and Language Processing (3rd ed. draft), Stanford University, 2024, pp. 1–23.

URL https://web.stanford.edu/~jurafsky/slp3/slides/4_NB_2024.pdf

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep Learning–based Text Classification: A Comprehensive Review, ACM Comput. Surv. 54 (3) (2021) 62:1–62:40. doi:10.1145/3439726.
URL https://doi.org/10.1145/3439726

[6] E. C. Moylan, M. K. Kowalczuk, Why articles are retracted: a retrospective cross-sectional study of retraction notices at BioMed Central, BMJ Open 6 (11) (2016) e012047, publisher: British Medical Journal Publishing Group Section: Ethics. doi:10.1136/bmjopen-2016-012047.
URL https://bmjopen.bmj.com/content/6/11/e012047

[7] E. R. Babbie, The Practice of Social Research, Cengage AU, 2020, google-Books-ID: KrGeygEACAAJ.

[8] S. Stretton, N. J. Bramich, J. R. Keys, J. A. Monk, J. A. Ely, C. Haley, M. J. Woolley, K. L. Woolley, Publication misconduct and plagiarism retractions: a systematic, retrospective study, Current Medical Research and Opinion 28 (10) (2012) 1575–1583. doi:10.1185/03007995.2012.728131.

[9] C. Candal-Pedreira, J. S. Ross, A. Ruano-Ravina, D. S. Egilman, E. Fernndez, M. Prez-Ros, Retracted papers originating from paper mills: cross sectional study, BMJ 379 (2022) e071517, publisher: British Medical Journal Publishing Group Section: Research. doi:10.1136/bmj-2022-071517.
URL https://www.bmj.com/content/379/bmj-2022-071517

[10] M. Gaudino, N. B. Robinson, K. Audisio, M. Rahouma, U. Benedetto, P. Kurlansky, S. E. Fremes, Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020, JAMA Internal Medicine 181 (8) (2021) 1118–1121. doi:10.1001/jamainternmed.2021.1807.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8111562/

[11] R. Watch, Retraction Watch Database User Guide (Jun. 2024).
URL https://retractionwatch.com/wp-content/uploads/2023/12/Building-The-Database.pdf

[12] R. Watch, Retraction Watch (Jun. 2024).
URL https://retractionwatch.com/

[13] J. A. Teixeira da Silva, Silent or Stealth Retractions, the Dangerous Voices of the Unknown, Deleted Literature, Publishing Research Quarterly 32 (1) (2016) 44–53. doi:10.1007/s12109-015-9439-y.
URL https://doi.org/10.1007/s12109-015-9439-y

[14] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, arXiv:2205.01833 [cs] (Jun. 2022). doi:10.48550/arXiv.2205.01833.
URL http://arxiv.org/abs/2205.01833

[15] R. Van Noorden, More than 10,000 research papers were retracted in 2023 a new record, Nature 624 (7992) (2023) 479–481, bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Scientific community, Publishing. doi:10.1038/d41586-023-03974-8.
URL https://www.nature.com/articles/d41586-023-03974-8

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (May 2019). doi:10.48550/arXiv.1810.04805.
URL http://arxiv.org/abs/1810.04805

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, . Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (85) (2011) 2825–2830.
URL http://jmlr.org/papers/v12/pedregosa11a.html

[18] D. Banks, Thoughts on Publishing the Research Article over the Centuries, Publications 6 (1) (2018) 10, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/publications6010010.
URL https://www.mdpi.com/2304-6775/6/1/10

[19] P. J. Steer, S. Ernst, Peer review - Why, when and how, International Journal of Cardiology Congenital Heart Disease 2 (2021) 100083, publisher: Elsevier. doi:10.1016/j.ijcchd.2021.100083.

URL https://www.sciencedirect.com/science/article/pii/S2666668521000070

[20] V. Warne, Rewarding reviewers sense or sensibility? A Wiley study explained, Learned Publishing 29 (1) (2016) 41–50, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/leap.1002. doi:10.1002/leap.1002.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1002

B ENGAGEMENT WITH DPP ACTIVITIES: COM61003

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

COM61003: Introduction to Responsible SLT Leadership ePortfolio

Student Name	Aaron Fletcher
Student Registration	230116573
Cohort Number	5
Academic Year	2023/24

Contents

Ethical research practice

Epigeum Completion Certificate

epigeum

Certificate

Number: 1991230383

This is to certify that

Aaron Fletcher
of University of Sheffield

Successfully completed the course

Research Integrity: Core

Good research conduct	100%
Irresponsible research practices	90%
Planning your research	80%
Managing and recording your research	80%
Data selection, analysis and presentation	100%

as part of the Epigeum Online Course System with a score of 88%.

Dated: 12 October 2023

epigeum

Certificate

Number: 1991230383

This is to certify that

Aaron Fletcher

of University of Sheffield

Successfully completed the course

Research Integrity: Core

Scholarly publication	80%
Professional responsibilities	90%
Communication, social responsibility and impact	90%

as part of the Epigeum Online Course System with a score of 88%.

Dated: 12 October 2023

Responsible Research and Innovation (RRI) Guest Lectures

1. *Summary of each lecture attended (max 1 page). Include date, title, and presenter.*
2. *Evidence of participation and completion of ORBIT training*

Add additional tables as required.

Lecture 1

Title	Equality, Diversity and Inclusion (EDI) webinar – the role of EDI in enhancing research and innovation within a competitive engineering environment.
Speaker	Professor Jim McLaughlin
Organisation	Director of the Nanotechnology and Integrated Bioengineering Centre
Date	29/9/2023
Narrative	The presentation was a 1-hour online talk about equality, diversity, and inclusion. This was an external lecture (The lecturer was from the University of Belfast). There were some short technical difficulties when starting the lecture, which resulted in poor lecture delivery, chiefly intermittent notification pings from the presenter's computer and slides not advancing correctly. This detracted from the presenter's ability to convey the lecture's important content. The online lecture format did not allow authentic audience interaction, so I could not ask questions. The presenter started with a short introduction outlining what would be discussed, which I found to be a good way to determine quickly if this lecture suited the audience. The lecturer used figures and relevant research well - each slide point had sufficient referencing. Notably, 90.7% of engineers were male, and 7.8% were BAME. As part of a subgroup that EDI affects, more could have been done to highlight some of the issues these communities face. Real-world examples and case studies could have strengthened arguments. Additionally, I wanted to focus more on the lecture title (*'Enhancing Research and Innovation within a Competitive Engineering Environment'*) as little was discussed about what precisely a competitive environment is and how EDI will help improve that. Ultimately, this lecture was designed to be a starting point for EDI within a research setting, with clear definitions provided, sometimes in excess, which hampered the development of more critical insights. It was great at outlining the key sectoral equality issues and the fact that if you measure EDI issues based on. who is working in your department, you don't necessarily understand the barriers which prevented people from working there in the first place. This lecture's relevance to me will likely be in two main areas: ensuring my eventual PhD topic addresses and overcomes any real or potential EDI issues and my role in promoting research roles for other EDI-affected persons. Take-home messages: ● In-person format preferred for giving lectures ● Overuse of definitions can detract from the message. ● Ensure technical issues are resolved before starting.

	• Use real-world examples to really strengthen arguments.

Lecture 2

Title	Should a Robot Speak (if so, why, when and how)?
Speaker	Professor Roger Moore, Professor of Spoken Language Processing
Organisation	Department of Computer Science
Date	17/10/2023
Narrative	The presentation was a 1-hour in-person lecture on human-robot interaction by a world-renowned specialist. The premise of the lecture was that researchers assume that interacting with a robot via speech is natural. Spoken language is the most complex form of communication we can see, and it has multiple components, not just automatic speech recognition systems, which need to be considered when creating these systems. Roger presented an approach to understanding how language is perceived, such as pragmatics (the why did someone say this based on their environment, energetics (the motivational context, i.e. shouting) and Synhronics (such as directing attention or coordinating behaviours). I previously had started to consider that speech does not follow the traditional view (stimulus-response) and is much more complex; however, this lecture has formalised that approach - the coupled response. He further explained that speech is more than just a single channel, as we utilise many questions to understand speech. He presented that each agent in a communication scenario has individual priors (i.e. a set of beliefs about the other agent in the communication), such as individuality, personality, demographics, and culture, which matter to the context. As ASR systems do not have those, and we as agents expect those things, this generates a mismatch of expectations and leads to confusion by the human agents. The question is, can ASR systems be created that approximate human agents so much that this mismatch of expectations can be bridged, or is it not possible? Novel concepts - such as the habitability gap (a mismatch between capabilities and expectations of the human users) and likened it to the uncanny valley. Take away points: • Consideration of the end user is important when designing services • Approximating humans might not be possible • A holistic approach is needed when designing these systems.

Lecture 3

Title	AI Technologies for Decision-Making: Challenges
Speaker	Prof Mirco Musolesi
Organisation	Department of Computer Science at University College London
Date	18/10/2023
Narrative	This hour-long online lecture was themed around the responsible use of AI technologies in decision-making scenarios, such as the economy, medical or other. Decision-making systems are used in many social scenarios, such as conflict resolution and peace maintenance, with the potential risk of applying technology to this domain resulting in an impact on humans. One cited example is that the next financial crash is likely to emerge from AI. The talk raised points such as these systems that don't have a vested interest in the outcome of these decisions (i.e., little ability to assess the human impact on the decisions they are generating). The talk delved into assumptions made by AI creators, such as: ● Biological - the brain processes information using biologically equivalent on/off switches. ● Psychological assumption - that the brain can be viewed as a device operating on a known rule set. ● Epistemological assumption - All knowledge can be formalised. ● Ontological Assumption - the world consists of independent facts that can be formalised with precision. The speaker outlined the decision-making process in machines and how it is one of the biggest problem areas, with generalised concepts such as machine autonomy, keeping humans in and out of the loop. Ultimately, decisions cannot be fair or equal unless the system has a stake in the outcome. Relevance to me: ● Remember to look for unintended consequences of any code produced. ● Assessment of this risk needs to be continuous.

Lecture 4

Title	Open Research Conversation: Sustainability and open data: Balancing environmental concerns
Speaker(s)	Tom Webb Chris Olga
Organisation	Department of Bioscience at University of Sheffield School of Advanced Studies University of London
Date	21/2/2024
Narrative	This lecture is part of Love Data week 2024. Marine biology has a large amount of data available to it now, allowing it to answer more questions then possibly before. You can ask questions such as if the 2023 heatwave caused issued for marine animals. Tom has noticed that lack of data in marine biology is not a large concern as the data is readily available. They have developed a process for this, such as creating an occurrence record, which can be combined with other measures, such as temperature records for that area, so you can compare occurrence records, with temperatures to see if there is any correlation. Different information on taxonomy and the characteristics of marine animals available. Loads of environment data available, typically through bio Oracle. Speaker talks about combining different sources of open source data to answer questions such as did the heat wave affect marine biodiversity. Second speaker talked about their role from a humanities background to data science, stating that humanity subjects typically have no open datasets. Vast majority of literally scholars have to use data closed off by paywalls. Developed a toolkit that was developed, primarily aimed to remind people that digital work does have an impact on climate change. While IT is not the most significant problem with greenhouse gases, that was checked in 2021, and it likely to increase as AI developed. Transport contributes to 14% of green house gases, so need to consider when travelling to conferences etc. Toolkit is aimed at how to decarbonize aspects of their work, how to change systems within the university, carbon literacy etc. Relevance to me: ● Need to consider the environmental impact when designing these sytems ● Should take steps to minimise technological debt.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Lecture 5

Title	Social Media and Ethics
Speaker(s)	Nicolas Gold, UCL Computer Science
Organisation	UKRIO
Date	21/2/2024
Narrative	This 1 hour webinar lecture is part of UKRI online free webinar series, which looked at he the ethics issues for public data and providing researchers with an ethical defense for their principles and practice. The speaker was very clear that this was not a place to garner information regarding legalities or specific studies. Anology of speaking a coffee shop was used, where conversations within a public space might be overheard. However, do the customers in the coffee shop consent to their conversations being listened to. People often have issues with this if their conversations are used for research into AI etc. For example, youtube users were suprised to be included in AI research (Hu 2019). It could be that the subjects, social media users, have a poor understnding of publically availabluty and implications so you need to consider the types of data, and risks presented to participants. Informative statistics were presented by the speaker: On twitter, 61% of twitter users were not aware of their research use and 65% of people think that we shouldn't use tweets without permission. This shows a user desire for information about how their data is being used The speaker raised the question of if people trusted the social media platforms. Short answer no, and they do not come out well for trust in their data. A study concluded users wanted meaningful transparency and we need to try to meet the expectations of the people's who data we use Discuss on Public vs Private Data occurred, again with an analogy being given. In this foot path analogy, land owned by gatekeepers (such as employers for employers, Headteachers for school, Charities for service users, Social media companies for their users) and we as researchers have to ask permission from the landowners to user their data. He raised issues of expectation, consent, contract, regulation and access to data, and outlined how we should respect the stakeholders involved Implied consent argument - If the service clearly explains the public disposition of that data and the available controls to a user, then the user gives implied constent when they post data. Relevance to me: ● I will likely be engaged in some social media work during my PhD period and need to consider how to ethically best use this.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Evidence of ORBIT Participation

ORBIT

Certificate of Attendance

This acknowledges that

Aaron Fletcher

has successfully completed the course

Foundations Course in Responsible Research and Innovation

on the 11/06/2024

MARTIN DEHEAVER, DIRECTOR

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Orbit PhD Technology Assessment report

RRI Maturity Report: Technology-assisted review within the medical domain
Name of Assessor: Aaron Fletcher
Date of Assessment: 13/06/2024

Use Case
This PhD focuses on improving the time it takes to undertake a systematic review within the medical domain. A systematic review is a methodology for searching and appraising data for a given medical topic that is repeatable, reproducible, and considers all available evidence. Specifically, I will focus on information retrieval within the systematic review process and use datasets, such as the ones provided by Cochrane and Clef, to see if we can improve the recall of these methods by integrating large language models within the continuous active learning process. The use case for this evidence will be by clinicians who are undertaking these systematic reviews.

The Sustainable Development Goals Affected by this Project are...

Unquantified Positive Impact

The provision of good healthcare depends on synthesising available evidence into a format that encompasses all available data. If information is not retrieved that would have affected the outcome of a systematic review, this could lead to the provision of worse quality services.

Unquantified Positive Impact

Systematic Review information is used to teach a range of professionals, from students to updating seasons practitioners of medicine.

Innovation Potential
Based on your assessment of the SDGs affected by your technology and the following table, the Innovation Potential is 3

This is because your project impacted 2 Sustainable Development Goals. 0 of these were quantified, and 2 were unquantified.

Number of SDGs affected by this technology.	Type of Impact		
	No Impact	Unquantified	Quantified
0	Innovation Potential 1		
1		Innovation Potential 2	Innovation Potential 4
2 or more		Innovation Potential 3	Innovation Potential 5

Technology Readiness Level

The target technology readiness level for your project is :

- TRL 1 : Basic Research. Principles postulated and observed but no experimental proof available

Calculation of RRI Intensity Level

Based upon the Innovation Potential and the Target Technology Readiness Level for your project, the calculated RRI Intensity Level, based on the table below, for this project is:

- TRL 1 : Basic Research. Principles postulated and observed but no experimental proof is available

	IP 5	IP 4	IP 3	IP 2	IP 1
TRL 1	RIL 5	RIL 2	RIL 2	RIL 1	RIL 1
TRL 2	RIL 5	RIL 2	RIL 2	RIL 1	RIL 1
TRL 3	RIL 5	RIL 3	RIL 3	RIL 2	RIL 2
TRL 4	RIL 5	RIL 3	RIL 3	RIL 2	RIL 2
TRL 5	RIL 5	RIL 4	RIL 4	RIL 3	RIL 2
TRL 6	RIL 5	RIL 4	RIL 4	RIL 3	RIL 2
TRL 7	RIL 5	RIL 4	RIL 4	RIL 4	RIL 2
TRL 8	RIL 5	RIL 4	RIL 4	RIL 4	RIL 2
TRL 9	RIL 5	RIL 5	RIL 5	RIL 4	RIL 2

Once the RRI Intensity level required for a particular project has been established by a research actor, be they prospective researcher or funding body, the question then comes as to what type of research project organisation has the right characteristics to deliver such an initiative. Clearly this evaluation covers a very wide range of parameters associated with research excellence etc, but we propose, that these parameters be extended to include due consideration of the RRI Maturity level of the organisation.

When combining these levels of RRI maturity with the categories and components of RRI, we arrive at a way of representing the RRI maturity model in the following form. We would like to underline that the differences between levels are not as clear-cut as the representation above may suggest.

Calculation of RRI Intensity Level from SDG Impact and Target TRL

Based upon the Innovation Potential and the Target Technology Readiness Level for your project, the calculated RRI Intensity Level, based on the table below, for this Project is 2

Level Five	Strategic	Organisation has adopted RRI as a component of its strategic framework and aims to ensure all R&D activities cover all (or most of) RRI components.
Level Four	Proactive	Organisation realises the benefits of RRI and seeks to proactively and increasingly integrate these into its business processes.
Level Three	Defined	Organisation has a definition of (components of) RRI and has integrated these into its business processes.
Level Two	Exploratory / Reactive	Organisation reacts to external pressures concerning aspects of RRI and experiments concerning appropriate processes.
Level One	Unaware	Organisation is not aware of RRI or its components and does not incorporate it into its processes.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Personal Development Project (PDP)

Description of your PDP and plan for what needs to be done to realise the PDP project and how it will be carried out over the following 3 years. In addition to your ideas, target communities, goals, etc don't forget to include information about:

- *Activities*
- *Timings*
- *Feasibility and risk*
- *RRI analysis*
- *Partners needed (if applicable)*
- *Funding needed (if applicable)*

(3-5 pages)

PDP proposal - Draft

Entrepreneurship and Business Guest Lectures

Summary of each lecture attended (max 1 page). Include date, title, and presenter.
Add additional tables as required.

Lecture 1

Title	Early Career Research Through Commercialisation
Speaker	Various: Laura Talboters, Joe Carruth, Ryan Bramley
Organisation	University of Sheffield: Research, Partnerships & Innovation
Date	23/11/2023
Narrative	This was a 1 ½ hour online lecture with contrasting information to case studies. The first lecture, given by Laura, was on how to apply research to commercial opportunities. She outlined previous schemes that have used CIP, which is no longer funded, and how it directly affects the current scheme (RISK) and why they believe it is better. She outlined what is available from the university, focusing on the available leadership skills/mentoring. Depending on the requirements, funding was available for ten projects, lasting 5 - 57 days. She focused on the impact of these projects and why they are a valid career choice for PGR / early researchers. This led very well into the two case studies that were presented next. The first one, on contextualising captions for deaf users, was compellingly presented. They outlined their motivation, how the scheme (CIP) had helped them realise that and the tangible benefits that it had had for them (an ongoing researcher job with the university) and their target population (changes in how deaf subtitles were processed, and involvement of deaf people within their production, and evidence provided for improvement in legislation). The following case study could have been presented better. No slide deck was available for this, and the result was an incoherent presentation that needed to show me why CIP had been integral to creating their business. The best case study, given by the Phlux owner, was clear and concise. It showed their story of creating a material which had potential applications, how they raised funds to survey the market through a funding body, was given mentors and support from Sheffield University in this process and how they went on to develop a product which has market value, protects UOS IP and also can feed back into further research. This set of lectures is an excellent reference for myself, and it has increased my knowledge and enthusiasm for commercial opportunities.

Lecture 2

Title	Decoding value in business
Speaker	Speaker: Klaas Molapisi Chair: Rea Nkhumise
Organisation	PwC South Africa
Date	24/1/2024
Narrative	I attended an online virtual meeting that lasted for about 1 ½ hours on applying value to business. The focus of the meeting was on entrepreneurs, and the presenter and people within the group gave introductions. Instead of discussing the definition of value, the conversation revolved around what value means to society and how it relates to business. The group talked about the principles or standards of behaviour that guide business and how business is a form of trade that should be focused on providing value. The discussion centred around the idea that values are what are applied to a business and how Apple provides value by making their products easy to use for creative purposes. The group also talked about how working from a point of value can reduce corruption, grow ethics, and create a fair-trade market. There are three levels of values, ranging from corruption (the worst place to be), capitalism (more of a concern), to sustainability (the best place to be). The main takeaway from the meeting was that businesses should prioritise working from a point of value rather than serving the investor at the cost of their values. When we have values, we establish clear principles, hold ourselves accountable, develop products from the point of impact, and create sustainable outcomes. Value doesn't necessarily create profitability, but it creates sustainability, ultimately what matters the most.

Lecture 3

Title	AI Technologies for Decision-Making: Challenges
Speaker	Prof Mirco Musolesi
Organisation	Department of Computer Science at University College London
Date	18/10/2023
Narrative	This hour-long online lecture was themed around the responsible use of AI technologies in decision-making scenarios, such as the economy, medical or other. Decision-making systems are used in many social scenarios, such as conflict resolution and peace maintenance, with the potential risk of applying technology to this domain resulting in an impact on humans. One cited example is that the next financial crash is likely to emerge from AI. The talk raised points such as these systems that don't have a vested interest in the outcome of these decisions (i.e., little ability to assess the human impact on the decisions they are generating). The talk delved into assumptions made by AI creators, such as: ● Biological - the brain processes information using biologically equivalent on/off switches. ● Psychological assumption - that the brain can be viewed as a device operating on a known rule set. ● Epistemological assumption - All knowledge can be formalised. ● Ontological Assumption - the world consists of independent facts that can be formalised with precision. The speaker outlined the decision-making process in machines and how it is one of the biggest problem areas, with generalised concepts such as machine autonomy, keeping humans in and out of the loop. Ultimately, decisions cannot be fair or equal unless the system has a stake in the outcome. Relevance to me: ● Remember to look for unintended consequences of any code produced. ● Assessment of this risk needs to be continuous.

Lecture 4

Title	Open Research Conversation: Sustainability and open data: Balancing environmental concerns
Speaker(s)	Tom Webb Chris Olga
Organisation	Department of Bioscience at University of Sheffield School of Advanced Studies University of London
Date	21/2/2024
Narrative	This lecture is part of Love Data week 2024. Marine biology has a large amount of data available to it now, allowing it to answer more questions then possibly before. You can ask questions such as if the 2023 heatwave caused issued for marine animals. Tom has noticed that lack of data in marine biology is not a large concern as the data is readily available. They have developed a process for this, such as creating an occurrence record, which can be combined with other measures, such as temperature records for that area, so you can compare occurrence records, with temperatures to see if there is any correlation. Different information on taxonomy and the characteristics of marine animals available. Loads of environment data available, typically through bio Oracle. Speaker talks about combining different sources of open source data to answer questions such as did the heat wave affect marine biodiversity. Second speaker talked about their role from a humanities background to data science, stating that humanity subjects typically have no open datasets. Vast majority of literally scholars have to use data closed off by paywalls. Developed a toolkit that was developed, primarily aimed to remind people that digital work does have an impact on climate change. While IT is not the most significant problem with greenhouse gases, that was checked in 2021, and it likely to increase as AI developed. Transport contributes to 14% of green house gases, so need to consider when travelling to conferences etc. Toolkit is aimed at how to decarbonize aspects of their work, how to change systems within the university, carbon literacy etc. Relevance to me: ● Need to consider the environmental impact when designing these sytems ● Should take steps to minimise technological debt.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Lecture 5

Title	Social Media and Ethics
Speaker(s)	Nicolas Gold, UCL Computer Science
Organisation	UKRIO
Date	21/2/2024
Narrative	This 1 hour webinar lecture is part of UKRI online free webinar series, which looked at he the ethics issues for public data and providing researchers with an ethical defense for their principles and practice. The speaker was very clear that this was not a place to garner information regarding legalities or specific studies. Anology of speaking a coffee shop was used, where conversations within a public space might be overheard. However, do the customers in the coffee shop consent to their conversations being listened to. People often have issues with this if their conversations are used for research into AI etc. For example, youtube users were suprised to be included in AI research (Hu 2019). It could be that the subjects, social media users, have a poor understnding of publically availabluty and implications so you need to consider the types of data, and risks presented to participants. Informative statistics were presented by the speaker: On twitter, 61% of twitter users were not aware of their research use and 65% of people think that we shouldn't use tweets without permission. This shows a user desire for information about how their data is being used The speaker raised the question of if people trusted the social media platforms. Short answer no, and they do not come out well for trust in their data. A study concluded users wanted meaningful transparency and we need to try to meet the expectations of the people's who data we use Discuss on Public vs Private Data occurred, again with an analogy being given. In this foot path analogy, land owned by gatekeepers (such as employers for employers, Headteachers for school, Charities for service users, Social media companies for their users) and we as researchers have to ask permission from the landowners to user their data. He raised issues of expectation, consent, contract, regulation and access to data, and outlined how we should respect the stakeholders involved Implied consent argument - If the service clearly explains the public disposition of that data and the available controls to a user, then the user gives implied constent when they post data. Relevance to me: ● I will likely be engaged in some social media work during my PhD period and need to consider how to ethically best use this.

Academic CDT / SpandH / NLP Seminar Series

1. *Summary of each seminar (max 1 page) including analysis of presentation quality (strengths & weaknesses). Include date, title, and presenter.*
2. *Details of when you presented your research topic as part of the main seminar series, CDT conference, or similar.*
   Include a copy of the slides or poster used.

Add additional tables as required.

Seminar 1 - NLP

Title	Language Resources and NLP Applications for Portuguese
Speaker	Professor Viviane Moreira
Organisation	Institute of Informatics at UFRGS
Date	27/9/2023
Narrative	This one-and-a-half-hour lecture covered two main topics: The professor's academic background and research areas (1 hour) and the explanation of the ARR (30 minutes). The lecture started with a clear justification of why Portuguese presents a particular challenge for NLP analysis - chiefly that while Portuguese is the 6th largest language among native speakers, it only represents 2% of online web content. In comparison, English is spoken by 4% of the world and is represented by 55% of online web content. The speaker briefly discussed their and their students' work, covering various topics. However, due to time constraints, there wasn't an in-depth discussion about any of the research. As a result, the presentation felt like a promotion of the department and students rather than a focused event. The abstract had suggested a more in-depth discussion, but this was not delivered. The second part of the lecture, the ARR, was an overview of the submission and review process for various top-tier conferences. While I currently have no plans to submit, the information presented will undoubtedly become useful as I start to want to present at these conferences. As NLP is my preferred field, it was interesting to get an overview of some current research topics; however, I gained little additional knowledge from this seminar about NLP. The speaker was an L2 English speaker. However, they presented very clearly and concisely, struggling with enunciation with few words. The presentation style did, however, detract from conveying information, as the speaker's over-reliance on visual prompts distracted from communication. Main takeaway points: ● Seminars with multiple students' research are inefficient for knowledge gathering. ● The submission process should be investigated closer to conference applications.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Seminar 2 - NLP

Title	What deep nets can & can't do with language and why
Speaker	Prof. Robert Berwick
Organisation	MIT
Date	4/10/2023
Narrative	The presentation was a 1-hour talk in-person/online lecture about how deep neural nets currently understand language and why they can't perform like humans on this task. This was an external lecture (The lecturer was from MIT and was being held as part of the workshop on biologically inspired models of language). The presentation style was one of the best I have seen (so far), particularly with the presenter's use of thorough examples and slide design stimulating active engagement. At no point did I feel that the lecture was repetitive and that the presenter added to the discussion about slide material rather than just repeating it. This lecture targeted post-graduates without defining basic terms and generally accepting that participants understood the domain (i.e., what a transformer is, and what a deep neural network is). This was appropriate, given the audience, which allowed a greater depth of material to be produced. The first compelling argument presented was that deep neural networks cannot generalise if it is not within the training data, while children can. The example given was GoogleNet's inability to generalise from full-colour photographs to black and white drawings, while children can. As deep neural networks are essentially remembering a cloud of data, anything outside of that data cloud, such as black and white drawings, cannot be generalised to, and such, there exists two solutions: More training data to cover all these eventualities (in which case, is this general intelligence?) or that the neural networks are not the solution to human-equal intelligence. One central argument by the presenter was that deep neural networks, such as chatGPT, do not universally approximate functions. This initially seemed bold, given that it is widely taught. The presenter presented some arguments for this case, such as rather than coordinate transformation of the space, it is divided into convex polytopes within the original space. If true, this currently is outside my means to understand the implications fully; however, it is an interesting point which I will look into further, and have been searching for related literature. This is relevant to me, as, given the medical NLP domain that I intend to enter, there will be medical scenarios which have yet to occur before - new diseases being discovered or reclassification of diseases (which happens fairly frequently!). This was a very informative lecture for me and has provided me with lots of references to consider: • Logical syntax and semantics; their linguistic relevance https://doi.org/10.2307/410891 • Emergent linguistic structure in artificial neural networks trained by self-supervision https://doi.org/10.1073/pnas.1907367117 Take-home messages: Understand and learn more about geometrical views of deep learning.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Seminar 3 - NLP

Title	Evaluating Automated Citation Screening in Systematic Reviews: Metrics, Review Outcomes, and Datasets
Speaker	Wojciech Kusa, Final Year PhD candidate
Organisation	EU Horizon 2020 project DoSSIER
Date	20./10/2023
Narrative	The talker presented an hour-long talk on their work on systematic review automation in the medical domain. His presentation used a mixture of discussion of their experiences and basic examples to outline their points. This proved very effective, especially with the complex topics he presented. He was an extremely fast talker, which meant that a lot of material could be covered, but I, as an audience member, could not comprehend his research findings. His research area is something that particularly interests me, having previously been in the medical domain and conducted systematic reviews myself, yet even with this background, I found myself needing help to keep up with some of the concepts that he presented. His argument on traditional evaluation metrics for systematic review paper generation (true negatives, work saved oversampling, confusion metrics, etc.) was particularly valuable. I wish he had spent more time on this topic. This felt like it could have occupied the entire time slot. He then discussed the actual outcomes of systematic reviews and available datasets for this area, including advertising his soon-to-be-released dataset. Overall, this lecture could have been less dense to communicate their ideas effectively. While I understand that the speaker was keen to share what they had learnt throughout their PhD experience, this felt like a series of talks on the topic area. The abstract did effectively communicate the topics to be discussed. An interesting thing this presenter did was that, as the presentation was online, he aligned his webcam with his face, so it looked to the audience as if he was speaking to them. This felt effective. Take-home messages: ● Consider reducing content to ensure the audience can follow along adequately ● Frequent checks on whether the audience follows through with examples or real-world scenarios. ● Enunciate and speak slowly, slower than my normal speaking pace.

Seminar 4 - NLP

Title	Query Automation for Systematic Reviews.
Speaker	Harry Scells, Alexander von Humboldt Research Fellow.
Organisation	Leipzig University, Germany
Date	03/11/2023
Narrative	This presentation was a 1-hour remote presentation on the topic of query automation for systematic reviews. He was moderately successful in relaying the key concept: improvement within systematic reviews can come at different levels of the process. It had some overlap with a previous NLP seminar which I had attended, which addressed how to evaluate the metrics of systematic reviews, so it felt somewhat complementary; however, this seminary was targeting systematic reviews at a higher level than before - through improving the search query of the initial screening and aiming to reduce the amount of time it takes for more in-depth reviews on papers. The contrast in approaches between the two seminars (automating research reviews vs reducing the number of papers to reviews) was great, as it showed how, from a single problem area, there are multiple different approaches which can be taken to enhance it. During the lecture, the presenter used a simple image to explain the entire process of a systematic review. He emphasised the significance of automating this process due to its cost and time requirements. The lecturer also highlighted the importance of conducting a systematic review and explained in detail how it should be done. He outlined the different stages, which include screening through a boolean search, abstract screening, full study screening, study synthesis, and finally, the preparation and dissemination of systematic reviews. The process is aimed at refining studies from around 30 million to only about ten studies. However, it is an expensive process that takes up to two years to complete. Harry's research centres around developing search strategies that can help reduce the need to screen large numbers of abstracts and full-text documents. Systematic reviews typically require complex queries that involve multiple subclauses a vast array of medical terms and ontologies that need to be carefully chosen. Harry's research aims to make the systematic review process more efficient and accurate by improving search strategy development. Boolean queries are essential for systematic literature reviews, and they can be created automatically using either the conceptual method (which uses human expertise) or the objective method (which is more algorithmic). Both methods use seed studies for guidance and to balance the number of studies retrieved. The conceptual method involves gradually expanding high-level concepts, while the objective method follows a fixed order for each stage. Although the objective method is usually better, both methods can improve precision.

Seminar 5 - NLP

Title	Taking stock of understanding and intelligence in LLMs... Then and now
Speaker	Allyson Ettinger
Organisation	Sheffield NLP group
Date	26/1/2024
Narrative	This was a 1 hour NLP seminar provided by the Sheffield NLP group. The lecturer started by examining existing approaches to evaluating LLMs and related work. The idea behind this was to provide a story of evaluation, which provided good structure to the presentation - past, present and future. She also discussed the historical pattern of pre-trained LLms of hype and bust, likely related to their beating evaluation tasks. However, upon closer inspection, they could improve on generalisation. The lecturer outlined how current LLMs are performing exceptionally well on current evaluation metrics and how this is linked to their consumption of these metrics in their training data, which casts doubts on whether these evaluation metrics represent true learning (i.e. understanding of the context of a question, and applying it to different scenarios, or rote learning (i.e. memorisation of the relevant facts). Good, easy-to-understand examples were provided, with different tests showing property knowledge (Property knowledge and property inheritance). She presented her research in the area, such as adding attractors (Items related to the question but not the answer) to problem tasks for pre-trained LLMs to solve, resulting in poorer performance. This reduction in performance did not match that of human performance reduction, suggesting that the pre-trained LLMs were extracting relevant information from the context of the question. She also explained that adding this distraction, particularly after the actual entity, further reduces performance. A thought I had yet to consider, which was presented, was the lack of importance that negation provides with the subsequent word prediction, as there is a relatively high entropy of subsequent words that could be next. This is particularly important concerning natural language generation and could explain why hallucinations exist within these models. The next section of the lecture focuses on what's next and how these evaluation metrics have changed. She presented solutions to the problem of pre-trained LLMs using contextual understanding to answer the question with the introduction of logic puzzles, which have been used in psychology previously. Overall, the presentation style was excellent, if a little rushed, due to the amount of time allotted. I would be happy to receive further lectures from this person. The major points that I took away from this lecture: ● Pre-trained LLMs can “cheat” evaluation tasks if they are trained on them ● Evaluation tasks are in an arms race against the models, we will likely require more complex evaluation tasks as time progresses. ● Just because something scores highly does not mean the model has good generalisability.

Seminar 6 - NLP

Title	Demystifying prompts in Language models via Perplexity estimation
Speaker	Hila Gonen
Organisation	University of Washington
Date	15/12/2024
Narrative	This was an hour-long external lecture from a professor at the University of Washington, who has received two best paper awards in Corenll 2019. This lecture was provided via online meeting. The topic for the seminar was very relevant and interesting to other projects I have undertaken recently - namely, the effect of prompting on large language models. Hila's presentation style was clear, particularly when exploring potentially challenging topics. It certainly catered to an audience with background knowledge of the topic area, as there was an assumption that perplexity was understood. I would have improved this delivery by spending more time on perplexity and how it is derived. I had to search for information about perplexity while listening to the lecture, which made me need help understanding some of the later content. Her topic and paper were highly interesting and novel, and it is one of the first talks I have been to that has attempted to quantify how prompt engineering works and even provide potential explanations as to why certain prompts work better than others. Using perplexity as a proxy, with the hypothesis that the lower perplexity of prompts results in more accurate scoring was great. I doubt its application to other models, as it does not necessarily match my observations - in that providing more data via the prompt (and hence increasing complexity) results in more accurate results. This could be a potential avenue for research; however, it is a challenge to work on the perplexity of a sentence when you don't have the corpora that the newer, closed models have been trained upon. Hila's generation of different prompts is extremely relevant for the mini project, in which our subteam is looking at prompt engineering. Their approach is to generate 4-10 seed sentences, translate them into another language, then back to generate more, and finally ask a large language model to generate variations upon that, resulting in 100+ variations of the prompt. This did introduce noise, with some prompts not being grammatically correct; however, as her work showed, this doesn't necessarily impact the accuracy of the results. Ultimately, this lecture was interesting and well-presented, and I particularly enjoyed her interaction with the viewers. In contrast, it was being presented, rather than awaiting for answers at the end of the presentation - it improved my understanding of the concepts!

Seminar 7 - Speech

Title	Inclusive speech technology: Developing automatic speech recognition for everyone
Speaker	Dr Odette Scharenborg
Organisation	Associate Professor SpeechLab / Multimedia Computing Group
Date	26/2/2024
Narrative	The main focus of a one-hour lecture on speech technology was on inclusive speech technology and responsible research. Speech technology is primarily trained on everyday speech and normal voices, relatively homogeneous sources. This causes technology to be only available to a chosen few and increases digital divides, violating linguistic rights as included in the Human Declaration of Human Rights. The goal is to mitigate bias in speech recognition systems and uncover the causes of bias. There are multiple causes of bias, including differences in the training data distribution from the test data and algorithms and a need for more diverse development teams. The lecture discussed ways to reduce bias against diverse speaker groups in a SOTA ASR system, such as focusing on data scarcity, using data more effectively, and current research. We have around 700 languages of the word; however, we only have 2% of that available as speech technology. Speech technology is primarily trained on normal speech and normal voices, a relatively homogeneous source. There are many sources of diverse speech - such as regional accent, cold, and dysarthria, but speech technology is unavailable for these people. This is bad because technology is only for the chosen few; it increases digital divides and different access to information, especially for low/illiterate people, disabled people and people who speak 1 of the 98% of the world's language. Linguistic rights, as included in the Human Declaration of Human Rights, state that it is a human right to communicate in one's native language. Regardless of what it is, technology needs to include various languages. Ideally, we need lots of speech with many different transcriptions, so we need language models that predict the order of the words in the language. We need lots of text, which is unavailable in many languages. Transferring high-resource systems to low-resource systems often does not work. There is bias in data and systems against different speaker groups. The goal is to uncover/mitigate bias in SOTA speech recognition systems. Imagine you have a test set of N=24 utterances from 4 speakers; 6/24 utterances need to be recognised. Imagine you have the same 25 % error rate; however, you are just misclassifying one type of language. This is biased against speaker groups. Depending on the database, sometimes female speech is better recognised than male speech. Speech from people between 18-30 is better recognised than speech from people outside. Adults > Children, Standard Accent> Regional accent, White speakers > Black Speakers. Different languages have different studies; however, there is a systematic difference between groups.

	There are multiple causes of this bias - firstly, if the training data distribution is different from the test data, you will have issues; algorithms tend to introduce bias, and you also need a diverse developers team. How well can a SOTA ASE system deal with diversity of speech? First, you have to quantify it. Quantification of the bias in SOTA ASR What is the cause/origins of the bias? Is bias dependent on the language? Is bias dependent on the DNN architecture? What is bias? We do not know how to define bias; it depends on different factors; we define it as the difference in error rates between two speaker groups. She looked at bias against gender, age, regional accents, and non-native accents. Trained SOTA hybrid TDNNF-HMM and E2E asr (corpus gesproken nederlands) then tested on dutch jason corpus. Found that female speech > male, teenagers > older adults > children Native speakers > non-native speakers Dutch regional speech > Flemish regional speech Worse performance for all the adults - because regional accents may be more robust for adults. Read speech > Human interaction speech. Compared to the different architectures -> Differences in WER were smaller with hybrid systems. How can we reduce bias against diverse speaker groups in a SOTA ASR system? Focus on data scarcity Using data more effectively. Domain adversarial training Data augmentation does not help with all improvements in speech ASR systems More data with speed/volume perturbations More speakers with pitch perturbations More speakers with new articulation patterns will increase the variability of your training data. To reduce WER and bias -> Add natural or artificial data -> Use data augmentation on standard speech. If you use data augmentation and use data more effectively, you can reduce bias. SOTA data augmentation and training techniques need to be generalised better for all types of diverse speech. A low error rate does not mean you have a low bias.

Seminar 8 - Speech

Title	Where Speech Technology Fits into AI Safety
Speaker	Jennifer Williams
Organisation	University of Southampton, South England Engagement Director, UK national EdgeAI Hub, Chair of ISCA security and privacy in speech communication
Date	04/06/2024
Narrative	This was a one-hour in-person lecture at the University of Edinburgh during the joint CDT UKRI conference. This was one of the few lectures that I have been able to attend in person, which was given in a lecture theatre in front of a sizeable audience. The presenter intended the talk to be non-technical, where they would not go into experiments and results, which I enjoyed less. I understood that the point was to be as general as possible, as the audience comes from an extensive background of knowledge/skill set; however, the audience was primarily students of CDT in NLP/Speech, so this was a missed opportunity. The lecture started with the history of speech technology, which framed the subsequent lecture nicely, and the fast improvement in the capability of these models. It did, however, verge somewhat from the topic and used LLMs as an example of this increase - which, while they are related, and there have been some improvements with speech technology due to the transformer approach, I would have to look at this further to see the impact that this has had on the speech domain. 1. 2016 Merlin speech synthesis 2. Language grounding 3. Neural vocoders 4. Neural speaker representations (X-vectors) and style tokens 5. VoicePrivacy Initiative a. Technical response to oncoming GDPR privacy concerns 6. Emotion recognition and paralinguistic analysis 7. Spoofing speaker verification 8. Partial spoofing detection 9. Multi-task neural networks (enhancement + recognition, speaker recognition, anti-spoofing) 10. Foundation models, ChatGPT, deep fakes and AI regulation The presenters' point that speech technology faces more public scrutiny than ever before was well made. They mention that the increased public engagement with these technologies (even for deepfakes/novelty sake) means that there is a growing awareness of their capabilities. She referenced multiple challenges speech technology faces, such as opening up the public to novel threats and vulnerabilities. A point that was particularly well made and resonant to my research area was that we tend to hold these LLMs/AI technologies to a higher standard than that we have of existing technologies, such as internet search. Both can return answers to questions we might not want to be freely available (such as how to build a bomb). However, internet search is almost expected to be able to return these illicit results, while AI models could be better if they do. This dual standard of accountability for these technologies is something that I want to explore further in future lectures within RRI.

Seminar 9 - Speech

Title	AI analysis of Voice to Aid Laryngeal Cancer Diagnosis
Speaker	Mary Paterson
Organisation	University of Leeds
Date	12/06/2024
Narrative	This was a one hour lecture on AI analysis of laryngeal cancer. General symptoms are a change in voice, dysphagia etc - they are very broad symptoms. 160,265 cases in men since 2021, 24,350 in female. The vast majority of these cancers are stage 1 / stage 4. Median survival time after stage 4 for 5 years is 30% verses 90% at stage one. Stage 4 usually necessitates removal of the voice box + long term radiation therapt. Lots of side effects of having this treatment. Ultimately we want this dignosed at stage 1 to reduce this harm. 2 week wait pathway. Start at PCP, 2 weeks wait then go to a specialist. The general idea is that using an AI system is inbetween the PCP + specialist, looking at classification / prioritisation of patients. SPecialists are seeing a huge number of patients but actually rate of having cancer is really low (5%). This can ease the load on specialists. Pinpoint test system. Applicaiton possibilities -> recording device in PCP, and recording at home. If using recording at PCP we have a more standardised system, which can be human vetted before. Barrier of entry to use a smartphone/having an environment that won't cause issues with classification/diagnosis. Started with feature extraction - MFCC - eGeMAPSv02- wav2vec2 feature states. Wav2Vec2 feature states were not interpretable.Wav2Vec2 performed better than MFCC?eGeMAPSv02. Potential contributing factors for the failure of the models on the external datasets. - Patient demographics - Recording length - Recording devices - Recording environments. Did the length of the recording affect the model, so shortened the healthy patients? ● MFCC - Shorter recordings did worse, longer recordings did better (generally). ● eGeMAPS2-v02 o not a big difference ● Wav2vec2 - needs around 5 seconds to get good performance. However it did not translate to the external dataset. Wanted to improve this: ● Investigate preprocessing to reduce the degradation on external data. ● Investigate more complex speech tasks. ● Investigate the effects of recording device. Audio was collected from patients using the cookie theft test. Used to reduce the chance of speaker revealing identifiable data. Recorded on two devices and a Nokia 110 phone. Collected more information for demographics, such as Age, Gender, BMI, smoking status, Caffeine and Alcohol consumption. Often get people who don't tick the boxes the for all symptoms. Because the data is patient reported, it is often is noisy. Inbuilt noise reduction software on phones can reduce the amount of data collected - as it often treats elongated vowels as background noise after a certain point. Differences in the devices exist and is

	easily found. The lecture was presented in a great and informative way, with high interactivity. uestions: Have the patients consented to patient being shared.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

UKRI CDT Conference Poster

TECHNOLOGY ASSISTED REVIEW
within medical domain

University of Sheffield

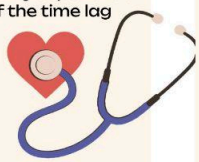
Author
Dr. Aaron Fletcher
ahafletcher1@sheffield.ac.uk
Supervisor
Dr. Mark Stevenson
mark.stevenson@sheffield.ac.uk
University of Sheffield

SLT Centre for Doctoral Training

UK Research and Innovation

ABOUT

As a **veterinary surgeon** with a decade of medical practice, my background may seem unusual for pursuing a PhD in NLP. During my practice, I aimed to improve the medical evidence base through **publishing a PICO review** on non-steroidal anti-inflammatory drugs and their potential to mitigate gastrointestinal side effects. My experience in medical research and awareness of the time lag between **conducting reviews and disseminating findings** motivated me to work on facilitating clinicians' access to the latest evidence through NLP.

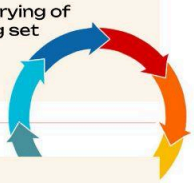


KEY TERMS

Systematic Reviews aim to identify, evaluate and collate findings relevant studies to a specific health issue, to make the available evidence easier to use.

TAR is the iterative process of retrieving documents from a collection, and then reviewing them until a substantial majority, or threshold has been met.

Active Learning is Interactive querying of the collection, adding to the training set with new labeled data.



AREA

The medical evidence base is expanding at an unprecedented rate, with a staggering **268.93% increase in publications from 2002 to 2022**. This growth poses a significant challenge, as it becomes impractical for a single human to thoroughly review all evidence returned from queries in a reasonable timeframe. Nevertheless, this is precisely what medical researchers are expected to do during the peer review process. As this body of knowledge continues to swell, the need for efficient tools that can review documents becomes increasingly urgent

To date, medical TAR research has focused on using techniques such as logistic regression with active learning. Could we **leverage more SOTA models,** such as Mamba or GPT, which have greater capacity for transfer learning, the ability to understand complex queries, improved text understanding, to achieve better medical TAR results? There is a plethora of AL techniques which can be used during the TAR process, however, **are some** more **superior** than others **within the medical dataset?** If so why?

OBJECTIVES

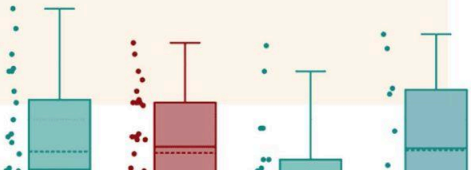
- Adapt SOTA models (e.g. Transformers/Mamba) for the Active Learning TAR approach within systematic reviews.
- Explore the effect of different active learning approaches on these models.
- Explore prompt engineering approaches on LLMs within TAR

DATASETS

CLEF eTAR is a collection of diagnostic test accuracy and intervensional systematic reviews.

Cochrane is a large collection systematic reviews.

PubMed is the largest collection of medical evidence.



This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]

C ENGAGEMENT WITH DPP ACTIVITIES: COM61004

COM61004: Introduction to Collaborative Research Practice for SLT ePortfolio

Student Name	Aaron Fletcher
Student Registration	230116573
Cohort Number	5
Academic Year	2023/24

Contents

Mini-Project

Each student will be expected to include

- *an individual log of contributions/activity organised on a chronological basis, i.e. a brief weekly summary of activity and their individual contributions. See the information about worklogs for additional guidance.*

Note: the group report will be submitted separately.
Add additional tables as required.

Your group	Team 1

Work log - week 1

Date (w/b)	23/10/2023 (2 ½ hours)

Narrative	- 23/10/2023 (1 hour)
	  - An initial meeting to discuss data generation was held. This was a freeform meeting where people brainstormed ideas for generating the data. Little consensus was made, however, we have scheduled a meeting for later in the week to address some questions raised - what menu items to include, different scenarios/scripts to produce audio recordings, and what participant characteristics we want to cover. A Trello board with tasks and meeting notes were created in a shared file.
	- 27/10/2023 Second meeting based on an agenda (1 & ½ hours).
	  - I have taken somewhat of a role of project chair/manager by facilitating the meetings to ensure that people's opinions are heard. Various items were discussed based on the agenda, and tasks were assigned to subteams to meet again to discuss findings. I have been tasked with creating the agenda for the third meeting and proposing a data management plan with Boxshaun and Yao - this is an area that I have little experience in (besides using GitHub for my projects), so I will investigate the university's advice on this and discuss with my colleagues.

Work log - week 2

Date (w/b)	30/10/2023 (4 ½ hours)

Narrative	- 31/10/2023 Sub team Meeting (1 hour)
	  - First meeting with sub-team, where an overall approach to the topic was outlined, with key components such as ASR, Named Entity recognition separated into two sub-groups. Each sub-group will work on researching SOTA approaches for their small task before passing it on to the other sub-team to verify.
	  - ***Task***: Research Named Entity Recognition and provide SOTA options for the project.
	- Data Management Plan (1 hour):
	  - In conjunction with the other DMP team members, a data management plan was outlined and will be sent to the library for review before submission.
	  - ***Task:*** Submit data management plan for review
	- 2/11/2023 Dataset recording Meeting (1 hour)
	  - Introduction by Stefen to the equipment that is available for us to use.

	- 3/11/2023 Data Generation Team Meeting (1 & ½ hours)
	  - The project has changed the scope to focus more on data generation than previously, as approximately 80 hours of recording will be needed for publication.
	  - The preliminary cut-off point for 14th February is set for finishing data generation.
	  - Discussion of continuing ethics application.
	  - Preliminary discussion of how the data will be generated/Experimental design, i.e. closed microphone on the server + customer, with an outside array of microphones to provide a differential.

Work log - week 3

Date (w/b)	6/11/2023 (3 hours)

Narrative	- 9/11/2023 - Sub-team meeting
	- The Data Management Plan was submitted for review by the library team / Stuart Wrigley , and feedback was received. A few changes were made based on this - such as not using Google Drive to mirror the backup, as the university research storage offers more storage and is backed up on multiple sites. This is a live document so further iterations will be made depending on requirements.
	- Following a meeting with the CDT management team, the project scope was scaled back - from 80 to 10 hours of training data.
	- The data management plan was finalised by all DMP team members on 9/11/2023. This will be submitted on Monday 13/11/2023.
	- Me and FP created a Gantt chart for the Project management meeting, which outlined key project phases and their ending times.
	- Subteam work:
	  - A further 1-hour meeting was held where we discussed our research. I had been focusing on a potential named entity recognition. I created a list of potentially interesting NER models to investigate based on the f1 score and the accuracy of hugging face models. Collectively, we were interested in using span markers (a NER model which, rather than categorising the word, does so over a range of words related to the category, which might be more applicable to our applications).
	  - Decided to attempt to implement one of these models with different encoders, as it will form the research question of what encoder (tokenizer) results in better performance with span NER tagging.
	  - To enable comparison of our results, we will attempt to use the same dataset (Berkley), which we will annotate with our additional class.

Work log - week 4

Date (w/b)	13/11/2023 (3 hours)
Narrative	Sub-team meeting: ● Discuss what we plan to present in the mini-project meeting catch-up, as the team has submitted into three parts (NER, ASR and Query Detection); each group will contribute a slide on their work so far. ● I have been attempting to generate some training data, and some testing data which represent the problem more and code up a way of splitting these up. Data generation meeting: ● The meeting was interrupted by a fire alarm! ● Reduced scope to address issues with time in producing enough data (40 hours down to 10 hours).

Work log - week 5

Date (w/b)	20/11/2023 (4 hours)
Narrative	Sub-team (*4 hours*): ● Formation of a slide deck for presentation for a mini project meeting. ○ Included proposed task topology, work undertaken by the ASR subteam, the NER subteam, and the query identification subteam. ○ Included metrics of current approaches so far. ● I uploaded a working example of a Packed Levitated Marker for Named entity recognition to the group Github, which got 74% accuracy. This was trained on the FoodBased NER corpus and tested on the Berkley Restaurant Corpus. ● Production of Gantt Chart for subteam. Data generation meeting - No meeting was held this week; however, some discussions were held in the Mini project meeting (*30 minutes*): ● Further reduction of data gathering scope. ● Discussion of experimental design. ● Discussion of ethics application. Training in the Kroto lab (*30 minutes*): Health and safety induction into the Kroto laboratory now enables us to get some quick data to assess.

Work log - week 6

Date (w/b)	27/11/2023 (3 hours)
Narrative	Data generation progress (1 ½ hours): ● Meeting 5 (*1 hour*) ○ We discussed publication, with the consensus being that those wanting to publish will continue following the cessation of the mini-project. ○ Established that we would want to record in the Kroto lab with scenarios. Dummy data is pending and will be recorded and shown at the next meeting. ● Speech transcription and annotation meeting (30 minutes): ○ Meeting with Minghui, established some brief outlines of what we want the annotation guidelines to look like (i.e. span tagging, item categorisation). ○ AF to investigate program platforms to use this. Subteam progress (*1 ½ hours*): ● Finalisation of subsections and upload of completed code to GitHub ● Relocation to subset task to intent classification. ○ Established the types of classes that we want to detect (Add, Remove, Edit) ○ Established a dataset that we will use to create mock models: ■ https://huggingface.co/datasets/xjlulu/ntu_adl_intent/viewer/default/train?q=food&p=1 ■ https://huggingface.co/datasets/silicone/viewer/maptask/train ○ Once completed, I will attempt an LSTM approach to the data and report findings to the subteam.

Work log - week 8

Date (w/b)	4/12/2023 (3 hours)
Narrative	● Data Generation Team: ○ Meeting with whole group (1 hour). ■ I presented annotation work done since the last meeting, which raised some issues with their creation. ○ Meeting with Stefen (30 minutes). ■ Discuss some easy changes to the experimental design that will improve the data quality, and arrange training sessions in the Kroto Lab. ○ Commence creation of annotation guidelines (1 Hour) ■ Discussed aspects of annotation guidelines and ultimately decided that span marking was better for the data type we

	would be using. We are planning on annotating on three levels ● Food type ● Query Type ● If to include within the dataset. ■ Different platforms were discussed to do this (Prodigy vs Borcorro). Ultimately, we decided to use Prodigy, AF + MZ will create annotation guidelines for the rest of the group and present it. ● Sub teamwork (30 minutes) ○ Implementation of the intent classification LSTM model.

Work log - week 9

Date (w/b)	11/12/2023 (5 hours)
Narrative	● Data Generation progress (*4 hours*) ○ Pilot recording in the Kroto Labs, including improving experimental design (2 hours) ■ Meeting with Stefan Goetze and other subteam members to set up an experimental design within the Kroto laboratory that closely matches real-world conditions. ○ Prepare a slide deck for the data generation and sub-team meeting with project directors, and attend the meeting (2 Hours). ■ Discuss the current project status. ■ Discuss potential changes to the annotation guidelines. ■ Share pilot recordings and compare the two experimental designs. ● Subteam Progress (*1 hour*) ○ Creation of LSTM approach to intent classification for open-sourced dataset (1 Hour). ■ 150 class classifiers which performed poorly (accuracy ~10%) ■ 12 class classifiers which performed well (accuracy 100%) ■ Updated to the shared GitHub repository.

Work log - week 10

Date (w/b)	18/12/2023 (6 hours)
Narrative	● Data Generation progress (*3 hours*): ○ Further pilot recording in the Kroto Labs, including improving

	experimental design (3 hours) ■ Generate pilot data, which can be used to further subteam approaches. ● Subteam Progress (*3 hours*): ○ A sub-team meeting (20/12/2023) will be held to discuss individual tasks over the holiday period and the formation of the research questions. I will focus on prompt engineering (1 hour) ○ Created a Python program to programmatically manipulate the recorded files and run various models on them, such as whisper, including prompt engineering (2 hours).

Work log - week 11

Date (w/b)	25/12/2023 (0 hours)
Narrative	● Holiday week

Work log - week 12

Date (w/b)	1/1/2023 (5 hours)
Narrative	● Subteam Progress (*4 hours*): ○ Research question formation (3 hours): ■ Formation of the prompt engineering research questions ■ Generated 5-6 research questions on prompt engineering, including background research, assessing implementation and write up. ■ RQ1 - Does providing a neogeolistic dictionary of words used in the transcription reduce WER for transcription. ■ RQ1 - Does providing longer term dependencies reduce WER for transcriptions ■ RQ3 - Does providing context improve accuracy? ■ RQ4 - Does providing more context improve accuracy? ■ RQ5 - Does chain-of-through prompting improve accuracy? ○ Subteam meeting (1 hour) ■ Meeting to discuss research questions, and explain to the group and discuss any potential challenges. ● Data Generation progress (*3 hours*): ○ Creation of annotation guidelines (1 hour) ■ Started drafting more formalised annotation guidelines within a shared google document.

Work log - week 13

Date (w/b)	8/1/2023 (3 hours)

Narrative	• Subteam Progress (*3 hours*)
	  ○ Creation of Research questions presentation
	  ■ The slide deck was created in the shared Google Drive in anticipation of the meeting on 16/1 outlining the project proposal (*1 hour*).
	  ■ Meeting with sub-team to discuss project proposal in person on Friday (*1 hour*).
	  ■ Further refinement of project proposal document (*1 hour*)

Work log - week 14

Date (w/b)	15/1/2023 (3 hours)

Narrative	• Subteam Progress *(2 & ½ hours)*
	  ○ Directors Meeting 3 (*1 & ½ hours*)
	  ■ Preparatory meeting before presentation to revise approach (30 minutes).
	  ■ Meeting with directors to outline project scope and email scoping document to all attendees (*1 hour*).
	  ■ Investigation and discussion of finalised order evaluation metric (such as graph edit distance, etc) (*30 minutes*).
	  • Generate some proof of concepts within Python to discuss with the group.
	• Data generation progress (1 hour).
	  ○ Advise the rest of the group to review/edit annotation guidelines
	  ○ Decide on purchase items for research.
	  ○ Update re: ethics application.

Work log - week 15

Date (w/b)	22/1/2023 (3 hours)

Narrative	• Data generation progress (4 hours).
	  ○ Finalisation of annotation guidelines (1 hour)
	  ○ Recording within the Kroto laboratory for more pilot recordings (2 ½ hours)
	  ■ This pilot recording identified some shortcomings with the research design, such as:
	  • The Research Coordinator needs to be more active in organising the research and ensuring quality recordings.

	● An outline needs to be used for the server flow to ensure that key points are being hit. ● The menu ordering system needs to be updated to current guidelines. ○ Generation of a research working log to keep participant characteristics (½ hour) and assign unique UID to each scenario/participant. Updated this to ensure that it is easy to use.

Work log - week 17

Date (w/b)	29/1/2023 (5 hours)
Narrative	● Recording audio scenarios (5 hours)

Work log - week 18

Date (w/b)	5/2/2023 (6 hours)
Narrative	● Recording audio scenarios (5 hours) ● Preparation of Director meeting notes (1 Hour)

Work log - week 19

Date (w/b)	12/2/2023 (7 hours)
Narrative	● Annotation masterclass meeting (1 hour) ○ A meeting presented by Minghui outlined the generated transcription and annotation guidelines and how to use the program Prodigy to transcribe the audio. ● Initial transcription test run (1 hour) ○ Tests were run where each person annotated transcripts to compare experiences and common challenges. ● Post-run meeting (1 hour) ○ The decision taken at this meeting is not to annotate food/quantity, and the intention to speed up this process. ○ Clarification of common issues, such as overlapping speech, which we decided not to translate. ○ We also decided not to transcribe dysfluencies. ● Subteam Meeting (1 Hour) ○ At the subteam meeting, we discussed scope reduction (removing the granular approach). ○ The code was generated for the prompt engineering approach. ● Creation of the prompt engineering approach code (2 Hours) ○ Creation of the code to run prompts through gemini + gpt, with test runs being created via the whisper transcription directly. ○ To determine evaluation metrics.

	• Transcription time (1 hour) ○ Transcription of the audio following above guidelines for this week

Work log - week 20

Date (w/b)	18/2/2023 (*4 hours*)

Narrative	• Creation of the prompt engineering approach code (*2 Hours*)
	  ○ Creation of the code to run prompts through gemini + openai, with test runs being created via the whisper transcription directly.
	  ○ To determine evaluation metrics.
	• Studio recording (*2 hours*)
	  ○ Recording more scenarios.

Work log - week 21

Date (w/b)	25/2/2023 (*3 hours*)

Narrative	• Transcription (*1 Hour*)
	  ○ Production of 10transcripts using prodigy
	• Prompt Engineering Coding (*2 Hours*)
	  ○ Further coding to complete this area

Work log - week 22

Date (w/b)	18/2/2023 (*3 hours*)

Narrative	• Project Meeting (*1 Hour*)
	  ○ Project Meeting with Stuart and Rob Gaizauskas to update them with the status of the project. Stark reminder that there are a few number of weeks left to produce the output (before easter!)
	• Meeting with Prompt Engineering Subteam (1 hour)
	  ○ Meeting with PE subteam to ensure that we are all on the same page with expectations set and deliverables mentioned.
	• Transcription (*1 Hour*)
	  ○ Production of 7 transcripts using prodigy

Work log - week 23-24

Date (w/b)	18/2/2023 (*12 hours*)
Narrative	● Run Experiments (*4 Hour*) ○ Run experiments for the PE team using code created throughout this process. The experiments are listed in the final project report. ● Meeting with Prompt Engineering Subteam (30 Minutes) ○ Meeting with PR subteam to clarify what the prompts are going to be (Me and FP) ● Project write-up (*7 hours 30 minutes*) ○ Collaborative write up of the project using Overleaf/Latex with the team.

Journal Club

Each student will be expected to record in their e-portfolio:

- *their commentary bullets about each paper*
- *brief write-up of outcomes of the discussion from each session They should indicate if they were the discussion chair*
- *A copy of the slides used at each session for which they were the presenter.*

Add additional tables as required.

Journal Club 1

Paper title	Speech Processing for Digital Home Assistants: Combining Signal Processing With Deep-Learning Techniques (2019)
Presenter	Jason Chan
Date	16/10/2023
Commentary bullets	**OVERVIEW:** Excellent starting point for students exploring speech concepts. **TARGET AUDIENCE:** It is vague; the paper presented a (kind of) systematic review and challenges within speech processing, yet it is in a magazine-style format. Readers are likely already engaged in speech processing, so is an “entry point” for this audience somewhat redundant? **SOTA:** Produced in 2019, so likely not. **ASSUMPTIONS:** ● Communication of information is only TTS; this is increasingly not the case as digital home devices increasingly have visual components, e.g. Google Hub. **WOULD HAVE LIKED:** ● An explicit discussion of the effect of multiple ASR systems in one environment. ● Discussion on how communication with these devices could be more natural and prescriptive. This could stem from the vagueness of what A “digital assistant” is. I.e. using “wake words” does not seem natural and changes my interaction with the device. **KEY POINTS:** ● Deep neural networks are not a universal solution for speech-processing tasks. ● Speech processing for digital assistants has yet to be solved. ● It is a great reference for speech processing challenges and (some) solutions.
Narrative	In this journal club, I was chairing the discussion. The presenter, Jason, covered the main aspects of the paper, outlining all the approaches given to speech recognition within the paper (Acoustic environment, Multichannel speech enhancement, automatic speech recognition, Text-to-speech synthesis and fully handfree interaction. This journal club reviewed a mini-review paper on approaches to key speech processing challenges circa 2019. Multiple points were raised during the discussion, such as who this paper was aimed at (it was concluded signal processing specialists who were interested in deep learning approaches). Some audience members opined this made the content less relatable to themselves, as the paper heavily relied on mathematical notation specific to the domain. Additionally, the magazine format made figures and references appear distant from the relevant text, hampering their understanding. While I did not necessarily agree (considering it was meant for production in a magazine, where it would have felt more appropriate), greater consideration by the authors of the main method of article consumption could have been given. **Action points to improve chairing:**

	• During the discussion, some people were more vocal about contributing than others, which, as the chair, is my role to adjudicate. To improve this, I asked targeted questions to people who had yet to have the opportunity to speak so far, ensuring that every person had contributed to the discussion. • Initially, I would have people read their points from the slide deck; however, this created a very prescriptive approach to the discussion (people were waiting their turn, etc). The discussion was better suited to more of a conversation, with my prompt questions allowing people to add to the discussion further. Going forward, a more informal approach is recommended. • Overall, it was a positive experience, which I require more experience with.

Journal Club 2

Paper title	An Overview of the SPHINX-II Speech Recognition System.
Presenter	Anthony Hughes
Date	31/10/2023
Commentary bullets	**Positives:** • Liked the introduction of key terms (Senones, etc). • Seems like a big ASR development at the time! • Uses a hybrid approach, which aligns with my cognitive systems bias on general AI! **Critiques**: • Multiple instances of abbreviations are used without instantiating first (e.g. MFCC, HMMs, LPC). • When introducing a novel term (Senone), a worked example is appreciated! • A clear overview of the SPHINX system is challenging due to self-citation and domain-specific terminology. • The paper only addresses out-of-vocabulary utterances beyond increasing the amount of data. • Language models used are of their time (i.e. better options exist now than n-grams!). • No consideration of noisy environments given - Would expect senone's will change with noise! **Questions:** • What exactly is a semi-continuous hidden Markov model? • Have hidden Markov models just been replaced by DNNs? • Are some senone's language unique? • Where are the codebooks? Could the author have provided more information or made them open source? Is this due to sponsorship restrictions?
Narrative	• Paul Gering chaired this journal club meeting. • This journal club had a slightly different approach than the previous one, with the facilitator (Anton) interjecting more than the previous session (to clarify issues with our understanding of the texts). • A key question was the motivation behind reading this paper, which seemed to be divided - motivations raised from it being a requirement for the CDT, reading because it is a foundational system of the topic area, and some people hand little motivation. This is important because, depending on the motivation, you will have different expectations of the article. • Was a conference paper, which might explain the formatting as typically, abstracts are submitted before the article is written. • Commercial application of SPHINX-2 was a quite successful dictation system.

Journal Club 3

Paper title	From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation.
Presenter	Yanyi Pu
Date	17/11/2023
Commentary bullets	• Very well laid out, structured referenced paper for word meaning representations. Provides a great starting point! • Not fully comprehensive, but as stated (due to the pace at which the field evolves), it is not intended to be. Would have liked search methodology, however! • Great explanation of how polysemy affects word embeddings and research done to overcome this. • Translational-based embeddings - This is a new concept; how would it deal with words without direct translation (i.e. Waldeinsamkeit)? • Four axes, not three, on pg 481. • Do humans discriminate between words linearly? • “Curse of multilinguality” - any information on why this is happening? Do people experience this curse? • Reporting Bias - Paper gives a great way of formalising this concept. • Good section on current challenges, such as degeneration issues in contextual embedding spaces.
Narrative	• Jack Cox chaired the journal club. • It was a very interesting and informative journal club despite the journal covering a large body of work. • The paper reviewed the majority of work that had been done on foundational topics within word embedding distributional models, such as how to embed meanings into Euclidean space, sense-aware embeddings and semantic knowledge injection to word embeddings. The journal club discussed: • What exactly is meaning and understanding (and a discussion of formal semantic language understanding vs distributional language understanding). • A discussion of symbolic grounding and how language can be very specific to the person using their language and grounded in their understanding. • Limitations of distributional language understanding approaches include whether we linearly distinguish between words / semantic meanings, their biases on training data and their inability to generate novel concepts. • What an optimal approach would look like - semantic knowledge injection through combining the two approaches. • Translational-embedding models: this was a new concept to most, and generally, people thought it would be an excellent way to encode semantics into models. We discussed how far apart the languages would need to be to encode semantic knowledge sufficiently (it seems like it would have to be not close enough but not too far apart. • People’s uses for the paper, with the consensus being a reference paper for word embeddings used to look up studies performed in the area, but it not being an exhaustive representation of this topic in the fast-moving domain.

Journal Club 4

Paper title	Inter-Coder Agreement for Computational Linguistics
Presenter	Aaron Fletcher
Date	8/12/2023

Commentary bullets	• Full link here.
	• Does Inter-Coder Reliability Matter?
	  ○ Low inter-coder reliability means unreliable datasets.
	  ○ If two coders produce similar results, then they have a similar internal understanding of the annotation scheme.
	  ○ Reliability is required for validity.
	  ○ If coders are inconsistent, then
	  ○ Annotation doesn't capture the truth.
	  ○ Some coders are incorrect.
	• Paper Contribution: Common notion
	  ○ Solution: Provide more!
	  ■ items $\{ i \mid i \in I \}$ of cardinality i
	  ■ categories $\{ k \mid k \in K \}$ of cardinality k
	  ■ coders $\{ c \mid c \in C \}$ of cardinality c
	  ■ Ao: Observed Agreement
	  ■ Do: Observed Disagreement
	  ■ Ae: Expected Agreement
	  ■ De: Expected Disagreement
	  ■ $p(\cdot)$: Probability of a variable
	  ■ $\hat{p}(\cdot)$: Estimate of the probability from observed data
	  ■ n with subscript to indicate the number of judgements of a type.
	• How to measure Inter-Coder Reliability
	  ○ Agreement without chance correction.
	  ○ Chance correlated coefficients.
	  ○ Coefficient S
	  ○ κ
	  ○ π
	  ○ κw
	  ○ Krippendorff's α
	  ○ Fleiss's Multi π
	  ○ Multi-κ
	  ○ BUT
	  ○ Which measure do we choose and why?
	• K, what is it good for ?
	  ○ Absolutely nothing :)
	  ○ K is any kappa-like coefficient (κ, α, π etc).
	  ○ No universally accepted standard.
	  ○ Medicine: >0.4
	  ○ CL: > 0.8
	  ○ The authors: No specific threshold

- Krippendorf themself thinks > 0.9!
- In ML/AI we are often generalising patterns, not hypothesis testing, and K is a poor predictor of ML success.
- K scores are ultimately tests for soundness of annotation scheme.

- Bias and Prevalence
  - Bias problem: Π/κ different when annotator marginal distribution are wildly divergent.
  - Prevalence problem: When a most of the items fall under one category, might lead to artificially high observed agreement by chance.
  - Authors suggest coefficient choice should be based on desired interpretation of change agreement not magnitude of divergence.
  - In reality we typically have a small subset of data where c > 1, which reflects the reliability of the annotation procedure, so Π / α more appropriate.
  - κ provides more information information regarding validity - it rewards biased annotators.
  - Can express overall annotation bias as the difference between κ and Π.
- Unitizing
  - The process of identifying units of annotation (typically linguistic units e.g. words, utterances, or noun phrases).
  - Segmentation/Topic Marking: Portions of text that constitute a unit because they are about the same topic (Think of our mini-project, where we are marking “2 burgers” as one topic!).
  - Agreement coefficients with segmentation likely to be lower:
  - Boundary/not boundary distinction: K = 0.647
  - TREC segmentation of broadcast news: K = 0.784 / 0.36
  - Identification of argumentative zones: K = 0.81
  - Conversational games: K = 0.59
  - People tend to agree on the bulk of segments, but not the boundaries!
- Anaphora
  - Where a grammatical substitute relates to a previous word or topic.
  - Lots of different types (some authors suggest 12 types!).
  - E.g.
  - Each fall, penguins migrate to Fiji. It happens just before the eggs hatch.
  - it is referring to the migration of the previous sentence
  - M: ....
  - first thing I’d like you to do
  - is send engine E2 off with a boxcar to Corning to
  - pick up oranges
  - as soon as possible
  - S: okay
  - M: and while it’s there it should pick up the tanker
  - Passonneau’s Proposal (using sets of mentions of discourse entities as labels).
- Word Sense Tagging
  - Bank can refer to a financial institution or the side of a river.
  - “I deposited money in the bank”, Bank should be tagged with sense of financial institution.
  - Requires dictionary: often coders will have different understandings of each word sense, and gets even more complicated with polysemous verbs.
  - Paper suggested use professional lexicographers and arbitration.

	○ Use coarser grained classification schemes, which group together dictionary senses (Wordnet) ○ Call has 28 fine-grained senses in Wordnet 1.7 ● The Main Takeaways ○ Report your inter-coder agreement, the maths isn't that scary! ○ You can classify intercoder correlation in terms of agreement or disagreement. ○ Justify statistical test(s) beyond it is standard for inter-coder agreement! This paper is an excellent reference. ○ But if we were to generalise … ■ c = 1 → No need ■ c = 2 → κ ■ c > 2 → Krippendorff's α ○ Don't get hung up on "perfect scores". ○ More coders, better results. ○ How often do we have c > 1 in the age of "big data"? ○ Why did they not provide a worked example for when c > 2? ○ What is the "sweet spot" for variability? Often, we use noise to increase the robustness of our models.
Narrative	● All participants well-received the paper, highlighting its usefulness as a reference when undertaking dataset annotation. ● The discussion focused on the variability of an acceptable K value and how it should be used. The consensus was that the actual value doesn't particularly matter. ● The group thought the conclusion could have been more explicit in their recommendations. ● The importance of expert coders was discussed. ● Is high agreement a requirement / relevant for the real world? ● The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing

Journal Club 5

Paper title	Efficient Estimation of Word Representations in Vector Space
Presenter	Paul Gering
Date	5/2/2024
Commentary bullets	• Main focus on the paper is clearly explaining how these models are computationally efficient, and why previous models, such as n-gram modelling are not. • Further optimisation given through the use of Huffman trees / hierarchical softmax given with a clear explanation. • Doesn't delve into why CBOW scores higher for syntactic accuracy (64%) vs Skip-Gram on Semantic accuracy (55%). • Questions: • With CBOW why are we sampling future words? When writing I don't tend to consider future words within a sentence, rather I attempt to convey a thought. • Both skip-gram and CBOW use context windows, what are the optimal lengths for these windows? • Would it account for homonyms/polysemes - don't think so! • Why did they choose 640 dimensions? • Is King - Man + Woman = Queen a valid evaluation metric?
Narrative	This was an hour-long journal club which was chaired by Minghui Zhao. This paper covers estimation of word representations in vector space. • Seminal Paper was generally well received. • Very good replicability within the paper. • Does this represent learning and understanding, human need a few examples, and we generalise it very well whereas the neural networks do not do this. • Discussed assumptions of the models - distributional hypothesis, linear relationships, context window. Doesn't consider long-term dependencies. • Listed the vectors which were released. • Bias within the models, discussed if they represent understanding or understanding. • If you evaluate the distant/top k predicted things then you are introducing more assumptions and more hyperparameters. Which improves the clarity of the causes of this. • Circular analysis potential present, which could present train data contamination. • Would have liked more analysis of the results in the paper. • This was about the time that GPUs were starting to be created that could make use of neural networks. • Focus in NLP was very different in 2013, SVM, etc. • Table 4 - Uncertainty over whether the comparisons are valid. Issue if it was a valid comparison. This is a proof of concept rather than a benchmarking paper. How much of the information is useful in the table. • Document/ word level meaning representations were popular at the time, so vector representations were used as it was the goal, rather than natural language generation.

Journal Club 6

Paper title	Efficient Methods for Natural Language Processing: A Survey
Presenter	Boxuan Shaun
Date	20/2/2024
Commentary bullets	• Reads like a cookbook on how to approach optimisation for NLP :). • Due to the attempt to cover everything, more is needed; however, the function of this paper is to provide readers with references on the various approaches to improving efficiency. • Would have enjoyed more discussion on pruning, particularly given how closely it closely matches what happens with humans - see synaptic pruning. • This very effective paper highlights that universities need more hardware capability to analyse these models. • No methodology was presented for their search strategy for papers (was this a systematic approach or cherry-picking of papers?) • Some newer approaches (prompting, knowledge distillation) are new and likely to be updated rapidly due to the field. Could it have been better to go more in-depth on these areas or potentially a series of papers? • Great general explanations of each optimisation concept! • I liked some practical examples of implementing some of these optimisation concepts, which might be out of the scope of this paper but would have made it a great resource! • Very little consideration for the interpretability of the models - does this have to come at the expense of efficiency?
Narrative	This was an hour-long journal club which Ian Kennedy chaired. The paper covers increasing the efficiency of the LLMs in NLP. • Participants discussed the breadth and context of this paper. It was noted that the paper came from a discussion in Germany. This potentially biases the reported papers to the participants of that discussion. • Due to the large breadth, there was limited discussion of each of the topics brought up in the paper. Some sections need to be more developed (such as the prompt engineering section), potentially owing to it being an emergent field at publication. • It provided a starting point for evaluating efficiency, yet a clear approach was needed. • Very little consideration of the increased interpretability of the models; does this have to come at the expense of efficiency? • Some participants didn't like the fact that opinion was not offered in the paper, while others offered too much opinion through the paper. • A significant issue with the paper was the need for a clear methodology for the search criteria. Was this a systematic review, or was this an opinion piece? What were the inclusion/exclusion criteria? • Discussion of the role of surveys and key differences between systematic reviews. surveys flirt too close to being systematic reviews to not explicitly state they have not effectively searched the search space, which can lead to misunderstanding.

Journal Club 7

Paper title	Position information in transformers: An overview
Presenter	Minghui Zhao
Date	7/3/2024
Commentary bullets	• Great survey paper on the different approaches to position information in transformer models, it excelled at highlighting variations of a single aspect (position information). • Unified mathematical notation is great! • Table 1 and 2 are great contributions to area, as it is a valuable reference for future work. • Clear logical layout to the paper, particularly the recurring concepts section (Reference point, injection method, fundamentals). • ¼ of a page of text explaining how a table is showing related papers (740)! Potentially could be more concise here. • Quite liked the point they made NOT to provide quantitative comparison for this survey, it would have been near impossible to recreate all these models on the same dataset, and evaluation metric. • Re: Character level processing - characters are semantically void, so would this have more of a downstream effect on polysemy tasks (assuming a non-infinite window length for positional information)
Narrative	This was an hour and a half long journal club which Tom Clark chaired. The paper covers the various approaches to position information within transformers. • Participants discussed the breadth and context of this paper. It was very focused on a small feature of transformers, which allowed it to be a great reference for information in that domain. • Discussion was had on why we need to encode positional information - for example, why order matters with text and noting that the transformer model is order invariant. • Participants discussed if positional information was even a pre-requestite for accurate understanding of sentences. Some argued that if we look at the start and the end of the sentences we can potential predict the entire sentence from that point (langauge modelling), and we do this as humans. I personally didn't agree with this, as we do not have proof on how humans encode information, and there is a potential infinite number of permutations that a sentence could take. I think positional information is important for comprehension by transformers. • Questions were varied from the group, with discussion on why they were investigating sinusoidal embeddings, despite them not having positional information. • Discussion on the different level of embeddings, such as character vs subword/word. I was interested in why character level embeddings might provide better results as they are semantically void, to which Anton suggested that it might be because of the fewer permutations that tokenizing at this level could bring, reduced out of vocabulary issues, and that they potentially could work better on smaller datasets, due to it producing more examples.

Journal Club 7

Paper title	Position information in transformers: An overview
Presenter	Minghui Zhao
Date	7/3/2024
Commentary bullets	• Great survey paper on the different approaches to position information in transformer models, it excelled at highlighting variations of a single aspect (position information). • Unified mathematical notation is great! • Table 1 and 2 are great contributions to area, as it is a valuable reference for future work. • Clear logical layout to the paper, particularly the recurring concepts section (Reference point, injection method, fundamentals). • ¼ of a page of text explaining how a table is showing related papers (740)! Potentially could be more concise here. • Quite liked the point they made NOT to provide quantitative comparison for this survey, it would have been near impossible to recreate all these models on the same dataset, and evaluation metric. • Re: Character level processing - characters are semantically void, so would this have more of a downstream effect on polysemy tasks (assuming a non-infinite window length for positional information)
Narrative	This was an hour and a half long journal club which Tom Clark chaired. The paper covers the various approaches to position information within transformers. • Participants discussed the breadth and context of this paper. It was very focused on a small feature of transformers, which allowed it to be a great reference for information in that domain. • Discussion was had on why we need to encode positional information - for example, why order matters with text and noting that the transformer model is order invariant. • Participants discussed if positional information was even a pre-requestite for accurate understanding of sentences. Some argued that if we look at the start and the end of the sentences we can potential predict the entire sentence from that point (langauge modelling), and we do this as humans. I personally didn't agree with this, as we do not have proof on how humans encode information, and there is a potential infinite number of permutations that a sentence could take. I think positional information is important for comprehension by transformers. • Questions were varied from the group, with discussion on why they were investigating sinusoidal embeddings, despite them not having positional information. • Discussion on the different level of embeddings, such as character vs subword/word. I was interested in why character level embeddings might provide better results as they are semantically void, to which Anton suggested that it might be because of the fewer permutations that tokenizing at this level could bring, reduced out of vocabulary issues, and that they potentially could work better on smaller datasets, due to it producing more examples.

Journal Club 8

Paper title	An Overview of Speaker Identification: Accuracy and Robustness Issues.
Presenter	Jack Cox
Date	21/3/2024
Commentary bullets	• It is well structured and provides great justification for its need, with clear explanations of verification vs identification. • Why would you need to "balance" FAR = FRR, isn't this particularly application dependant - assuming you want higher FRR in say banking applications? • Features section is a good recap of information learnt in COM 6502 speech processing! • This tutorial covers a lot of ground, missing data, gaussian mixture models, universal background model, SVM etc - Might this had been better focusing on one aspect of these rather than every aspect. Do we really need basic principles of SVM separating hyperplanes for example (Figure 11 page 33)? • GMM-SVM/GMM-UBM perform better with mixture increase, so why does author state that "the importance of using a GMM-UBM approach when confronted with limited amounts of training data " - What am I missing here?
Narrative	This was an hour-and-a-half-long journal club that Anthony Hughes chaired. The main discussion stemmed from the purpose of this paper, in that it attempted to cover an extensive range of topics, yet crucially, it needed to provide more detail in each section. A good example is the necessity for discussion of support vector machines in this paper, which went into great detail despite not being the primary focus of this paper. People who were more interested in the speech technology side were interested in this paper, as it provides a good reference for fundamental topics (such as Fourier transformations etc. As someone uninterested in speech, this offered little exciting content for me. The paper's definition of speaker verification vs identification was not as clear for me as it was for other readers. After a brief discussion, we collectively concluded that this might be due to people not considering this a set problem (i.e., verifying against a known number of people for access vs verifying against potentially every person), hence why complexity would increase with the latter task. Stefan chose this paper for the journal club to expose us to a more foundational paper and to introduce us to approaches that we might not enjoy when writing papers. I took from this paper to not promise the premise of a tutorial when not giving exactly that, and also to ensure that my papers are more focused on a narrow area.

Journal Club 9

Paper title	Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond
Presenter	Fritz Peters
Date	15/4/2024

Commentary bullets	• Felt that the introduction was very underwhelming:
	  ◦ Failed to provide a clear concrete example of how causality is different from prediction (to the point where I had to look elsewhere!)
	  ▪ Assumed that causality is what would happen to Y if we change X, whereas prediction is knowing X predict the next Y.
	  ▪ Good description of confounding variables.
	  ◦ "We cannot rely on the usual assumption that training and test data are identically distributed"
	  ▪ Besides the lack of clarity about what this means, does anyone actually assume this?
	  ◦ The "Blackbox metaphor" of DNN is tired and I am ready for it to never be published again
	  ▪ It actively paints DNNs as "unknowable"
	  ▪ Models are not trained for reasoning, just predictors.
	  • When we make a cup of tea, we rarely reason why we put the milk in after the tea has brewed, we do it because we always have done. Only when it goes wrong, do we start to reason why it was done that way. Why is it a shock to anyone that models that predict do the same? .
	• Good mathematical background given, especially as someone who had little causality experience, along with clear mathematical notation.
	• Provides me with a new approach to looking at NLP and approaching it in a way that might lead to more interpretable results.
	• Style: Lots of words used to outline what will be explained (at some point…), rather than bait the reader, structure your content better.

Narrative	This was an hour long session with Fritz Peters presenting and Yao Xiao chairing the journal club. The presentation style was well given, with key points of the paper abstracted which really improved understanding. While this paper had a really interesting topic, the majority of participants did not enjoy how this paper presented arguments, with the consensus being interesting area, however poor execution. The authors did however present a convincing reason why casuality needs to be considered when creating NLP applications. A shortfall, I felt, was that saying that current NLP models don't consider this is a bit reductive as well, they are not designed to do this. As no comparative evaluation models were provided which showed the improvement that casualty has over correlational approaches, it still remains to be seen if this approach is better. Additionally attempting to shoe horn in causality as an improvement on correlational models wasn't fully evaluated. When we do things in real life we don't always consider causality, when say making a cup of tea, we rather rely on previous experiences to create predictions about the process. The paper would have benefited from more examples, particularly comparative examples between causality and correlation. It did however achieve its goal for us to consider modelling from a different approach.

Journal Club 10

Paper title	Self-Supervised Speech Representation Learning: A Review
Presenter	Yao Xiao
Date	13/5/2024
Commentary bullets	• A great resource for speech researchers interested in self-supervised speech representation Learning. • Had to use external resources to more clearly define the exact difference between semi-supervised and self-supervised learning. • Why use the term generative approaches when it already exists (and it widely used)? • How would contrastive learning approach multiclass problems? • Good demonstration of increasing dataset resulting in diminishing returns after a certain point/ choice of dataset etc. • Zero-resource speech technologies was interesting, as it is is obviously possible for speech to be learnt without a joint representation (i.e. people who are blind from birth), however in that scenario they still have other joint representations of the speech (just not text). • Great outline as to why speech problem is different to nlp.
Narrative	This was an hour and half long in person journal club presented by Yao, and chaired by Boxuan. Yao's presentation style was impressive, and she managed to reduce the more complex parts of the paper into an easier to understand format, without trivialising the content. The presentation itself provided me with an accurate definition of codebook vectors, which was something that had been previously brought up in the journal clubs but not fully explained - a fixed-sized table of embedding vectors learnt by generative models. What I found particularly interesting about the discussion was the contrastric predictive coding, which is a n -1 problem where you group your classes into whats included and excluded. This approach might not work with wave2vec which is a n binary classification problem. Furthermore determining true positive and negative samples with CPC would surely still be a semi-supervised machine learning approach, as it would need labeled instances to begin with. Yao had no questions on her presentation. In general the group liked the paper, and felt it provided a well structured. It was noted by Anton (the JC chair) that the maths had been very simplified in the paper as it is meant to provide a general approach to self-supervised learning, which at times, comes at the expense of accuracy. There was some people who did not like some of the tables, which were presented poorly. We discussed if multimodal joint representations is a prerequisite for unsupervised learning in speech (which, can be done by blind from birth people. We discussed the reasons for needing this type of learning, given that if you add in all the possible variations of the formants of speech with the general population, it quickly scales to enormous scale. We also discussed if layers represent specific learning, or if this is just us wanting there to be a pattern to this.

Journal Club 11

Paper title	Attention Is All You Need
Presenter	Ian W Kennedy
Date	13/5/2024

Commentary bullets	• For a given hyperparameter h, wouldn't the linear projection within the multihead attention of VKQ be very similar? Additionally, what is the effect of higher values of h (besides increased computational power)?
	• Are transformers just finite hidden state RNNs? If so, given unlimited computational power, would RNNs be more performant than transformers, as it captures more information about the hidden states?
	• This was a paper builds on what I would consider foundational papers for modern NLP, and was for people who are interested in NLP, and had a presumed background knowledge (i.e. Seq2Seq & Jointly Learning to align and translate, which was great for me!). Can see value for speech people in it to, considering lots of these approaches have been adopted by them.
	• Interesting comments by the reviewers in NIPs 2017.
	  ○ They mention the papers lack of statistical testing conducted on the ablation studies on hyperparameters.
	  ○ "While none of the underlying techniques here are strikingly novel in themselves, the combination of them and the details necessary for getting it to work as well as LSTMs is a major achievement".
	• While the title is meant to be catchy, is attention really all you need?
	• While it reduces the context vector bottleneck, it doesn't completely remove it, meaning that large context window documents would still be challenging (despite claims of infinite length context windows…).
	• What other features are needed to improve performance - reasoning, structural inductive biases?

Narrative	This was an hour-long in-person discussion with Jason Chan, who chaired the meeting. Discussion centred around whether the paper justified its choices well. While the paper was not intended to be a review paper, it did not have much analysis of the topics presented within the paper (such as why cos/sine was used over techniques such as CNN positional encodings or rotary positional encodings). While the paper conducted ablation studies on the number of heads used for attention, more discussion was needed about why this information performance decreases after a certain number of heads. Additionally, we discussed whether attention is the only mechanism used for the next token prediction, as it does not mimic well what humans do within sentence generation. The presentation style was short yet very informative, which was very suitable for the group that had previous knowledge of this paper, which is considered one of the most impactful papers in the field. The discussion also occurred on whether the issue of fixed context vectors being a bottleneck is fully solved with the multihead attention / positional attention mechanism.

Meta-MOOC

For each block, record

- *the topic you worked on (including links to the relevant Meta-MOOC platform pages) and the other person / people you worked with*
- *A brief description of your individual contributions to each topic you have been involved with, including:*
  - *a list of the resources (websites, videos, lecture slides, etc) you considered for inclusion in the Meta-Mooc on this topic plus an indication of which you chose to include and why, and which you chose to exclude and why*
  - *an indication of what additional material (annotations, comment, structuring) you contributed to the Meta-Mooc on this topic.*
- *Make sure you timestamp each part of the block narrative to show how your work built up over the 8-week block period*
- *See the information about worklogs for additional guidance*

Add additional tables as required.

Block 1

Topic	Speech & NLP in Healthcare
Group	Aaron, Fritz, Cliodhna, Paul
Dates	Before December
Slides	• Page: Challenges for Speech and NLP in healthcare. • Fixed broken links/grammatical issues. I renamed/restructured the page to address speech centricity. Five new challenges of NLP within healthcare were added. • Refactored existing linked programme, which produces up-to-date reviews of all publicly available clinical NLP.

Block 2

Topic	Dialogue Systems
Group	Aaron, Fritz, Jack
Dates	December to 23 February 2024

Slides	Completed Jobs:
	• Frame-based dialogue systems: merged with Dialogue-state architecture
	  ○ updated/fixed inks
	  ○ Added resources section
	• Dialogue System Issues
	  ○ Created new page containing table with links to resources on ethical/implementation issues with dialogue systems
	• Examples of dialogue systems
	  ○ Reviewed resources embedded within prose on Chatbots page
	  ○ Created table incorporating these resources and my comments
	  ○ Found and added resources relating to ChatGPT, and further resources on ELIZA
	
	Things to do:
	• Restructuring remaining pages to remove large blocks of text
	• Further resources in examples of dialogue systems, particularly for spoken dialogue systems
	• Expansion of dialogue system issues page

Block 3

Topic	Prompting Methods in NLP
Group	Aaron, Ian, Tom
Dates	23rd February till 29th April 2024
Slides	Completed Jobs: ● Added new page under techniques outlining Zero/Few shot approaches ● Along with relevant FS/ZS papers, and video lectures found online. ● Under techniques, added links to the existing sub-pages, and links to four review papers on prompt engineering. ● Updated the work log for this page ● This page requires lots of work still, as the area is currently undergoing a lot of research! Things to do: ● Restructuring remaining pages to remove large blocks of text ● Reformatting Prompt Tuning and Calibration

Technical Training

- *A list of all relevant training courses (e.g., HPC, RSE, programming, ML, etc) undertaken with evidence of completion. Include course title, provider, duration, dates.*
- *Reflection on lessons learned in the context of your group work and research project.*
- *A review of where the learned skills have been applied, with evidence.*

Add additional tables as required.

Training Course 1 - 2 ½ hours

Title	Basics of Supervised Machine Learning
Provider	Research Computing Training
Duration	2 ½ Hours
Dates	6/10/2023
Narrative	This training course was a lecture combined with practical elements, focusing on machine learning and supervised learning basics. It covered foundational aspects, such as features, target values, and classifying different machine learning approaches (supervised, unsupervised and reinforcement). The lecturer outlined how linear regression describes the data trend and how it can be used to predict quantities. They also presented classifiers as a way of distinguishing data points. The lecturer's high-level overview of how to approach machine learning tasks was appropriate and useful, and breaking each stage down and presenting the reasons behind each was informative. I found the amount of knowledge I had retained from my previous MSc module on machine learning and felt it helped me understand some of the more complex questions presented in the training course. The training course was aimed at people who may have yet to gain experience with the statistical methods presented however, as a recap, it felt sufficiently useful. The packages used were sklearn, which I had less experience with than tensor flow however, this was a relatively easy learning curve. Some topics I had previously learned were covered - such as ● Regression vs Classification. ● Data splitting for training/test datasets. ● Data normalisation and standardisation. ● Using trained linear regression models to predict new values. ● Use of evaluation statistics, such as R squared, MSE, etc. ● Concepts of overfitting/underfitting and its impacts on generalisation. New concepts gained from the session ● Importance of training R squared being similar to test R squared, and its value in generalisability. ● Multicollinearity is a concept I had a vague idea about; however, I can now understand how multicollinearity inflates values such as R squared, and using other metrics such as adjusted R squared would be better. ○ I.e. the more independent variables you add to a machine learning model, the more accurate your model becomes. The lecture/training style of the presenter was very good and inclusive. The session felt rushed at no point, and their use and interactivity were exceptional. Every person in the room contributed.

Training Course 2 - 2 ½ hours

Title	Advanced Python: Language Features
Provider	LinkedIn Learning
Duration	2 ½ hours
Dates	15/10/2023
Narrative	This was an online lecture format, which was provided through the Linkedin Learning platform. I elected to do this training to improve my Python coding ability, which, while I think is currently satisfactory, I aim to improve the comprehensibility of my future code. The lecture format allowed for ad-hoc training, which I enjoyed, as I completed a section at a time, with a break in between. The lectures covered: ● Python coding style, current standards such as PEP 8. ● Some idiosyncracies of Python, such as strings vs bytes processing, boolean type casting and the walrus operator (an operator which can be used to increase comprehensibility of while loops). ● Advanced functions of Python, such as how to implement document strings (which, while I had previously understood, did explain its importance in helper functions) and lambda functions (a way of creating a small localised function which increases comprehensibility). ● Comprehensions (List, Dictionary and Set). I have previous knowledge of them but needed to understand their syntax fully. ● Classes of objects, such as enumerations (a way of binding values to a symbolic variable to increase comprehensibility and reduce computation), overriding unique methods, such as __repr__ and __string__, and how to enable object operations, such as obj1 + obj2 with overriding of comparison operations. ● Structural Pattern Matching, such as match, pattern guards, and sequence patterns. The lecture has taught me new Python features and will improve my ability to communicate with other researchers. I have started implementing features, such as list comprehensions in the lab sessions for each course I am on and conforming to a Python coding style.

LinkedIn Learning

Advanced Python: Language Features

Course completed by Aaron Fletcher
Oct 15, 2023 at 07:37PM UTC • 2 hours 20 minutes

Top skills covered

Python (Programming Language)

Head of Content Strategy, Learning

COURSE COMPLETION

Certificate ID: 6db5fb616c2473e55bd597689c6c778645d071810b4d7fddeaf1d8674d055d6d

Training Course 3 - 2 ½ hours

Title	Temporal Analysis in Python
Provider	Research Computing Training
Duration	2 ½ Hours
Dates	27/10/2023
Narrative	This was an in-person training session on the use of Python for temporal analysis of data. This is something that I was very keen to participate in, as my background is in medical clinical notes, which follow a temporal pattern; I was looking to gain further insight into this and how it might further my understanding in that area. The session started with a brief introduction to the problem, with examples being lifted from astrophysics. The majority of the terms outlined I had covered in my other module (Signal Processing), however, this formed as a refresher/revision for that. Some terms were not the same, however, that was because multiple key terms were different in each speciality. The session was highly interactive, with the presenter posing a mixture of theoretical and practical questions, such as the role of imputation/ how to deal with data gaps. From previous reading, this is quite a controversial topic, and the imputation of missing data (i.e. the act of determining a missing data value from related data) is an area that I need to research further. A key novel point addressed was that conversion into a stationary process (i.e. detrending through differencing) removes observations from the dataset, which can have unintended effects (such as removing a key factor in your analysis). The next part included a more intuitive understanding of the Fourier transformation, which is something that I will likely be using extensively over the next few months. I had previously explained this through mathematical terms, however, this presentation showed it through images and allowed me to understand the Fourier transformation graphs based on their input diagrams. A particularly impactful aspect of this training was that the presented actively asked questions, and if we raised questions, it showed us how to test what we thought through the use of Google Colab. Key topics covered: • Temporal data, and how it differs from other types of data • Stationary vs Non stationary data • Fourier Analysis • Hanning Windows • Types of noise • Practical applications of this through NLTK other packages within python.

Training Course 4 - 3 hours

Title	Supervised Machine Learning 2
Provider	Research Computing Training
Duration	3 Hours
Dates	15/1/2024
Narrative	This was an online training session provided by the research computing training at Sheffield University as a follow-up to the supervised machine learning I undertook in October 2023 (Basics of Supervised Machine Learning). The session covered the history behind the neural network, its attempts to approximate the human biology of the neuron, and early entries into the domain, such as perception. It covered basics such as the overarching topology of the domain and attempted to outline the differences between artificial intelligence and machine learning. The teaching catered to people with limited or no knowledge of neural networks and did not focus on the mathematical principles behind them. This meant the training was less relevant to me than formal machine learning training I had already undertaken. What was new to me was the TensorFlow playground, which the lecturer linked. It is an online toy program that allows users to add additional neurons to see its effects on a classification problem. It had different issues and succinctly outlined why solving non-linear classification problems with logistic regression is difficult. The lecturer gave problems to solve during the session, with work examples. The examples were relatively basic, yet they were a good introduction to using the tensorflow package, which I have limited experience with compared to the other packages, such as Pytorch. As other research papers can use Tensorflow, it was relevant to my overall PhD from this aspect. This was an online session, which felt less interactive compared to the in-person session and could discuss problems and ask questions. The lecturer answered the questions presented; however, due to the online nature, I got less answers to these questions and interactions. I still have questions regarding the impact of depth and width on neural networks and how to approximate the neuron by removing/ editing neurons not used to solve the issue (such as apoptosis from human biology). The refresher on activation functions felt relevant and useful. Key topics covered: • Approach to classifiers vs prediction • Hyperparameter optimisation • Activation functions • Supervised machine learning

Training Course 5 - 2 ½ hours

Title	Python: Design Patterns
Provider	LinkedIn Learning
Duration	2 ½ hours
Dates	23/06/2024
Narrative	I recently completed a two-and-a-half-hour course on LinkedIn Learning that focused on design patterns in Python. I enrolled in this course to improve my ability to communicate my code effectively with others, especially because I am involved in open research. My code must be well-organized and easy to understand, ensuring accessibility for others. The training course covered various approaches to designing patterns within Python, such as creational patterns, structural patterns, and behavioural patterns. Creational patterns include factories, abstract factories, singleton, builders, and prototypes. Factories are used when you are uncertain about the types of objects the system needs or when your application needs to decide what is used at runtime. Aibstract factories take this further when users expect to receive related objects at runtime. Singletons are used when you only want one object to be created from a class and to share global variables, which allows multiple object instances to share the same state. Builder patterns attempt to prevent the anti-pattern of telescoping constructions (where an object has a complex number of constructors) and prototype clone objects according to a prototypical instance. Structural patterns include decorators, proxies, adapters, composites, and bridges. Decorators add additional functions to established functions or objects without modifying their code. Proxies attempt to postpone object creation in situations with high resource requirements and create objects when necessary. Adapters convert the interface of classes into another one the client expects. Composites maintain a tree data structure to represent part-whole relationships. Finally, bridges attempt to untangle unnecessarily complicated class hierarchies. Behavioural patterns include observers, visitors, bridges, and more. Observers can monitor for changes in the subject by constructing a one-to-many relationship between the subject and the observer. Visitors add new features to a class hierarchy without changing it. Iterators allow sequential access to elements of an object without changing it, allowing augmentation of it before iteration. Moving forward, I plan to use decorators, as they allow me to modify existing codebases, which means I can reuse other people's code without modifying it too much. Additionally, I have already used the iteration pattern to simplify my classes, only giving access to the required information when iterating. This LinkedIn Learning course has been useful in developing my coding skills.

LinkedIn Learning

Python: Design Patterns

Course completed by Aaron Fletcher
Jun 23, 2024 at 08:35PM UTC • 2 hours 12 minutes

Top skills covered

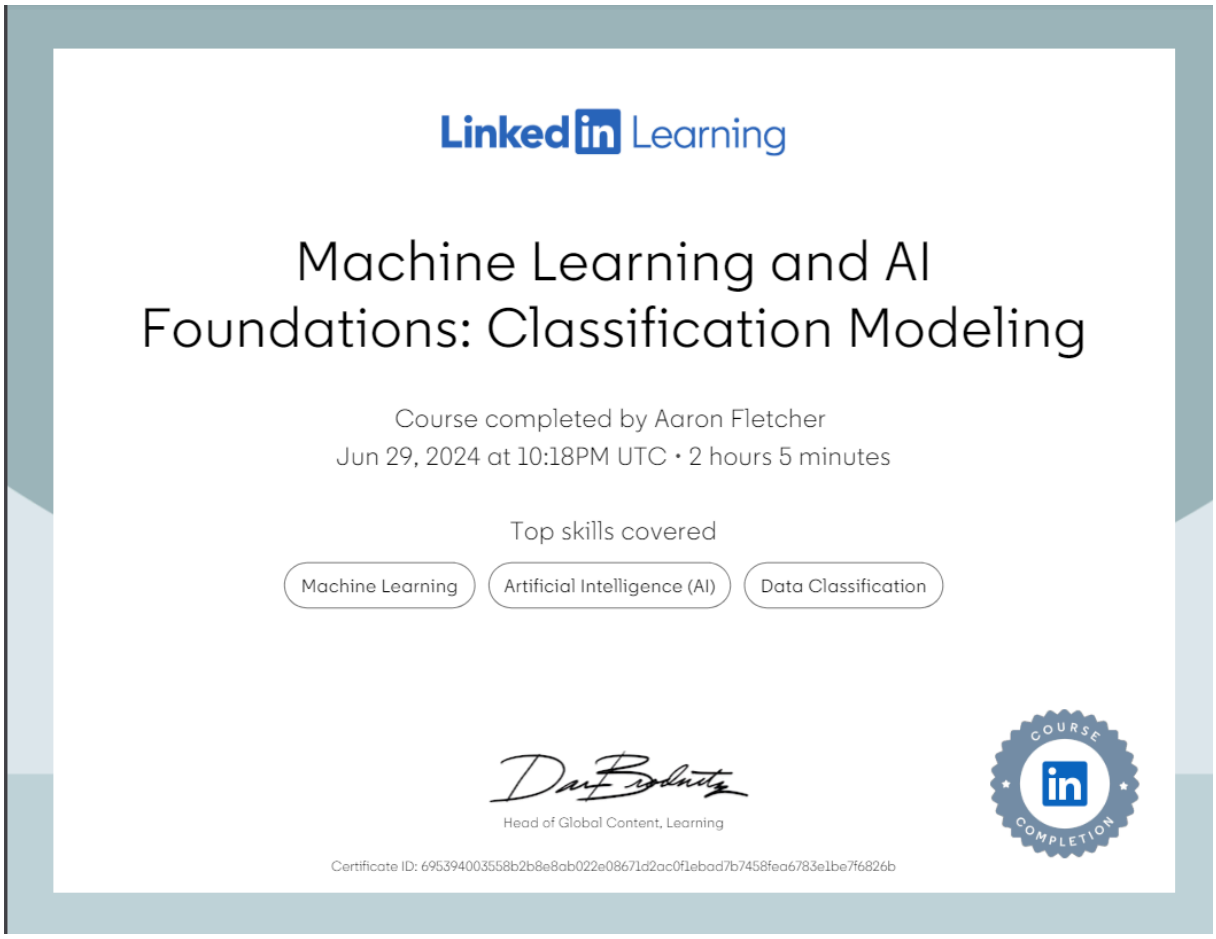
Design Patterns | Python (Programming Language)

Head of Global Content, Learning

Certificate ID: aa8c8bfaebc377502f94e618b85b0128346935732a1e1bafae22738a3ff6a1de

Training Course 6 - 2 hours

Title	Machine Learning and AI Foundations: Classification Modeling
Provider	LinkedIn Learning
Duration	2 Hours
Dates	29/06/2024
Narrative	This was a 2-hour online course linked to learning, which covered the basics of classification modelling. I chose this course to ensure that my understanding of classification topics matches that of other sources. The topics covered in the learning were: Classification Strategies Model Evaluation ● Confusion Matrics ● Lift Charts ● Gains Charts Classification Algorithms ● Stepwise Discriminant ● Logistic Regression ● Decision Trees ● KNN ● Linear SVMs ● Neural Networks ● Baysian Networks In honesty, this course was poorly given. While it provided little information that I was not aware of (which would mean that I should have selected a better course) by the time that the course was outlining its core teachings, about ¼ of the training course had been completed. For future training, I would ensure that I understand exactly what is being taught and ensure that I am the target audience. Additionally, some of the questions given in the chapter quiz were just plain incorrect (Groups other than healthcare professionals and social scientists can use linear regression, for example!).



LinkedIn Learning

Machine Learning and AI Foundations: Classification Modeling

Course completed by Aaron Fletcher
Jun 29, 2024 at 10:18PM UTC • 2 hours 5 minutes

Top skills covered

Machine Learning | Artificial Intelligence (AI) | Data Classification

Head of Global Content, Learning

Certificate ID: 695394003558b2b8e8ab022e08671d2ac0f1ebad7b7458fea6783e1be7f6826b

COURSE COMPLETION

References

[1] Peter Kranke. "Evidence-based practice: how to perform and use systematic reviews for clinical decision-making". In: *European Journal of Anaesthesiology* 27.9 (Sept. 2010), pp. 763–772. ISSN: 1365-2346. DOI: 10.1097/EJA.0b013e32833a560a (cit. on p. 7).

[2] *Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009)*. Type: Web Page. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009 (visited on 07/29/2024) (cit. on p. 7).

[3] Jennifer A. Swanson, DeLaine Schmitz, and Kevin C. Chung. "How to Practice Evidence-Based Medicine". In: *Plastic and reconstructive surgery* 126.1 (July 2010). tex.pmcid: PMC4389891, pp. 286–294. ISSN: 0032-1052. DOI: 10.1097/PRS.0b013e3181dc54ee. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4389891/ (visited on 07/29/2024) (cit. on p. 7).

[4] Gordon H. Guyatt et al. "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations". In: *BMJ (Clinical research ed.)* 336.7650 (Apr. 2008). tex.copyright: © BMJ Publishing Group Ltd 2008, pp. 924–926. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.39489.470347.AD. URL: https://www.bmj.com/content/336/7650/924 (visited on 07/29/2024) (cit. on p. 7).

[5] "Cochrane Handbook for Systematic Reviews of Interventions". In: (cit. on p. 7).

[6] Asghar Ghasemi et al. "Scientific Publishing in Biomedicine: A Brief History of Scientific Journals". In: *International Journal of Endocrinology and Metabolism* 21.1 (Dec. 2022). tex.pmcid: PMC10024814, e131812. ISSN: 1726-913X. DOI: 10.5812/ijem-131812. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10024814/ (visited on 07/29/2024) (cit. on pp. 7, 9).

[7] Jeremy Howick. "Front Matter". In: *The Philosophy of Evidence-Based Medicine*. John Wiley & Sons, Ltd, 2011, pp. i–xiv. ISBN: 978-1-4443-4267-3. DOI: 10.1002/9781444342673.fmatter. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444342673.fmatter (visited on 07/29/2024) (cit. on p. 7).

[8] Gehad Mohamed Tawfik et al. "A step by step guide for conducting a systematic review and meta-analysis with simulation data". In: *Tropical Medicine and Health* 47.1 (Aug. 2019), p. 46. ISSN: 1349-4147. DOI: 10.1186/s41182-019-0165-6. URL: https://doi.org/10.1186/s41182-019-0165-6 (visited on 07/29/2024) (cit. on pp. 8, 9).

[9] Ian Shemilt et al. "Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews". In: *Systematic Reviews* 5.1 (Aug. 2016). tex.pmcid: PMC4989498, p. 140. ISSN: 2046-4053. DOI: 10.1186/s13643-016-0315-4 (cit. on pp. 8, 9).

[10] Melita J. Giummarra, Georgina Lau, and Belinda J. Gabbe. "Evaluation of text mining to reduce screening workload for injury-focused systematic reviews". In: *Injury Prevention* 26.1 (Feb. 2020). tex.copyright: © Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ., pp. 55–60. ISSN: 1353-8047, 1475-5785. DOI: 10.1136/injuryprev-2019-043247. URL: https://injuryprevention.bmj.com/content/26/1/55 (visited on 07/29/2024) (cit. on p. 8).

[11] Miguel Marques Antunes et al. "Preoperative statin therapy for adults undergoing cardiac surgery - Marques Antunes, M - 2024 — Cochrane Library". In: (). ISSN: 1465-1858. URL: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD008493.pub5/full (visited on 07/29/2024) (cit. on p. 9).

[12] B. Nussbaumer-Streit et al. “Resource use during systematic review production varies widely: a scoping review”. In: *Journal of Clinical Epidemiology* 139 (Nov. 2021), pp. 287–296. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2021.05.019. URL: https://www.sciencedirect.com/science/article/pii/S0895435621001712 (visited on 07/29/2024) (cit. on p. 9).

[13] Justin S. Smith et al. “Less is more: Sampling chemical space with active learning”. In: *The Journal of Chemical Physics* 148.24 (May 2018), p. 241733. ISSN: 0021-9606. DOI: 10.1063/1.5023802. URL: https://doi.org/10.1063/1.5023802 (visited on 07/30/2024) (cit. on p. 10).

[14] Steven C. H. Hoi et al. “Batch mode active learning and its application to medical image classification”. In: *Proceedings of the 23rd international conference on Machine learning*. ICML ’06. New York, NY, USA: Association for Computing Machinery, June 2006, pp. 417–424. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143897. URL: https://doi.org/10.1145/1143844.1143897 (visited on 07/30/2024) (cit. on p. 10).

[15] Maura R. Grossman and Gordon V. Cormack. “Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review Annual Survey”. In: *Richmond Journal of Law and Technology* 17.3 (2010), [i]–48. URL: https://heinonline.org/HOL/P?h=hein.journals/jolt17&i=471 (visited on 07/31/2024) (cit. on p. 10).

[16] Dana Angluin. “Queries and Concept Learning”. In: *Machine Learning* 2.4 (Apr. 1988), pp. 319–342. ISSN: 1573-0565. DOI: 10.1023/A:1022821128753. URL: https://doi.org/10.1023/A:1022821128753 (visited on 07/31/2024) (cit. on p. 10).

[17] Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. *A Novel Question-Answering Framework for Automated Abstract Screening Using Large Language Models*. tex.copyright: © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution-NonCommercial-NoDerivs 4.0 International), CC BY-NC-ND 4.0, as described at http://creativecommons.org/licenses/by-nc-nd/4.0/. June 2024. DOI: 10.1101/2023.12.17.23300102. URL: https://www.medrxiv.org/content/10.1101/2023.12.17.23300102v3 (visited on 06/27/2024) (cit. on p. 10).

[18] David D. Lewis and William A. Gale. “A Sequential Algorithm for Training Text Classifiers”. In: *SIGIR ’94*. Ed. by Bruce W. Croft and C. J. van Rijsbergen. London: Springer, 1994, pp. 3–12. ISBN: 978-1-4471-2099-5. DOI: 10.1007/978-1-4471-2099-5_1 (cit. on p. 10).

[19] Pengzhen Ren et al. *A Survey of Deep Active Learning*. Dec. 2021. DOI: 10.48550/arXiv.2009.00236. URL: http://arxiv.org/abs/2009.00236 (visited on 07/31/2024) (cit. on p. 11).

[20] Ron Artstein and Massimo Poesio. “Survey Article: Inter-Coder Agreement for Computational Linguistics”. In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004 (visited on 07/31/2024) (cit. on p. 11).

[21] Tim Pearce, Alexandra Brintrup, and Jun Zhu. *Understanding Softmax Confidence and Uncertainty*. June 2021. DOI: 10.48550/arXiv.2106.04972. URL: http://arxiv.org/abs/2106.04972 (visited on 07/31/2024) (cit. on p. 11).

[22] Dan Wang and Yi Shang. “A new active labeling method for deep learning”. In: *2014 International Joint Conference on Neural Networks (IJCNN)*. July 2014, pp. 112–119. DOI: 10.1109/IJCNN.2014.6889457. URL: https://ieeexplore.ieee.org/document/6889457/?arnumber=6889457 (visited on 07/31/2024) (cit. on p. 11).

[23] Chuan Guo et al. *On Calibration of Modern Neural Networks*. Aug. 2017. DOI: 10.48550/arXiv.1706.04599. URL: http://arxiv.org/abs/1706.04599 (visited on 07/31/2024) (cit. on p. 11).

[24] Burr Settles. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences, 2009. URL: https://minds.wisconsin.edu/handle/1793/60660 (visited on 07/31/2024) (cit. on p. 11).

[25] Gordon V. Cormack and Maura R. Grossman. *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*. Apr. 2015. DOI: 10.48550/arXiv.1504.06868. URL: http://arxiv.org/abs/1504.06868 (visited on 06/27/2024) (cit. on pp. 11, 13).

[26] A.M. Cohen et al. "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification". In: *Journal of the American Medical Informatics Association : JAMIA* 13.2 (2006). tex.pmcid: PMC1447545, pp. 206–219. ISSN: 1067-5027. DOI: 10.1197/jamia.M1929. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447545/ (visited on 08/01/2024) (cit. on p. 12).

[27] Gerbrich Ferdinands et al. "Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records". In: *Systematic Reviews* 12 (June 2023). DOI: 10.1186/s13643-023-02257-7 (cit. on p. 12).

[28] Gordon V. Cormack and Maura R. Grossman. "Evaluation of machine-learning protocols for technology-assisted review in electronic discovery". In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. Gold Coast Queensland Australia: ACM, July 2014, pp. 153–162. ISBN: 978-1-4503-2257-7. DOI: 10.1145/2600428.2609601. URL: https://dl.acm.org/doi/10.1145/2600428.2609601 (visited on 07/31/2024) (cit. on p. 12).

[29] Gordon V. Cormack and Maura R. Grossman. "Systems and methods for conducting a highly autonomous technology-assisted review classification". U.S. pat. 20160371261A1. Individual. Dec. 2016. URL: https://patents.google.com/patent/US20160371261A1/en (visited on 08/04/2024) (cit. on p. 14).

[30] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 08/04/2024) (cit. on p. 13).

[31] Eugene Yang et al. *Goldilocks: Just-Right Tuning of BERT for Technology-Assisted Review*. Jan. 2022. DOI: 10.48550/arXiv.2105.01044. URL: http://arxiv.org/abs/2105.01044 (visited on 06/27/2024) (cit. on p. 13).

[32] Y. Xu et al. *Forget Me Not: Reducing Catastrophic Forgetting for Domain Adaptation in Reading Comprehension*. Nov. 2020. DOI: 10.48550/arXiv.1911.00202. URL: http://arxiv.org/abs/1911.00202 (visited on 08/04/2024) (cit. on p. 13).

[33] *ielab/goldilocks-reproduce*. June 2024. URL: https://github.com/ielab/goldilocks-reproduce (visited on 07/31/2024) (cit. on p. 13).

[34] *BMC Medical Research Methodology*. BioMed Central. URL: https://bmcmedresmethodol.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research-article (visited on 11/13/2024) (cit. on p. 22).

[35] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. *LinkBERT: Pretraining language models with document links*. 2022. arXiv: 2203.15827[cs.CL]. URL: https://arxiv.org/abs/2203.15827 (cit. on p. 23).

[36] Xinyu Mao, Bevan Koopman, and Guido Zuccon. "A Reproducibility Study of Goldilocks: Just-Right Tuning of BERT for TAR". In: *Advances in Information Retrieval*. Ed. by Nazli Goharian et al. Vol. 14611. Cham: Springer Nature Switzerland, 2024, pp. 132–146. ISBN: 978-3-031-56065-1 978-3-031-56066-8. DOI: 10.1007/978-3-031-56066-8_13. URL: https://link.springer.com/10.1007/978-3-031-56066-8_13 (visited on 07/31/2024) (cit. on pp. 15, 23).

[37] Yu Gu et al. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. version: 6. Sept. 16, 2021. arXiv: 2007.15779. URL: http://arxiv.org/abs/2007.15779 (visited on 11/13/2024) (cit. on p. 23).

[38] Di Jin et al. *What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams*. version: 1. Sept. 28, 2020. arXiv: 2009.13081. URL: http://arxiv.org/abs/2009.13081 (visited on 11/13/2024) (cit. on p. 23).

[39] Dan Hendrycks et al. *Measuring Massive Multitask Language Understanding*. Jan. 12, 2021. arXiv: 2009.03300. URL: http://arxiv.org/abs/2009.03300 (visited on 11/13/2024) (cit. on p. 23).

[40] Carol Lefebvre et al. "Cochrane handbook for systematic reviews of interventions". In: *Oxfordshire, UK: The Cochrane Collaboration* (2011) (cit. on p. 26).

[41] Jo Akers, R Aguiar-Ibáñez, and A Baba-Akbari. "Systematic reviews: CRD's guidance for undertaking reviews in health care". In: *University of York* (2009) (cit. on p. 26).

[42] Simon Briscoe, Alison Bethel, and Morwenna Rogers. "Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study". In: *Research Synthesis Methods* 11.2 (July 4, 2019), p. 169. DOI: 10.1002/jrsm.1355. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7079050/ (visited on 11/13/2024) (cit. on p. 26).

[43] *MECIR Manual — Cochrane Community*. URL: https://community.cochrane.org/mecir-manual (visited on 11/13/2024) (cit. on p. 26).

[44] Evangelos Kanoulas et al. "CLEF 2017 technologically assisted reviews in empirical medicine overview". In: *CEUR Workshop Proceedings* 1866 (Sept. 2017), pp. 1–29. ISSN: 1613-0073. URL: http://ceur-ws.org/Vol-1866/ (visited on 07/31/2024) (cit. on p. 15).

[45] Evangelos Kanoulas et al. "CLEF 2018 technologically assisted reviews in empirical medicine overview: 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018". In: *CEUR Workshop Proceedings* 2125 (July 2018). ISSN: 1613-0073. URL: http://www.scopus.com/inward/record.url?scp=85051077484&partnerID=8YFLogxK (visited on 07/31/2024) (cit. on p. 15).

[46] Evangelos Kanoulas et al. "CLEF 2019 technology assisted reviews in empirical medicine overview". In: *CEUR Workshop Proceedings* 2380 (Sept. 2019). tex.copyright: cc_by. ISSN: 1613-0073. URL: https://strathprints.strath.ac.uk/71253/ (visited on 07/31/2024) (cit. on p. 15).

[47] Jonathan De Bruin et al. *SYNERGY - Open machine learning dataset on study selection in systematic reviews*. In collab. with Jonathan De Bruin and Rens Van De Schoot. 2023. DOI: 10.34894/HE6NAQ. URL: https://dataverse.nl/citation?persistentId=doi:10.34894/HE6NAQ (visited on 07/31/2024) (cit. on p. 15).

[48] Adam Roegiest et al. "TREC 2015 Total Recall Track Overview". In: *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*. Ed. by Ellen M. Voorhees and Angela Ellis. Vol. 500-319. NIST Special Publication. National Institute of Standards and Technology (NIST), 2015. URL: https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf (visited on 07/31/2024) (cit. on p. 16).

[49] Maura R. Grossman, G. Cormack, and Adam Roegiest. "TREC 2016 Total Recall Track Overview". In: 2016. URL: https://www.semanticscholar.org/paper/TREC-2016-Total-Recall-Track-Overview-Grossman-Cormack/126240dedd75626fd736f0485d06f1f516517e54 (visited on 07/31/2024) (cit. on p. 16).

[50] David D. Lewis et al. "RCV1: A New Benchmark Collection for Text Categorization Research". In: *Journal of Machine Learning Research* 5 (Apr 2004), pp. 361–397. ISSN: ISSN 1533-7928. URL: https://www.jmlr.org/papers/v5/lewis04a.html (visited on 07/31/2024) (cit. on p. 16).

[51] Alison O'Mara-Eves et al. "Using text mining for study identification in systematic reviews: a systematic review of current approaches". In: *Systematic Reviews* 4.1 (Jan. 2015), p. 5. ISSN: 2046-4053. DOI: 10.1186/2046-4053-4-5. URL: https://doi.org/10.1186/2046-4053-4-5 (visited on 08/02/2024) (cit. on p. 17).

[52] Guy Tsafnat et al. "Systematic review automation technologies". In: *Systematic Reviews* 3.1 (July 2014), p. 74. ISSN: 2046-4053. DOI: 10.1186/2046-4053-3-74. URL: https://doi.org/10.1186/2046-4053-3-74 (visited on 08/12/2024) (cit. on p. 17).

[53] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. 1st ed. Cambridge University Press, July 2008. ISBN: 978-0-521-86571-5 978-0-511-80907-1. DOI: 10.1017/CBO9780511809071. URL: https://www.cambridge.org/core/product/identifier/9780511809071/type/book (visited on 08/03/2024) (cit. on p. 17).

[54] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. Dec. 2020. DOI: 10.48550/arXiv.2004.05150. URL: http://arxiv.org/abs/2004.05150 (visited on 08/06/2024) (cit. on p. 19).

[55] Yikuan Li et al. "A comparative study of pretrained language models for long clinical text". In: *Journal of the American Medical Informatics Association* 30.2 (Feb. 2023), pp. 340–347. ISSN: 1527-974X. DOI: 10.1093/jamia/ocac225. URL: https://doi.org/10.1093/jamia/ocac225 (visited on 08/06/2024) (cit. on p. 19).

[56] Manzil Zaheer et al. *Big Bird: Transformers for Longer Sequences*. Jan. 2021. DOI: 10.48550/arXiv.2007.14062. URL: http://arxiv.org/abs/2007.14062 (visited on 08/06/2024) (cit. on p. 20).

[57] Jason Priem, Heather Piwowar, and Richard Orr. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. June 2022. DOI: 10.48550/arXiv.2205.01833. URL: http://arxiv.org/abs/2205.01833 (visited on 06/14/2024) (cit. on p. 21).

[58] Malte Ostendorff et al. *Enriching BERT with Knowledge Graph Embeddings for Document Classification*. Sept. 2019. URL: http://arxiv.org/abs/1909.08402 (visited on 08/06/2024) (cit. on p. 21).

[59] Benjamin Wolff, Eva Seidlmayer, and Konrad U. Förstner. *Enriched BERT Embeddings for Scholarly Publication Classification*. May 2024. DOI: 10.48550/arXiv.2405.04136. URL: http://arxiv.org/abs/2405.04136 (visited on 08/06/2024) (cit. on p. 21).

[60] Parishad BehnamGhader et al. *LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders*. Apr. 2024. DOI: 10.48550/arXiv.2404.05961. URL: http://arxiv.org/abs/2404.05961 (visited on 08/06/2024) (cit. on p. 21).

[61] Yinqiong Cai et al. *CAME: Competitively Learning a Mixture-of-Experts Model for First-stage Retrieval*. Nov. 2023. URL: http://arxiv.org/abs/2311.02834 (visited on 08/16/2024) (cit. on p. 22).

[62] Hideitsu Hino and Shinto Eguchi. *Active Learning by Query by Committee with Robust Divergences*. Nov. 2022. DOI: 10.48550/arXiv.2211.10013. URL: http://arxiv.org/abs/2211.10013 (visited on 08/16/2024) (cit. on p. 22).