Natural Language Processing Methods for Detecting and Measuring the Impact of Scientific Work Beyond Academia

by

James Edward Ravenscroft

Thesis

A thesis submitted to the University of Warwick in partial

fulfilment of the requirements for the degree of

Doctor of Philosophy in Urban Science

Warwick Institute for the Science of Cities

September 2022

Contents

List of Tables

List of Figures

Acknowledgments

Firstly I would like to thank both of my PhD supervisors Prof. Maria Liakata and Dr. Amanda Clare for their continued guidance over almost a decade of working together. Thank you for taking me under your combined wings as a nervous undergraduate and for all of the support you have both given me ever since! I'm eternally grateful for the mentoring and guidance you have both imparted, your personal support and friendship and of course for introducing me to Lindyhop. A special thanks to Prof. Rob Proctor for providing a friendly face and helping me to navigate the Bermuda-triangle of uni admin during Maria's transition to QMUL.

I'd like to thank one of my co-authors and closest friends Dr Daniel Duma whose acquaintance I first made at LREC 2016. I appreciate all of your technical, professional and personal advice and your eternal optimism and showers of flattery have never failed to cheer me up over the years since.

I'd like to say a heartfelt thank you to my professional and academic collaborators including Arie Cattan, Prof. Ido Dagan, Albert Webson and the Turing Institute NLP group for patiently bouncing ideas around with me, reading paper drafts and helping me debug various models.

A huge thanks to Doug Ayres, Phil Westcott and the wider team at Filament who have been so patient with my insistent promises of "just one more year" for about 6 years while I've been working on my PhD part time in between business meetings, working lunches and occasional server fires.

Thanks to my personal network, family and friends for all your support through the good times and bad over the last few years. In particular, thanks to my mum Lesley Ravenscroft for proof-reading this document and getting involved with some of the data collection work too!

Last but by no means least I want to say thank you to my loving and ever patient wife Amy who, having met me after I began my studies in 2015, has only ever known me to be a PhD writing, zoom-call-taking workaholic. I'm so grateful to have you in my life and for your love and support over the last few years and for encouraging me to continue writing when I've been close to giving up. I hope that my completion of this PhD will free up more time for us to spend together.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author. Parts of this thesis have been published by the author:

- Ravenscroft, James, Amanda Clare, and Maria Liakata. *'Detecting and Measuring Scientific Impact with Multi-Task Few-Shot Learning'*. Pending Review, Transactions of the Association for Computational Linguistics, 2022.

- Ravenscroft, James, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. *'CD^{2}CR: Co-Reference Resolution across Documents and Domains'*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. https://www.virtual2021.eacl.org/paper_main.134.html

- Ravenscroft, James, Amanda Clare, and Maria Liakata. *'Measuring Prominence of Scientific Work in Online News as a Proxy for Impact'*. arXiv, 28 July 2020. https://doi.org/10.48550/arXiv.2007.14454

- Ravenscroft, James, Amanda Clare, and Maria Liakata. *'HarriGT: Linking News Articles to Scientific Literature'*. In Proceedings Of the 56th Annual Meeting Ofthe Association for Computational Linguistics-System Demonstrations, 19–24, 2018. http://aclweb.org/anthology/P18-4004

- Ravenscroft, James, Maria Liakata, Amanda Clare, and Daniel Duma. *'Measuring Scientific Impact beyond Academia: An Assessment of Existing Impact Metrics and Proposed Improvements'*. Edited by Lutz Bornmann. PLOS ONE 12, no. 3 (9 March 2017): e0173152. https://doi.org/10.1371/journal.pone.0173152

Abstract

Scientific research has a profoundly important impact on our society and the environment. However, the multifaceted nature of this impact makes it particularly difficult to measure and, as shown in this thesis, it cannot be measured using traditional academic impact metrics that focus on counting citations and publications. Furthermore, existing societal and environmental impact metrics are only applicable to one scientific discipline or geography or are expensive processes run irregularly by government agencies.

This thesis investigates natural language processing methods for identifying and measuring societal and environmental scientific impact and how such impact is reported in the news. A novel regression task and model are presented for identifying and quantifying this impact based on text extracted from scientific papers and news articles that discuss them. This is enabled by developing methods for linking and comparing news articles with academic papers that they discuss, whilst accounting for the structural and linguistic differences between the two types of document. Text encoding strategies for representation and comparison of long documents are also a focus of the thesis. A new cross-domain, co-reference resolution task between news articles and scientific papers is introduced so that co-referring entities may be used as anchors for aligning the two types of documents. Through comparisons of news article excerpts and sentences from corresponding scientific papers, it is shown that scientific discourse structure and argumentation in scientific papers is a likely predictor of which information will be presented prominently in news articles.

This work introduces several novel natural language task settings for which no pre-existing data sets exist. This has necessitated the production of new human-annotated datasets which were built using bespoke annotation tools that use semi-supervised learning to accelerate the labelling process and minimise the cognitive load of the task on the annotator. The thesis also makes use of low resource approaches including few-shot and multi-task learning to facilitate the development of accurate models with small data-sets. The resulting annotated data-sets, annotation tools and guidelines along with state-of-the-art machine learning models are all made available as open assets.

This thesis contributes new ways to measure societal and environmental impact of scientific work and help scientists and funding bodies understand how work is being used by others, justify the spending of public funding and inform better public engagement.

Abbreviations

AI Artificial Intelligence

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

BOW Bag of Words

CD^{2}CR Cros-Domain Cross-Document Co-reference Resolution

DOI Document Object Identifier

GloVe Global Vectors

IR Information Retrieval

JIF Journal Impact Factor

JSD Jensen-Shannon Divergence

NDCG Normalized Discounted Cumulative Gain

NER Named Entity Recognition

NLP Natural Language Processing

PET Pattern Exploitative Training

PLM Pre-trained Language Model

REF Research Excellence Framework

URL Uniform Resource Locator

1

Introduction

1.1 Motivation

We all benefit from the impacts of scientific progress every day. From the food we eat, to the medicines we take, to the smart phones that provide us with far more than voice communication, modern life would be unrecognisable without the impact of countless iterations of scientific progress over hundreds of years. Many people take these benefits for granted. However, there is growing interest amongst researchers and funding bodies in understanding and measuring the effects of scientific work upon the world around us. For scientists, the knowledge that one's work is impactful provides personal satisfaction and motivates continued scientific research and engagement. For research funding bodies, whose decisions are increasingly scrutinised in today's competitive economic climate, understanding the impact of funded research can help to justify spending and secure further investment in promising work.

The need for measuring the impact of scientific works within academia is widely recognised, with a number of widely-used instruments like h-index (Hirsch, 2005) and JIF (Garfield, 2006). These take into account the number of citations that the work receives or the prestige of the journal that the work was published in. However, these metrics only paint a partial picture of impact: the academic community's familiarity with the work. In order to see the full impact of the work, one must look beyond academia for evidence of the social, economic and political changes that scientific progress brings about. For example, research papers that lead to a new cancer treatment pathway; an efficient, low-carbon industrial process with multi-billion-dollar savings or to long-lasting changes in law & policy are clearly more impactful than the sum of their citations. We designate these types of impact 'comprehensive impact'.

Whilst it is clear *why* one should examine comprehensive impact, the *what*, *where* and *how* to measure it are much less clear. Although there are existing approaches for measuring this type of scientific impact (which we explore in depth in chapter 4), they leave much to be desired, requiring lengthy manual evaluation exercises or relying heavily upon difficult to access knowledge and data. Furthermore, today's metrics do not help us understand *what* makes a given scientific work more or less impactful outside of academia. Tools that help scientists to understand why a piece of work was impactful could also help them to generate more impact with their future works, improving public interest in lesser known disciplines and helping secure new funding.

This thesis sets out to explore how modern Natural Language Processing (NLP) and machine learning (ML) techniques can be used in combination with abundant and

freely-available data resources to identify and explain examples of scientific impact beyond academia. Our work involves the combination and processing of large, heterogeneous data sources and requires us to seek answers to a number of open methodological questions within NLP (many of which we outline in Section 1.3 below).

1.2 Research Questions

RQ1: To what extent does the academic impact of scientific work relate to its impact beyond academia (comprehensive impact)?

This research question and sub-questions are addressed in Chapter 3 (**RQ1.1, RQ1.2**), in which we provide a literature review of current widely-used scientometrics, and Chapter 4 (**RQ1.3**), in which we explore the carry out an empirical analysis to characterise a selection of scientometrics and the relationships between them.

- **RQ1.1** What are commonly used metrics for the academic impact of scientific work?
- **RQ1.2** What metrics exist for measuring impact beyond academia and, how widely are they used?
- **RQ1.3** What statistical relationships exist between widely used academic impact metrics and scientific impact outside of academia?

RQ2: How does evidence of impact beyond academia affect a scientific work's performance in terms of existing comprehensive impact metrics?

This research question and sub-questions are addressed in Chapter 5 in which we discuss sources of non-academic scientific impact and develop automated systems for identifying news articles that discuss scientific work in order to provide such evidence. We then carry out further empirical experiments, characterising the relationship between non-academic scientific impact and the presence or absence of corresponding news coverage. This research question relates to **CH1** data collection and annotation with multiple annotators, discussed in section 1.3 below.

- **RQ2.1** What data sources provide evidence of non-academic impacts of scientific work?
- **RQ2.2** How can we automate the extraction and linking of such evidence to the scientific work it relates to?
- **RQ2.3** In terms of current comprehensive impact metrics, how does the presence or absence of evidence of comprehensive impact affect a work's associated score?

RQ3: How can we identify semantically-similar statements in topically-aligned news articles and scientific papers despite the disjoint language use and grammatical styles of the two document types?

This research question and sub-questions are addressed in Chapter 6 in which we characterise the respective lengths, vocabularies and purposes of scientific papers and the news articles that discuss them before exploring techniques with which to compare them in order to understand information flow between the two document types. This research question relates to **CH4**: modelling long documents, discussed in section 1.3 below.

- **RQ3.1** What are the statistical characteristics of news articles and scientific papers and, how do they differ?
- **RQ3.2** What techniques can we use to best represent news articles and scientific papers for comparison?
- **RQ3.3** What contextual information do the document representations provide?

RQ4: How does information flow from scientific papers to the news articles that discuss them and what is the connection between REF impact scores and how works are presented in the news?

This research question and sub-questions are addressed in Chapter 7 in which we explore the ways in which the structure of scientific papers and the news articles that discuss them are linked to perception of importance. We develop techniques for identifying which information from scientific papers presented as most important within news articles that discuss them and explore how this relates to the works' non-academic impact levels. This research question is linked to Challenge **CH2** Cross-domain information alignment discussed in section 1.3 below.

- **RQ4.1** How do the structures and styles of news articles and scientific papers affect the perceived importance of information presented within them?
- **RQ4.2** Which information in scientific papers is most often presented as most important in news articles?
- **RQ4.3** What characteristics are different for newspapers with known links to REF Impact Case Studies?

RQ5: Can we identify co-referring entities, such as people, institutions and companies across news articles and scientific papers and use them to align semantically similar sentences despite the stylistic and linguistic differences between the documents?

This research question and subquestions are addressed in Chapter 8 in which we develop a cross-document co-reference resolution data set linking scientific papers and the news articles that discuss them. We then run a series of experiments on our new data set, using state-of-the-art models and provide an in-depth error analysis of our model. This research question is linked to Challenges **CH1**: Data Collection and Annotation with Multiple Annotators and **CH2**: Cross-domain Information Alignment which are discussed in more detail in section 1.3 below.

- **RQ5.1** How does cross-document co-reference resolution differ for documents in separate domains?
- **RQ5.2** How can pairs of co-referent phrases in news articles and scientific papers be identified and annotated efficiently and reliably?
- **RQ5.3** How well do current state-of-the-art cross-document co-reference resolution models perform at the cross-domain task and where do they struggle?

RQ6: How can we detect and quantify comprehensive impact implied by language use in scientific news articles and the academic works that they discuss?

This research question and subquestions are addressed in Chapter 9 in which we define a new regression-based scoring task for assessing the non-academic impact associated with sentences in news articles and we construct an associated dataset. We then train models to carry out the task, making use of state-of-the-art few-shot learning techniques to achieve good statistical performance despite the modest size of our training dataset. This research question is linked to Challenge **CH1**: Data collection and annotation with multiple annotations and **CH3**: Low resource learning, both discussed in detail in section 1.3 below.

- **RQ6.1** What is the current state-of-the-art approach for detecting impactful sentences in scientific news articles and press releases and what are its limitations?
- **RQ6.2** Can we define an appropriate task for quantifying comprehensive impact score based on text and can we reliably collect corresponding labelled data from multiple annotators?
- **RQ6.3** Is it possible to train a statistically performant, low-resource model for detecting and scoring text that implies comprehensive scientific impact?

1.3 Challenges

This thesis addresses the complex task of detecting and measuring evidence of comprehensive scientific impact in supporting documents aligned with scientific publications. Addressing this task required us to also address the following technical natural language processing challenges:

1.3.1 CH1 Data collection and annotation with multiple annotators

Collecting annotated data for use in supervised machine learning settings is a time-consuming and challenging task requiring the coordination of multiple human workers to build a dataset in a reasonable timeframe. Ensuring that data is consistent is the most significant challenge: if human annotators have developed divergent mental models of the task at hand, then their annotations will also likely be divergent which would impede model training. In this thesis, we assemble labelled datasets in Chapters 5, 8 and 9 and build easy-to-use annotation tools and illustrated guidelines documents which help annotators to label data consistently.

1.3.2 CH2 Cross-domain information alignment

Scientific news articles and the academic publications that they describe are written in very different styles and using different vocabularies despite sharing the same subject-matter (as we discuss in Chapter 6). In order to understand how faithfully a news article represents a scientific issue, we need to be able to align and compare sentences that describe the same thing. In this work we investigate two ways of achieving this: by using similarity between two sentences (see Chapter 7) and by using co-referring mentions of entities as contextual anchors (see Chapter 8).

1.3.3 CH3 Low resource learning

Supervised machine learning typically requires very large volumes of annotated data in order to train models which can generalise well to unseen inputs. In Chapter 9, we work with a dataset containing only $\sim$400 annotated examples (we used up all available time and budget owing to the significant amount of work that goes into collecting these annotations, hiring multiple annotators, iterating and revising guidelines over time and the difficulty of the annotation task itself). We make use of state-of-the-art transfer learning and multitask learning techniques in order to build a set of models that are able to carry out the impact scoring task with good statistical performance.

1.3.4 CH4 Modelling long documents.

Modelling and representing long sequences of text, such as scientific papers and news articles like those that we explore in this thesis, remain challenges within the field of Natural

Language Processing. Long documents can be represented using low fidelity encoding strategies like bag-of-words vectors, but contextual information about word order, relations between words and topics are necessarily lost. Modern deep learning approaches that make use of attention mechanisms are often unable to represent long documents due to their significant memory requirements. In this thesis we explore a number of document representation strategies (Chapter 6) and discuss the pros and cons of sentence-level representations for our comprehensive impact scoring task (Chapter 10).

1.4 Thesis Outline & Contributions

This thesis follows a traditional structure and is comprised of 10 chapters that build upon each other progressively:

- **Chapter 1** (this chapter) lays out our motivation for this work and the objectives and research questions that we tackle.
- **Chapter 2** provides relevant background information about machine learning, the fundamentals of natural language processing and more recent neural language modelling approaches.
- **Chapter 3** addresses **RQ1.1** and **RQ1.2** by providing background and a literature review for the field of scientometrics: the study of metrics and measures of scientific success. It defines types of scientific impact and provides a survey of scientometric research related to this thesis.
- **Chapter 4** focuses on **RQ1.3**, exploring the relationships between traditional scientific impact metrics, altmetrics and non-academic scientific impact metrics. We combine a number of heterogenous data sources to build a open-access citation network and facilitate a statistical comparison of metrics.
- **Chapter 5** addresses **RQ2** and **CH1**, exploring possible sources of evidence for comprehensive impact before focussing specifically on the scientific journalism and the creation of the HarriGT Corpus: a collection of news articles linked to scientific papers that they discuss. We also contribute a novel annotation platform that uses machine learning models and third party citation networks to help users link news articles to the scientific papers that they discuss. Finally, we carry out a statistical analysis and establish the presence of a relationship between mentions of scientific work in the news and achieving more comprehensive impact.
- **Chapter 6** addresses **RQ3** and **CH4**, reviewing and comparing document encoding approaches for representing news articles and scientific papers. We discuss the advantages and disadvantages of each strategy and evaluate similarity between pairs of news and science articles in an information retrieval task.

- **Chapter 7** explores **RQ4** and **CH2**. We carry out structural analyses of news articles and scientific papers to understand how authors use the structures of these documents to present important information prominently. We carry out sentence-wise comparisons of paired documents and reveal a relationship between the location of information in a scientific paper and how prominently it is presented in a news article.
- **Chapter 8** addresses **RQ5**, **CH1** and **CH2**. It defines a novel cross-document co-reference resolution setting for resolving references between news articles and scientific papers. We present an annotation tool for our cross-domain co-reference resolution task that uses a pre-trained language model to assist human annotators. We train a state-of-the-art co-reference resolution model on our dataset and carry out an error analysis on the resulting model.
- **Chapter 9** addresses **RQ6**, **CH1** and **CH3**. Following a pilot study using existing state-of-the-art comprehensive impact detection techniques, we define a new regression task for impact and build an annotation tool and we annotate sentences from news articles and scientific papers collected in previous chapters. Due to the time-consuming nature of the annotation task, we explore low resource learning strategies including few-shot learning and multi-task learning and train a series of models that achieve good regression performance and vastly outperform the legacy model.
- **Chapter 10** outlines our findings and achievements and discusses possible directions that future work could take.

2

A Background on Machine Learning (ML) & Natural Language Processing (NLP)

> *"I think perhaps the most important problem is that we are trying to understand the fundamental workings of the universe via a language devised for telling one another when the best fruit is"*
>
> *Terry Pratchett*

Natural Language Processing (NLP) is concerned with the application of computational methodologies and algorithms to natural languages (e.g. languages that humans would normally use to communicate such as English or Mandarin) in order to automatically interpret and extract meaning from it. Whilst many of the fundamentals of NLP come from philosophers and linguists who predate the modern information era, the modern discipline is thought to have started in the 1940s and 1950s with the publication of seminal works by Turing, Chomsky and Dostert (Nye, 2016; Hutchins, 2004, 1997). NLP is a subfield of both Computer Science and Linguistics.

NLP has experienced a boom in recent years thanks to the wide availability of modern Graphical Processing Units (GPUs) that offer large scale parallel mathematical calculations that have facilitated new state-of-the-art machine learning models that can perform challenging language tasks quickly and accurately.

Russell et al. (2010) observed that most applications of NLP fall into two broad groups of use cases:

- The facilitation of communication between computers and humans (e.g. chatbots) or two sets of humans (e.g. automatic translation tools);
- Knowledge acquisition and information extraction in which we aim to summarise, search, categorise or score text *en masse* in order to identify patterns and signals;

In this thesis we are primarily concerned with methods from the latter of these two groups which we apply to the task of helping us to understand scientific writing and the comprehensive impact therein.

The field of NLP is vast and many of the works and techniques described in this thesis are built upon many hundreds of person-years of combined effort. Unfortunately, there is not enough space to sufficiently and adequately describe all supporting works from their original theoretical bases which themselves consists of multiple books, papers and theses. Therefore, we include descriptions of machine learning approaches that we either make use of directly in our work or that are complementary to those that we do use in some way (e.g. that provide historical context for machine learning techniques that we deploy). Our

work assumes an undergraduate level understanding of probability and statistics, geometry, arithmetic, calculus and algorithms. For a more comprehensive background on NLP, we recommend Manning and Schiitze (1998) and Bird et al. (2009).

2.0.0.1 Chapter Summary

In this chapter we discuss a number of machine learning and NLP techniques that we use in subsequent chapters:

- We begin by describing some fundamentals of machine learning which play a significant part in today's state-of-the-art NLP landscape and by illustrating a typical NLP machine learning workflow;
- We explore approaches for representing text in a machine-readable way, from simplistic encoding of one word at a time to complex models of entire sentences and documents that take into account linguistic context and structure;
- We outline a number of relevant machine learning models used to make predictions about text inputs;
- We describe a number of common NLP tasks and related work which we build upon later in this thesis. Furthermore, we also describe appropriate corresponding evaluation metrics;

2.1 Natural Language Processing & Machine Learning

Machine Learning is a sub-field of Computer Science and Artificial Intelligence concerned with building systems that can 'learn' from data rather than requiring explicit programming like traditional software. Many natural language applications have extremely varied and diverse inputs which may be difficult or even impossible to encapsulate and process using traditional imperative or even declarative programming approaches. Machine learning is a tool in the NLP practitioner's arsenal that allows them to build applications which can process and respond adequately for unseen natural language inputs.

Machine learning approaches fall into different categories depending upon the type and availability of data used to train the system. In this thesis we focus on supervised and unsupervised learning.

2.1.1 Supervised Learning

Supervised learning models learn to predict outputs (usually represented as y) that correspond to a given input (usually represented x). Supervised algorithms are trained using a training data set x_{train} which consists of n tuples of example inputs and desired outputs

$(x_0, y_0), (x_1, y_1)...(x_n, y_n)$. A separate test set x_{test} in the same format is used to evaluate the performance of the supervised learning model.

Tasks that generate a real-valued output, for example prediction of price, temperature or house prices, are collectively called regression problems. Tasks that generate a discrete or categorical output, for example whether an email is spam or not spam or whether a photo is of a cat or a dog, are known as classification problems.

2.1.2 Unsupervised Learning

Unsupervised learning algorithms take a set of data containing only inputs and find structure in the data. A typical example of unsupervised learning is clustering in which alike data points are grouped together into the same cluster and dissimilar data points are grouped into different clusters.

Dimensionality reduction algorithms are used to reduce a large feature space into a smaller more manageable data representation, often to reduce overfitting (see section 2.2). These algorithms are unsupervised and we often use them to simplify machine-readable representations of text (as discussed in section 2.5.4).

2.1.3 Machine Learning Workflow

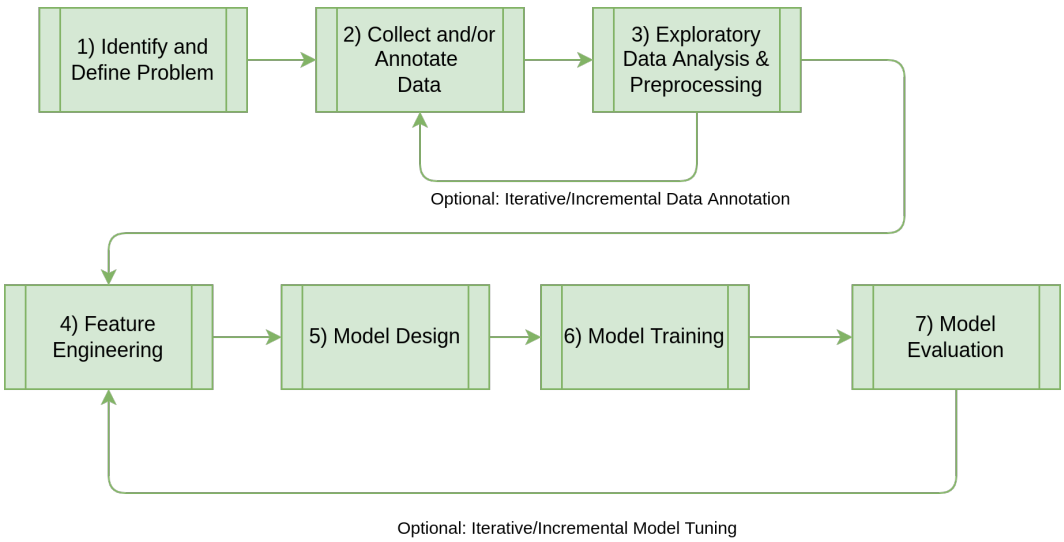


Figure 2.1: Typical Machine Learning Workflow

This thesis focuses heavily upon the application of supervised and unsupervised machine learning algorithms to an heterogenous collection of text data in order to develop new ways to measure and understand comprehensive scientific impact. In Figure 2.1 we outline a typical generalised workflow for both supervised and unsupervised models which we describe in depth below. Our workflow is similar to the general approach for working

with text in a machine learning context proposed by Nguyen et al. (2019). The approach outlined here is a generic one and operationalisation of such a workflow is highly dependent on a number of empirical aspects such as data quality, algorithm and hyperparameter selection as well as practical limitations around computation speed. Readers experienced with machine learning workflows may wish to skip ahead to section 2.5.

1. We must understand the problem and what approach is most appropriate for solving it. This will depend upon the availability of data - i.e. labelled or unlabelled and the type of output to be identified - i.e. continuous or categorical. In some cases a machine learning approach may be inappropriate. If there is very limited data available or the problem is so subjective that humans cannot reach consensus on the desired output then it may not be possible to train a machine learning model to solve the problem. Likewise, if the problem and its constraints and inputs can be expressed easily it may be more appropriate to write a program or algorithm to solve it than to use machine learning.

2. During this step we aim to collect input/output tuples for supervised learning problems or simply a set of inputs for unsupervised learning problems. For well explored NLP problems such as Natural Language Inference, Semantic Textual Entailment or Co-reference Resolution (which we explore in more detail in section 2.8), large datasets may already exist or it may be possible to combine existing datasets to generate larger training sets. For unknown or novel problems we may have to carry out manual data annotation. This usually involves defining a data collection task and asking humans to manually assign desired outputs to example inputs. In order to ensure that human annotators provide consistent annotations, a set of annotation guidelines may be compiled and distributed to help them to understand the task and provide guidance for how to handle confusing or contentious data examples.

3. Exploratory data analysis and preprocessing involves carrying out an initial examination of the dataset that has been collected. This typically involves carrying out statistical checks to characterise the data. For categorical data, this may include checking whether data points are evenly distributed across each of the classes. For unlabelled data or input/output pairs with continous outputs, this might include checking whether data is normally distributed and identifying outliers. In NLP we normally carry out a number of preprocessing operations which we describe in section 2.3. If human annotation was part of the data collection process in Step 2 then we may also calculate inter-annotator-agreement metrics such as Cohen's Kappa (Cohen, 1960) or Fleiss Kappa (Fleiss, 1971) to understand how consistent the collected dataset is. If data is collected inconsistently, it can prevent machine learning models from converging and limit their statistical performance. In these cases, there may be optional iteration over steps 2 and 3 in order to revise annotation guidelines and re-annotate

inconsistently labelled data.

4. Once data has been explored and preprocessed, we implement strategies for representing the input data in machine-readable ways that are compatible with machine learning algorithms. We more comprehensively describe approaches for representing text in section 2.5.

5. We decide which machine learning algorithm we intend to use to how best to configure the algorithm to solve the task. Many machine learning algorithms can be configured via the use of hyperparameters which dictate the speed, quality and resource usage of the learning process. Unlike model parameters which are learned automatically as part of the training process, hyperparameters must be set manually before training and can have a significant impact on the final performance of the model. In section 2.6 we describe in detail the machine learning models that we use in this thesis and the hyperparameters that they depend upon.

6. The machine learning model is taught to infer outputs for unseen input data by incrementally observing (x, y) pairs from the training data (or x samples in unsupervised settings) and updating its parameters.

7. We use a test set that was held back during training to evaluate how well the model training worked (see section 2.4 for detail on test sets and model evaluation). Each input from the training set is passed through the model and an associated output y' is generated. For supervised models where a correct y answer is known, accuracy, F1 score or regression metrics such as MAE may be calculated. Evaluation of unsupervised models is slightly more difficult due to the lack of true labels against which to compare. However, metrics such as perplexity or coherence may provide an indication of model fit. We discuss evaluation metrics in more detail in Section 2.4 below.

8. After model evaluation, investigators may choose to revisit steps 4-7 and attempt to improve performance by changing the features, machine learning algorithm and hyperparameters. This process may be repeated a number of times.

2.2 Underfitting and Overfitting

Machine Learning models are approximations (or rather *models*) of the functions that they are trained to emulate. Simple models have few parameters which may limit their ability to accurately capture the relationships between inputs and outputs. This phenomenon is called Underfitting. The opposite phenomenon, overfitting, occurs when a model is *too* complex for the problem it is learning to emulate, learning spurious relationships between

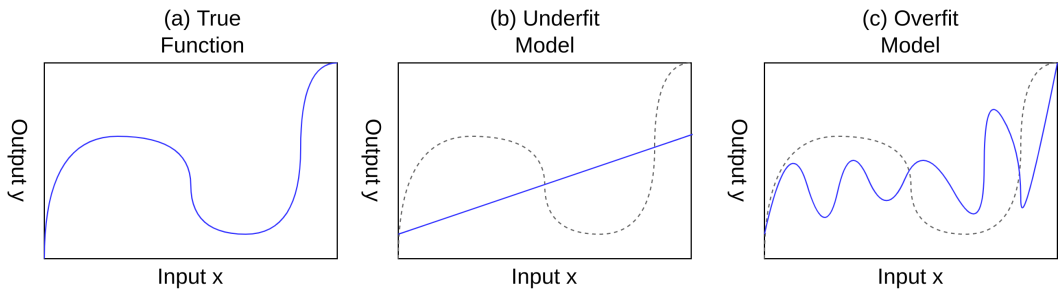


Figure 2.2: A simplified illustration of a regression function (a) and models that have underfit (b) and overfit (c) to the true function.

inputs and outputs that don't generalise well to new unseen data. Figure 2.2 illustrates these phenomena.

Underfitting and Overfitting are particularly important to consider when building NLP ML systems because of the complexity of natural language, the fidelity and number of features required to model it coupled with the complexity of modern language models (which we discuss in depth in section 2.6).

2.3 Preprocessing Text

When working with NLP machine learning models, the size and complexity of the vocabulary used to train the model can have a significant effect on the performance of the model; if we model with a simple vocabulary and remove too many words we are likely to underfit and if we model with a large vocabulary and capture every word we may overfit. This can be managed with appropriate use of the preprocessing steps described in this section and by picking appropriate text representations which we discuss in more detail in section 2.5.

Here we describe a number of commonly used text pre-processing strategies that can be applied before the input is vectorized and passed to a model, which help with normalisation of documents and reduction of vocabulary size and complexity.

2.3.1 Tokenization

Tokenization is the process of splitting a document into individual tokens or words. Some tokenizer implementations will separate contractions (e.g. they're $\rightarrow$ [they, 're]). Documents may be tokenized into single words but some models may tokenize into bigrams (two word chunks), trigrams (three-word chunks) or n-grams (where $n > 3$). Tokenization is usually carried out as an initial step before other preprocessing steps described below are carried out.

2.3.2 Case Normalization

In many NLP applications, the casing of letters in words is not important. Changing all words to lower or upper case can reduce complexity and noise by effectively de-duplicating word features. For example, rather than requiring separate representations for 'Scientist' and 'scientist', both forms can be represented as 'scientist'.

2.3.3 Stopword Removal

Stop words are commonly occuring words such as *the*, *as* or *but* which appear with a high frequency in natural language but which are unlikely to provide any value to downstream machine learning models (Manning and Schiitze, 1998). Removal of stop words is a very common preprocessing step which reduces noise and computational complexity by reducing the number of features that must be parameterised by models. Many open source NLP software libraries such as NLTK (Bird et al., 2009) and spaCy (Honnibal and Montani, 2017) provide lists of stop words in different languages which, whilst convenient for analysts, may inadvertently impede model performance issues due to surprising omissions or inclusions and should thus be used with care (Nothman et al., 2018). Removal of Stopwords may be unnecessary or even detrimental in some use cases and the effects should be tested as part of the model tuning workflow.

2.3.4 Stemming & Lemmatization

Stemming and Lemmatization reduce noise and complexity by normalising different surface forms of a word into a single form through simplification. Stemming strips affixes from words such that 'government' $\rightarrow$ 'govern' and 'executive' $\rightarrow$ 'execut' (Bird et al., 2009). Lemmatization takes stemming a step further by requiring that the resulting form is a known word in a dictionary. Like Stopword Removal, Stemming and Lemmatization can sometimes make performance worse and training with or without using them can be tested during the model tuning phase of the workflow.

2.3.5 Pattern-Based Replacement

In order to minimise vocabulary size and control overfitting, it is often useful to replace specific words and strings that occur rarely with a general placeholder. For example, adding specific website addresses or specific numerical quantities to the model vocabulary which could increase its size significantly with little benefit since the model is unlikely to encounter these same websites or exact numbers again in other documents. However, by using simple pattern matching techniques like Regular Expressions to replace all URLs with <WEB> and all numbers with <NUM>, we allow the model to generalise its behaviour for all websites and numbers.

2.4 Evaluation of Machine Learning Models

Evaluating machine learning models involves estimating how well a model that has been trained on a dataset may generalise to new unseen data. This process is facilitated by the separation of labelled data into disjoint training and test sets (as discussed in section 2.1.3 which allows us to approximate the performance of the model on a new unseen set of data. Evaluating our model on the same data that was used to train it would give an inflated indication of performance since the model will have overfit on the training examples, internalising the answers within its parameters; analogous to using worked algebra problems from the text book that a student used to learn mathematics as the exam questions used to evaluate their final end of year grade.

In this section, we describe common practices for selection of evaluation sets and then describe a number of commonly used machine learning metrics for supervised classification and regression use cases. In Section 2.8 we also describe some less commonly used task-specific metrics alongside their respective tasks.

2.4.1 Selection of Test, Validation and Train Datasets

It is fairly common to hold back 30% of the data for testing and train on 70%. If the model needs to be tuned (i.e. steps 4-7 in Figure 2.1 are repeated multiple times), the workflow may lead to overfitting of the model to the test set as the hyperparameters are tuned to increase performance on the same dataset. In this scenario, a common strategy is to split the dataset into three parts: a training set of 60%, a validation set of 20% and a final test set of 20%. The model can be fine-tuned on the validation set and once tuning is complete, a final estimate of the model's true performance can be made using the previously unseen test set.

In practice, datasets are usually separated through pseudo-random sampling without replacement. If data is unbalanced, stratified sampling may be used to ensure that the test, train (and validation) sets have comparable label distributions.

2.4.2 Cross-Fold Validation

Cross-fold validation can be used in the event that only small amount of labelled data is available for training and testing. In cross-fold validation, the data is randomly split into k folds (common values are 3, 5 and 10). $k-1$ folds are used to train the machine learning model and the remaining fold is used to evaluate the model. This process is repeated k times yielding k models. The results from each evaluation may be averaged together to approximate performance for the full data set.

2.4.3 Metrics for Classification

2.4.3.1 Accuracy

Accuracy is the most simplistic of classification metrics, giving a high level view of how well a model is performing. It is simply the proportion of correctly labelled examples versus the total number of examples:

$$\text{Accuracy} = \frac{y'_{correct}}{y'_{total}} \tag{2.1}$$

Accuracy is not appropriate for evaluating imbalanced problems. For example, in a binary classification task with 70 samples where $y = false$ and 30 examples where $y = true$, the model would achieve a 70% accuracy if it labels every example true. For multi-class classification tasks (for example, assigning a document 1 of 12 possible topic labels), top-level accuracy does not provide information about the model's performance with respect to each class.

2.4.3.2 Confusion Matrix & True/False Positive/Negative

When we evaluate classification models it is useful to assign each item i in the model output data into four sets:

- **True Positive (TP)** describes items that are annotated AND predicted as belonging to the class of interest (i.e. $y_i = true$ and $y'_i = true$)
- **True Negative (TN)** describes items that are annotated AND predicted as not belonging to the class of interest (i.e. $y_i = false$ and $y'_i = false$)
- **False Positive (FP)** describes items that are predicted to belong to the class of interest but annotated as not belonging to that class (i.e. $y_i = false$ and $y'_i = true$)
- **False Negative (FN)** describes items that are predicted not to belong to the class of interest but annotated as belonging to that class (i.e. $y_i = true$ and $y'_i = false$)

These sets can be expressed visually as a confusion matrix which plots actual labels against predicted labels as illustrated in Figure 2.3 which shows a confusion matrix for a binary task with 160 items in the evaluation set (sum of cells), 70 true positives, 20 false negatives, 40 false positives and 30 true negatives.

2.4.3.3 Precision, Recall, F1-Score

The metrics precision and recall help us to understand how well the model performs at filtering out false positives or false negatives.

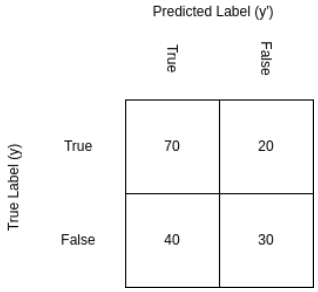


Figure 2.3: Example of a confusion matrix

Precision is defined as:

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{2.2}$$

Recall is defined as:

$$\text{Precision} = \frac{|TP|}{|TP| + |FN|} \tag{2.3}$$

For the classification task illustrated in Figure 2.3:

$$\text{Precision} = 70/(70 + 20) = 0.778 \tag{2.4}$$

$$\text{Recall} = 70/(70 + 40) = 0.636 \tag{2.5}$$

For a given model, these two metrics often oppose. We can increase precision at the expense of recall or *vice versa* dependent on our goal. In a medical test for a treatable but deadly illness setting we may want to optimise for recall at the expense of precision; more people who are not sick test positive and get invited for more invasive testing but we reduce the chance that someone who is sick is not detected. In a mail versus spam detection setting we may want to optimise for precision at the expense of recall; we'd like to keep the inbox as clean as possible by limiting false positives ($y = spam$, $y' = mail$).

It is common to summarise model performance at Recall and Precision using F-Measure or F1-Score which is the harmonic mean of the per-class precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.6}$$

2.4.3.4 Applying Classification Metrics to Multi-Class Problems

The metrics illustrated above can also be trivially applied to multi-class problems as well as binary classification problems. We simply calculate TP, TN, FP, FN with respect to each class in our dataset. Take Fisher (1936)'s Iris dataset in which we use measurements of

different flowers and use them to predict which species each sample belongs to (setosa, versicolor, virginica)

- **True Positive (TP)** describes items that are annotated AND predicted as belonging to the class of interest (e.g. class of interest is **setosa**, $y_i = setosa$ and $y_i' = setosa$)
- **True Negative (TN)** describes items that are annotated AND predicted as not belonging to the class of interest (e.g. class of interest is **setosa**, $y_i = versicolor$ and $y_i' = versicolor$)
- **False Positive (FP)** describes items that are predicted to belong to the class of interest but annotated as not belonging to that class (e.g. class of interest is **setosa**, $y_i = virginica$ and $y_i' = setosa$)
- **False Negative (FN)** describes items that are predicted not to belong to the class of interest but annotated as belonging to that class (e.g. class of interest is **setosa**,$y_i = setosa$ and $y_i' = versicolor$)

Likewise we can plot a multi-class confusion matrix by adding additional rows and columns for each class as illustrated in Figure 2.4:

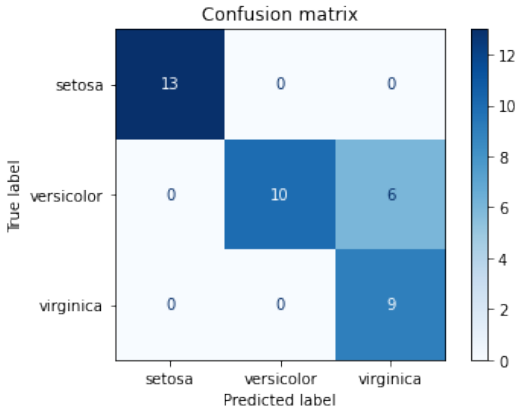


Figure 2.4: Example of a confusion matrix for a multi-class classification task

We can calculate Precision, Recall and F1 for each individual class. For **versicolor** (based on data in Figure 2.4):

$$Precision = 10/(10+6) = 0.625 \tag{2.7}$$

$$Recall = 10/(10+0) = 1.0 \tag{2.8}$$

$$F1 = \frac{2 \times 0.625 \times 1.0}{0.625 + 1.0} = 0.769 \tag{2.9}$$

2.4.3.5 Micro versus Macro Average

For Multi-class problems we can summarise overall performance by re-combining per-class Precision, Recall and F1 through averaging.

Macro Averaging involves calculating Precision, Recall and F1 for each class and taking the mean.

Macro average Precision for the iris dataset in Figure 2.4:

$$Precision_{setosa} = 1.0 \tag{2.10}$$

$$Precision_{versicolor} = 0.625 \tag{2.11}$$

$$Precision_{versicolor} = 1.0 \tag{2.12}$$

$$Precision_{macro} = \frac{1.0 + 0.625 + 1.0}{3} = 0.875 \tag{2.13}$$

Micro Averaging involves calculating TP, TN, FP, FN for each class and summing together the sizes of each respective set before calculating the metric of interest.

Micro average Precision for the iris dataset in Figure 2.4:

$$\text{TP}_{setosa} = 13, \text{FP}_{setosa} = 0 \tag{2.14}$$

$$\text{TP}_{versicolor} = 10, \text{FP}_{versicolor} = 6 \tag{2.15}$$

$$\text{TP}_{virginica} = 9, \text{FP}_{virginica} = 0 \tag{2.16}$$

$$\text{TP}_{total} = 32, \text{FP}_{total} = 6 \tag{2.17}$$

$$Precision_{micro} = \frac{32}{32 + 6} = 0.842 \tag{2.18}$$

For an unbalanced multi-class dataset, **macro averaged** metrics, which assign equal importance to all classes regardless of the number of samples assigned to them, will highlight poor performance in smaller classes but may give a more pessimistic view of overall performance than **micro averaged** metrics.

2.4.4 Metrics for Regression

In chapters 4 and 9 we train and evaluate regression algorithms. Here we briefly describe the metrics with which we measure their performance.

2.4.4.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a simple error metric for regression models defined as $MAE = \frac{\sum_i(|y_i - y_i'|)}{||y||}$. MAE is reported in the same unit as the target variable (e.g. if we are predicting house prices in £ then MAE would tell us "by how many £ the regressor is incorrect on average"). This can be both a benefit and a limitation of MAE; it makes it easier for subject-matter-experts to interpret (e.g. mortgage lenders) but may make it more difficult to interpret for those who are not familiar with the task that the regressor is trained on.

2.4.4.2 R^2 Coefficient

The R^2 coefficient is a task agnostic can be used to measure the performance of a regressor. It gives a broad sense of how well the regressor is able to estimate the target variable without requiring the reader to have background knowledge of what a good or bad absolute error value might be (as opposed to MAE above). The R^2 coefficient is defined as:

$$R^2 = 1 - \frac{SS_E}{SS_T} \tag{2.19}$$

where $SS_E = \sum_i (y_i - y_i')^2$, the total sum of squared residuals and $SS_T = \sum_i (y_i - \hat{y})^2$, the total sum of squared distances between each y value and the mean of y values ($\hat{y}$). R^2 is bounded $[1, -\infty]$ where a value of 1 indicates that the function is perfectly predictive of the dependent variable (sum of squared residuals, SS_E, is zero). In the special case that function always predicted the mean value of y, $\hat{y}$ then $\frac{SS_E}{SS_T}$ would be 1 and $R^2 = 0$. Thus, a value $0 \leq R^2 \leq 1$ indicates a function that is 'better than the mean'. Residual errors can be arbitrarily large if R^2 can be arbitrarily negative.

2.5 Representing Text

In order to train machine learning models to make predictions about text, we must convert it into a machine readable format that is compatible with ML models. Models are mathematical in nature and require numerical inputs. Thus we must convert text inputs such that they can be represented via numerical vectors. In this section we outline commonly used text representation approaches.

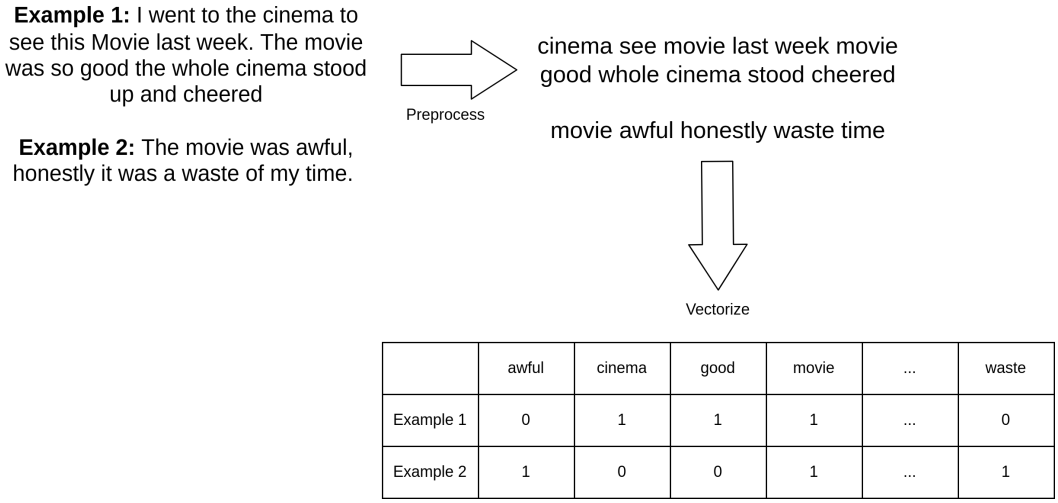


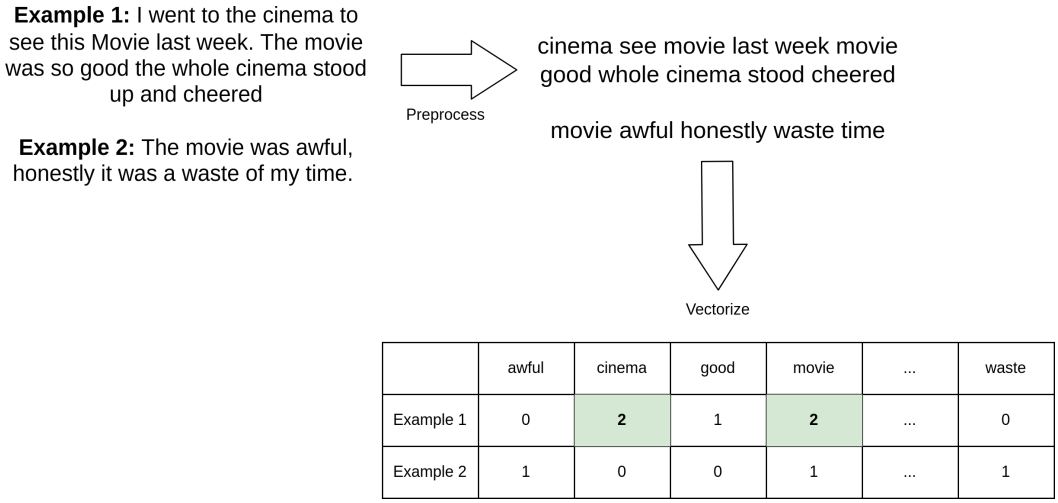
Figure 2.5: Two movie reviews are preprocessed and vectorised into bag of words (BoW) vectors

2.5.1 Bag-of-Words Representations

Bag of words (BoW) representations are one of the simplest representations of text that are still widely used in baseline experiments at the time of writing. Firstly, each input example i in the list of inputs D are tokenised and normalised as outlined in section 2.3. Next, a model vocabulary V is established by finding the union of all words occurring in all input examples excluding those removed during preprocessing. Following this a vector for each input example $\mathbf{x}_i$ is initialized with the same length as V such that $dim(\mathbf{x}_i) = dim(V)$. Then, we iterate through each word in the vocabulary and set the vector value with the corresponding offset offset to 1 if the word appears in the document or 0 if it doesn't.

$$\forall_{i \in D} \forall_{j \in V} \; \mathbf{x}_i^j = \begin{cases} 1, & \text{if } V_j \in i \\ 0, & \text{otherwise} \end{cases} \tag{2.20}$$

This process is illustrated in Figure 2.5. Bag-of-words vocabularies can be constructed with single tokens but may also include two or three token expressions in order to preserve more meaningful signals from multi-word phrases and natural word co-occurrence. For example, encoding the compound noun 'capital city' as a single item in a BoW vector may confer more meaning than independently capturing 'capital' and 'city'. Encoding multiple words together as a single entry in the model vocabulary is known as *ngram encoding*, single words are unigrams, two word phrases bigrams, three word phrases trigrams and so forth. Different ngram lengths can be captured together in the same model such that 'capital' 'city' and 'capital city' could all be encoded by the same model to maximise flexibility.



	awful	cinema	good	movie	...	waste
Example 1	0	**2**	1	**2**	...	0
Example 2	1	0	0	1	...	1

Figure 2.6: Two movie reviews are preprocessed and vectorized into count vectors. The words cinema and movie appears twice in Example 1 - the corresponding vector cells are highlighted green for clarity.

2.5.2 Count Vectors

Count vectors are an extension of BoW vectors where the number of occurrences of each ngram in each document is captured such that word repetition can be modelled. This is helpful for modelling long documents where word repetition may be more prominent. For example, a model that classifies magazine articles by topic may benefit from knowing that 'computer' occurs multiple times in an article about software development and once in an article about archaeology ("they dug up the artifact and used computer aided design to imagine what it looked like 1000 years ago"). Each cell in vector $\mathbf{x}_i^j \in [0, \infty]$ corresponds to the number of times the ngram at offset j appears in example i. This approach is illustrated in Figure 2.6

$$\forall_{i \in D} \forall_{j \in v} \; \mathbf{x}_i^j = wordfreq(i, j) \tag{2.21}$$

2.5.3 TF-IDF

As discussed in section 2.3, removing stopwords can help reduce noise in a model by ensuring that ngrams that commonly appear in almost all documents and which are unlikely to provide useful information to NLP models are removed. However, within specific use cases or language domains, there may be use case specific stop words. For example, a model trained to classify movie reviews is unlikely to gain useful information from knowing that the word 'movie' appeared in the review. Whilst one could build a use case specific list of stop words by manually examining the input documents, Term Frequency - Inverse Document Frequency (TF-IDF) provides a more scalable automated alternative approach (Jones, 1972).

The aim of TF-IDF is to score each ngram's importance based on how often it appears in a given document (as per count vectors above) but moderate the score by the number of documents it appears in. Term Frequency (TF), how often the word j appears in the document i is defined as:

$$TF_i^j = wordfreq(i, j) \tag{2.22}$$

Inverse Document Frequency (IDF) assigns a score to each word j based on how many documents within training the corpus the ngram appears and is defined as:

$$df(j) = \sum_{D}^{i} \begin{cases} 1, & \text{if } v_j \in i \\ 0, & \text{otherwise} \end{cases} \tag{2.23}$$

$$IDF^j = log \frac{|D|}{df(j)} \tag{2.24}$$

These terms are multiplied together to give the TF-IDF score for each ngram j with respect to each document i:

$$\forall_{i \in D} \forall_{j \in v} \mathbf{x}_i^j = TF(i, j) \times IDF(j) \tag{2.25}$$

2.5.4 A Note on the Limitations of BoW/ngram Representations

The simplicity of the BoW embedding variants explored above serves as a double-edged sword. On the one hand they are a convenient, efficient and intuitive way to encode text for NLP use cases. On the other hand, they are limited in terms of the fidelity with which they can represent natural language. BoW models do not encode any information about word meaning or relationships between words which means that downstream machine learning algorithms must learn to parameterise each word in the model vocabulary from scratch; although stemming and lemmatization can help by normalising different surface forms and participles of the same words, BoW vectors cannot provide any signal about semantic relationships between synonyms, antonyms, hypernyms etc. Furthermore, the necessity of fixing the model vocabulary at training time prevents BoW-based models from inferring outputs based on previously unseen words[1]. BoW embeddings are also unable to represent word order, which means that "I hope Biden wins and Trump loses" and "I hope Trump wins and Biden loses" would have identical BoW representations using any of the techniques discussed above, which is clearly problematic.

Without careful management, BoW representations can also become very large and sparse. Every n-gram that passes the preprocessing stage (see section 2.3) becomes an additional dimension in the model's input vector representation. In NLP task settings that

[1] in practice, previously unseen words are simply not represented in the document vector and an output is inferred based on words from the input document that were present in the training set.

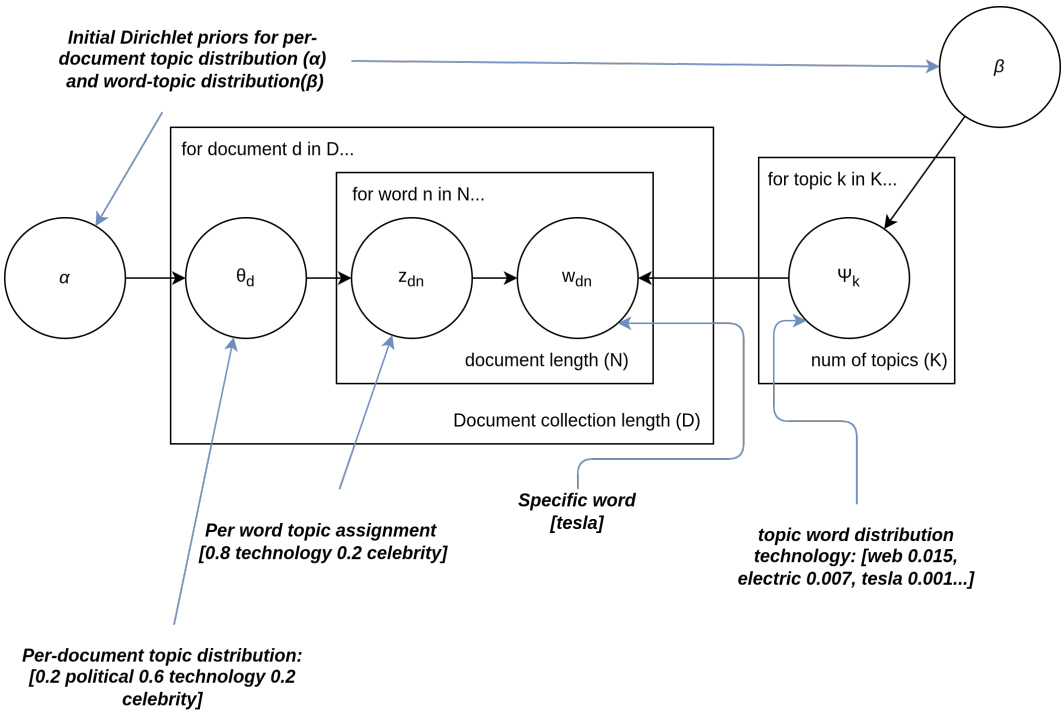


Figure 2.7: Annotated plate notation view of LDA model

involve large and complex vocabularies, such as processing scientific papers, model vocabularies can reach tens of thousands of ngrams (Liakata et al., 2012a) and many of these features may only appear in a relatively small proportion of the corpus leading to overfitting. In some cases, dimensionality reduction techniques such as Principal Component Analysis (Tipping and Bishop, 1999) or topic models such as LDA (discussed below in section 2.5.5) can be used as a preprocessing step applied to the bag-of-words vector before it is passed to the model in order to improve model performance and reduce complexity.

In the following sections we explore more powerful text representation approaches and models which can alleviate some of these issues at the expense of greater complexity and compute cost.

2.5.5 Topic Models

Topic models are a family of algorithms that analyse texts with the aim of identifying and extracting linguistic themes from them. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a commonly used probabilistic topic modelling approach which is simple, efficient and easy to use. LDA is based on the assumption that k probability distributions over words, or 'topics' exists for a collection of documents D and that each document d in the collection can be generated from a mixture of these topics (Blei, 2012).

The LDA process is visualised as a plate diagram in Figure 2.7. α and β are distributions used to randomly initialise θ, the probability distribution of topics over documents

and ψ, the probability distribution of words over topics respectively (Blei et al., 2003). Distributions are then updated iteratively through a Gibbs Sampling process.

The number of topics k is a hyperparameter that must be set before the algorithm is fit. The value of k is usually obtained through empirical testing. If LDA is trained with too few topics, mixture representations of documents may be too general making identification of alike documents very difficult (e.g. papers about the discovery of a new type of insect and the discovery of a new type of digestive enzyme in cows may both be associated with high likelihood with a topic approximating general biology rather than distinct entomology and biochemistry). However, with too many topics, the model begins to learn highly specific topics that are only representative of a small proportion of documents and unlikely to generalise well to new documents outside the corpus, essentially overfitting to the training data. In the ideal case, there should be enough topics in the model to represent the subjects discussed in the corpus but not enough that the model is able to memorize documents by assigning specialised topics to them.

LDA models can be used as an exploration tool and combined with visualisation tools like LDAVis (Sievert and Shirley, 2014) to provide interpretable views of patterns within a collection of documents. LDA topic distributions can also be used as document representations for downstream machine learning tasks, providing a form of dimensionality reduction and helping with word meaning disambiguation (i.e. the word cell as in cell phone may be strongly associated with a technology topic and cell as in stem cell associated with a biology topic).

2.5.6 Context-Independent Neural Word Embeddings

Context-independent neural word embeddings represent words or phrases as real-valued N-dimensional vectors in a latent space (where N is a hyperparameter set dependent on the size of the vocabulary and training corpus). Semantically similar words are assigned vectors with a strong cosine similarity and *vice versa*. Compared with BoW approaches where the dimensionality of each representation is equal to the size of the vocabulary $|V|$, word embedding representations are much denser (i.e. $|N| << |V|$). Word embeddings confer the advantage of being able to provide the semantic relationship information that they encode to downstream models. Furthermore, word embeddings can be conditioned on large unlabelled text collections, incorporating and encapsulating semantic similarity between words that are not in a task specific training data and thus providing more robust and generalisable representations of unseen words to downstream models. These models are predicated on the distributional theory of semantics (Harris, 1954), specifically, that words that are used and occur in similar contexts tend to have similar meanings. Seminal work by Collobert and Weston (2008) and Collobert et al. (2011) demonstrated that a dense neural network (which we describe in more detail in section 2.6) could be used to train universal word embeddings that can be used for a number of downstream NLP tasks. However,

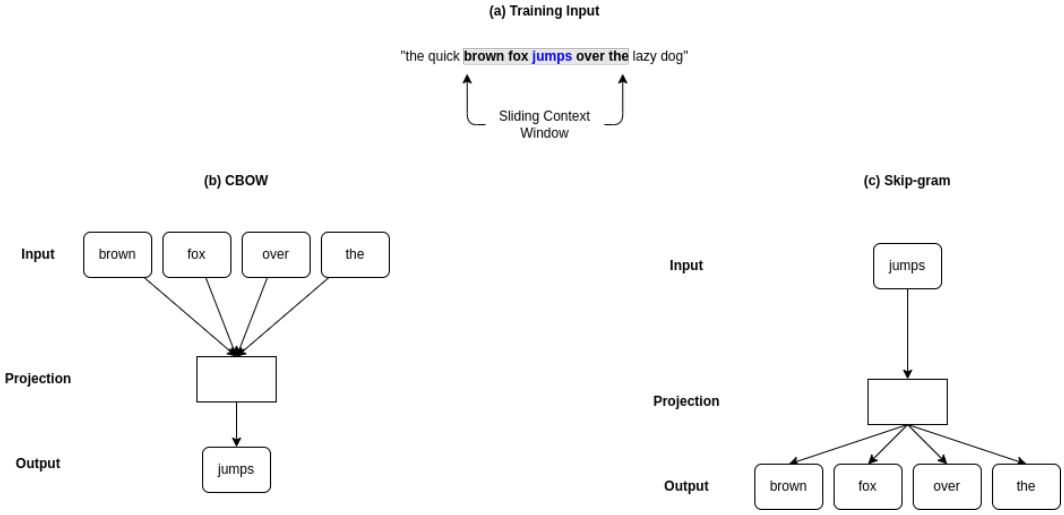


Figure 2.8: Model Architecture Diagrams for CBOW (b) and Skip-gram(c)

subsequent work by Mikolov et al. (2013b). arguably popularised word embeddings within the NLP community.

2.5.6.1 word2vec & Predictive Word Embeddings

word2vec (Mikolov et al., 2013b) takes the principals explored by Collobert and Weston (2008) and expands on them, contributing novel log-linear efficiency models that vastly reduces the computational cost of calculating word vectors. The approach works by training a simplified feed forward neural network (as discussed in section 2.6) to predict a word based on its surrounding context (CBOW, Figure 2.8.B) or to propose neighbouring context words based on an input (Skip-gram, Figure 2.8.C). In both cases word contexts are sampled from unlabelled input documents via a sliding context window (Figure 2.8.A), the size of which is a hyperparameter of the model that may be tuned.

After model training the word vectors, stored in the projection layer, may be extracted and used as inputs for downstream tasks. The two approaches capture slightly different information in their projections with CBOW working slightly better on semantic tasks and Skip-gram working much better on semantic tasks. word2vec projections pre-trained on $\approx$ 6B words from the Google News corpus are available on the author's website[2] which has led to word2vec being downloaded and used to boost state-of-the-art performance for a large number of NLP tasks. Skip-gram vectors exhibit "additive compositionality" whereby vectors may be used in simple alebraic operations. For example, $\text{vector}(\text{king}) - \text{vector}(\text{boy}) + \text{vector}(\text{girl}) \approx \text{vector}(\text{queen})$. This property of the vectors is caused by the skip-gram training objective (Mikolov et al., 2013a).

[2] `https://code.google.com/archive/p/word2vec/`

2.5.6.2 GloVE & Co-occurence Matrix Embedings

Global Vectors or GloVe (Pennington et al., 2014a) builds on top of word2vec and integrates both global corpus statistics and local context information. The model constructs a co-occurence probability matrix P_{ij} (in which a cell represents the probability that word j occurs in the same context as word i) for the training corpus. Then, vectors in a randomly initialised projection space are optimised via a regression objective such that words that are most likely to co-occur are assigned similar vectors and words that are unlikely to co-occur are assigned dissimilar vectors. Like word2vec, GloVe embeddings pre-trained on a large corpus 42B token corpus (the Common Crawl dataset[3]) are made available on the author's website [4] and have achieved popularity within the NLP community.

2.5.6.3 Representing Word Sequences with Word Vectors

As opposed to BoW document representations which assign a dimension in the model's input vector to each word in its vocabulary, word2vec and GloVe-based word projections assign a dense, multidimensional vector to each word. Therefore, we must address the best way to combine these vectors into a single input for a model. A common approach is to feed each word into neural models that accept sequential inputs (which we will explore in more detail in section 2.6) and encourage the model to learn a strategy for pooling individual word vectors. Recent work (Shen et al., 2018) shows that for many NLP tasks, simply taking the element-wise average of each word vector in an input sequence can provide a strong baseline. Both sequential models and average-vector representations of text inputs struggle to represent long inputs (e.g. full documents) and we discuss alternative strategies for longer inputs later in this thesis (see sections 6.3.2.5 and 6.6).

2.5.6.4 Out-of-Vocabulary Words, Sub-word Embeddings & Byte-Pair Encoding

A major advantage of neural word embedding approaches over BoW-based approaches is that they can encode semantic relatedness between words that do and do not occur in downstream training corpora allowing their corresponding models to better generalise to unseen texts. However, these approaches still learn to represent a fixed vocabulary which means that models may still need to omit missing words from their input (many models that use static neural inputs are trained with an unknown or 'UNK' token as a general placeholder for unknown words whilst some simply omit the missing word all together). Sennrich et al. (2016) proposed a solution to this problem inspired by the Byte Pair Encoding (BPE) compression algorithm (Gage, 1994) in which commonly occurring sequences of characters are used to assemble a dictionary in which longer sequences are represented. They propose training building vector projections of full words but also representing sub-word informa-

[3] `http://commoncrawl.org/`
[4] `https://nlp.stanford.edu/projects/glove/`

tion. They break words down into character ngrams (e.g. 'eating' → 'eat', 'ing.') with special character '.' representing the end of a word. New, out of vocabulary words can then be represented as vectors by summing together their constituent character ngrams. Subsequent work by Bojanowski et al. (2017) popularised subword representation by extending Skip-gram (Mikolov et al., 2013b) with a similar BPE scheme and making pre-trained models widely available[5].

2.5.7 Contextual Representations

Words can have different meanings in different contexts. One of the primary limitations of the word vector approaches outlined above is that they are context independent or static. Each word in the model vocabulary is associated with a single word vector and words with multiple definitions or meanings are encoded with a single definition (in practice the corresponding vector is likely to be closely aligned to the word's most common meaning). Attempts have been made to retroactively add word sense disambiguation to static word vectors (Trask et al., 2015; Orkphol and Yang, 2019). However, contextualised representations, which build upon context independent word vectors, have been shown to provide state-of-the-art performance for many NLP tasks by taking into account the context of the input tokens (Devlin et al., 2019; Cattan et al., 2020; Held et al., 2021; Wright and Augenstein, 2021; Bommasani et al., 2021).

Contextual representation models take into account long sequences of tokens (e.g. full sentences or short paragraphs) and generate vectors for each word that vary depending on the meaning of the sentence.

2.5.7.1 ELMo

Embeddings from Language Models (ELMo) (Peters et al., 2018a) introduced a novel biLSTM model (see section 2.6) that generates vectors which encode context-specific word sense and part-of-speech information by processing full sentences.The authors' primary contribution was the notion of pre-training general purpose language encoders rather than using context independent static word embeddings as input and learning task-specific encoders from scratch. ELMo is pre-trained on a next-word-prediction language modelling task similar on a very general purpose word dataset (Chelba et al., 2014) and shown to produce vectors that are useful for improving the performance of a large number of downstream tasks. Peters et al. (2018a) use character-level representations of words via a CNN layer (see section 2.6.4) allowing the model to construct vectors for previously unseen terms.

[5] `https://fasttext.cc/`

2.5.7.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), released shortly after ElMo extend contextual embeddings in two key ways. Firstly, they use a novel Transformer-Encoder architecture (see section 2.6.8.2) which allows their model to encode long sequences of text whilst taking into account word position and grammatical dependencies and without the downsides of RNNs (specifically: difficulty learning long-range dependencies between words and avoiding the computation bottleneck caused by dependencies between sequential states, discussed further in section 2.6.8). Secondly, BERT introduces two novel, unsupervised, pre-training tasks beyond the simple language modelling (next-word-prediction) used by Peters et al. (2018a):

- A masked word prediction task in which a small percentage of tokens in the input sequence are masked or 'hidden' during the encoding step and predicted by the model in a softmax layer over the model's vocabulary. The flexibility afforded by attention-based encoding (see 2.6.8) allows any token in the sequence to be masked, whereas RNN-based models are limited by their sequential nature (the missing token must always be last in the sequence or else it is observed 'early' by the model and can be guessed trivially).
- A next *sentence* prediction task in which two sentences are passed to BERT at the same time and the model must guess which sentence came first. BERT allows two sentences to be passed as part of the same sequence and includes a special control character (denoted $[SEP]$) to identify sentence boundaries.

BERT also incorporates sub-word information through a BPE-like encoding scheme (see section 2.5.6.4), allowing it to process previously unseen words.

After pre-training, BERT can be trained to carry out new, previously unseen NLP tasks by attaching task-specific output layers to the model and carrying out back-propagation. This process is known as 'fine-tuning', a type of transfer learning which we discuss in section 2.7.

2.5.7.3 BERT Variants

Since its publication, BERT has become so popular that it has even inspired its own sub-discipline: BERTology (Rogers et al., 2020) (although this name is likely tongue-in-cheek). The model has provided a starting point for fine-tuning a number of state-of-the-art NLP models and inspired a number of BERT model variants.

A key limitation of BERT is that it is a very large model with between 110 and 340 million learnable model parameters depending on model 'flavour'. Variants like ALBERT (Lan et al., 2020) and DistilBERT (Sanh et al., 2020) aim to reduce the size of the model

without reducing statistical performance and by doing so yield similar sized models with similar or better performance than the original formulation.

RoBERTa, or Robustly Trained BERT (Liu et al., 2019a) is a modification of BERT that use the same model architecture but amends the training regime to incorporate more data, improve the masked token prediction task by re-assigning the masked tokens within a given sentence during training and by completely removing the next sentence prediction task. RoBERTa has been shown to consistently outperform BERT across a number of NLP tasks. At time of writing, RoBERTa is regularly used as a base model for emerging NLP works (Cattan et al., 2020; Schick et al., 2020; Gao et al., 2021; August et al., 2020; Wright and Augenstein, 2021) and we use RoBERTa as the starting point for a number of the experiments in this thesis.

2.5.8 Metrics for Text Similarity

Gauging the similarity between feature vectors is a common application within NLP, facilitating information retrieval (how similar is this document to the input query?) and visualisation tasks (how closely together should these points be plotted?). In chapters 6, 8 and 7 we apply similarity metrics to document representations in order to identify related documents. We briefly describe metrics used for these comparisons below.

2.5.8.1 Cosine Similarity

Cosine similarity is a geometric measure of similarity often applied to compare document representations constructed via neural text embedding models (see sections 2.5.6.3, 2.5.7). The cosine similarity between two multidimensional vectors is defined as the cosine of the angle between ($cos(\theta)$) them which can be derived via Euclidean distance and Pythagoras' cosine rule (for vectors $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \cdot \mathbf{B} = ||\mathbf{A}||||\mathbf{B}||cos(\theta)$). Cosine similarity can be defined as:

$$cos\text{-}similarity(\mathbf{A}, \mathbf{B}) = cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||} \tag{2.26}$$

Cosine similarity has the same function bounds as $cos(\theta) \in [-1, 1]$. However, cosine similarity is most often used for comparing vectors in positive space with bounds $\in [0, 1]$. A higher number $\rightarrow 1$ indicates that two vectors are close to orthogonal (i.e. they are dissimilar in positive space). A low number $\rightarrow 0$ indicates that two vectors lie in a similar direction (i.e. they are similar).

Cosine Distance is the complement of cosine similarity in positive space, defined as:

$$cosine\text{-}distance(\mathbf{A}, \mathbf{B}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||} \tag{2.27}$$

Cosine Distance is often used in contexts where authors wish to emphasise the similarity

rather than the difference between two vectors (e.g similar vectors will have a cosine distance $\rightarrow 1$ and dissimilar vectors have a cosine distance $\rightarrow 0$). Like Cosine Similarity, Cosine distance is also bounded $[0, 1]$.

2.5.8.2 Jensen-Shannon Divergence

Jensen-Shannon Divergence (JSD), sometimes referred to as Information Radius (Manning and Schiitze, 1998) is a measure for understanding the similarity between two probability distributions P and Q derived from Kullback-Leibler divergence (Csiszar, 1975). JSD is defined as

$$JSD(P, Q) = \frac{D(P||M) + D(Q||M)}{2} \tag{2.28}$$

Where $M = \frac{P+Q}{2}$, the pointwise mean of the two distributions P & Q and D is the Kullback-Leibler Divergence between distributions P and Q defined as follows:

$$D(P||Q) = \sum_{i}^{||P||} P_i log \frac{P_i}{Q_i} \tag{2.29}$$

JSD is an alternative to cosine similarity that is often used for comparing vectors representing probability distributions (e.g. distributions from topic models (section 2.5.5) (Aletras and Stevenson, 2014; Wartena, 2013; Lee, 1999) rather than geometric latent vectors such as those produced by neural embedding approaches (sections 2.5.6 and 2.5.7).

JSD is bounded $[0, 1]$ where values $\rightarrow 1$ show that two distributions are very different and values $\rightarrow 0$ show that distributions are similar.

A full understanding of JSD and its definition based on information theory first principles is not required in order to understand our work and is provided in this thesis. However, readers who are interested may find the Essential Information Theory chapter in Manning and Schiitze (1998), which describes these concepts, JSD (referred to as IRAD in the book) and KL Divergence, an excellent starting point.

2.6 Models

In this section we describe a number of machine learning models that we apply to specific tasks in later chapters, or that provide useful background context for understanding related work. The section covers modelling of classification and regression tasks and incrementally builds from simpler models to more complex models.

2.6.1 Linear Models

Linear models are a class of simple statistical models for regression and classification tasks that are easy to understand and fast and cheap to run. They are often used as baselines, but

their simplicity limits them to simpler NLP tasks.

2.6.1.1 Linear Regression

Linear Regression is a simple regression model that attempts to model a scalar relationship between a dependent value y and one or more independent input variables x. The model fits to the data by creating a "line of best fit" that directly intersects or comes as close as possible to as many data points as possible by finding the global minimum for the sum of residuals (or 'error' - distance between the line and the data points as illustrated in Figure 2.9). The most common approach is Ordinary Least Squares (OLS) whereby we attempt to minimise the sum of squared residuals which has the effect of preventing positive and negative residuals from cancelling each other out.

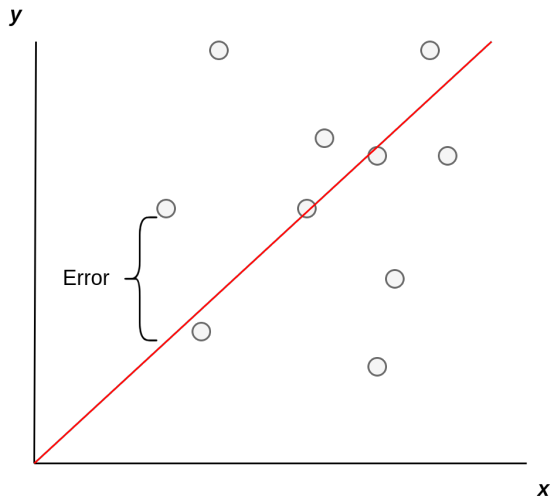


Figure 2.9: Illustration of a dataset plotted in x and y and a line of best fit generated by the model

2.6.1.2 Logistic Regression

Despite its name, Logistic Regression is a classification modelling approach which attempts to separate a set of data belonging to one of two classes (designated $y = 0$ and $y = 1$) using the logistic function (or sigmoid function) as illustrated in Figure 2.10. The location (μ) and scale (s) parameters of the sigmoid function are learned using maximum likelihood estimation over each (x, y) pair in the training set D_{train} and the corresponding y' prediction by the model. The sigmoid function produces an S-shaped output $0 \leq y' \in \mathbb{R} \leq 1$ which can be interpreted as the model's confidence that a given input x belongs to the $y = 1$ class:

$$y' = \sigma(\sum_{i=1}^{n} \mathbf{w}_i \mathbf{x}_i + c) \tag{2.30}$$

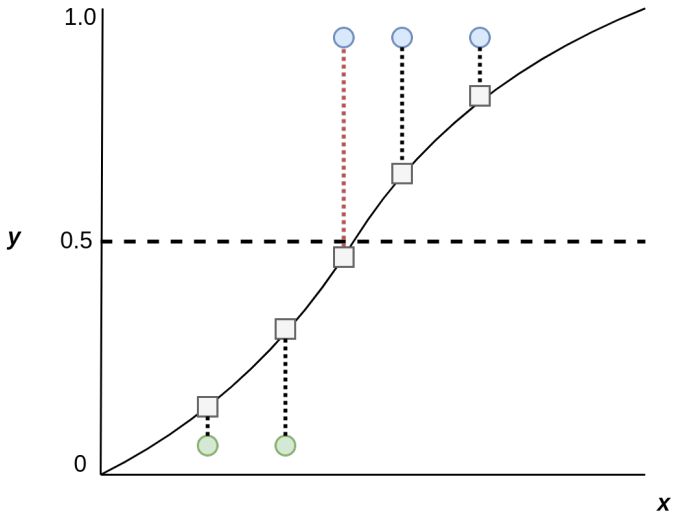


Figure 2.10: Illustration of a logistic regression model with a logistic function that separates classes $y = 1$ (blue) from $y = 0$ (green) along the decision threshold $y = 0.5$. One of the examples is misclassified (red dotted line) as the predicted data point falls below the decision threshold.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.31}$$

Model loss is calculated in equation 2.32 and partial derivatives are used to iteratively update the model parameters $\mathbf{w}$ and c. The loss is minimized through gradient descent.

$$L_{log}(y, y') = -(ylog(y') + (1 - y)log(1 - y')) \tag{2.32}$$

2.6.2 Support Vector Machine

Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) are a class of linear supervised learning models that can be used for classification or regression use cases. For a given input set $(\mathbf{x}_1, y_1)...(\mathbf{x}_n, y_n)$ where x is a p dimensional vector and y corresponds to one of two classes ($y = 1$ and $y = -1$), an SVM attempts to find a $(p - 1)$ dimensional hyperplane $\mathbf{w}$ which separates the two classes such that $\mathbf{w}^T\mathbf{x} - b = 0$. In order to identify this hyperplane we also define two other hyperplanes where $\mathbf{w}^T\mathbf{x} - b = 1$ and $\mathbf{w}^T\mathbf{x} - b = -1$ respectively that separate the two classes of data (anything above $\mathbf{w}^T\mathbf{x} - b = 1$ or below $\mathbf{w}^T\mathbf{x} - b = -1$ belongs to class 1 or -1 respectively). We train the model by maximising the margin between these two vectors. The data points that lie on or near these class hyperplanes are the support vectors and are used to define the decision function.

The principle of Support Vector Regressors (Drucker et al., 1996) (SVRs) is similar to that of a Support Vector Classifier (SVC) but with a slightly different learning objective.

[6] https://commons.wikimedia.org/wiki/File:SVM_margin.png

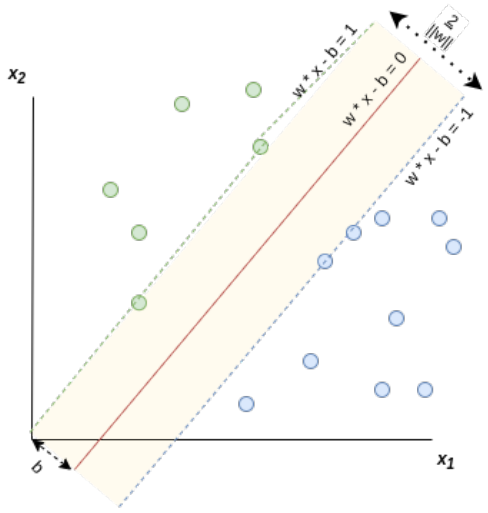


Figure 2.11: Illustration of a support vector classifier, diagram heavily inspired by Larhmam[6]under Creative Commons Share-Alike License

We seek to find a line that best fits the data set $D = (\mathbf{x}_0, y_0), ..., (\mathbf{x}_n, y_n)$ where $y \in \mathbb{R}$, again using support vectors from $\mathbf{x}$ to find the optimal solution. We attempt to find a hyperplane $f(x) \rightarrow y'$ that has at most ϵ deviation from the true targets y (Smola and Schölkopf, 2004).

We define a hyperplane:

$$f(x) = \mathbf{w}^T\mathbf{x}$$

$$\text{subject to} \begin{cases} y_i - \mathbf{w}^T\mathbf{x} - b \leq \epsilon \\ \mathbf{w}^T\mathbf{x} + b - y_i \leq \epsilon \end{cases} \tag{2.33}$$

We illustrate this in Figure 2.12. We optimise the solution by minimising $\frac{1}{2}||w||^2$.

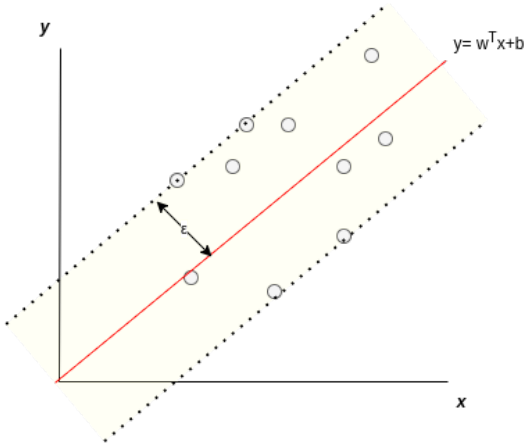


Figure 2.12: Illustration of a support vector regressor.

Support vector machines provide powerful predictive capabilities and usually outperform linear methods for text classification and generalise well for high dimensional sparse feature spaces associated with text classification, particularly when working with BoW features (Joachims, 1998).

2.6.3 Neural Models

Over the last decade, artificial neural networks (ANNs), specifically, deep neural models consisting of multiple, complex hidden layers, have seen vast adoption from the NLP community and the broader ML community due to their ability to model very complex problems and outperform many of the models listed above (in many but not all circumstances). The availability of cheap modern Graphics Processing Units (GPUs) and other types of tensor co-processors have made it feasible to train and run many of the highly complex and computationally expensive neural network architectures that have been theorised and tested on a small scale over the last few decades.

Many of the experiments and models described in this thesis are built using neural models, so it is important that we fully explore the theory behind them as a prerequisite to understanding these experiments. We devote a significant proportion of this chapter to the fundamentals of neural networks, back propagation and some of the more recent model architectures.

This section is heavily influenced by the excellent Primer on Neural Networks by Goldberg (2015).

2.6.3.1 Neurons

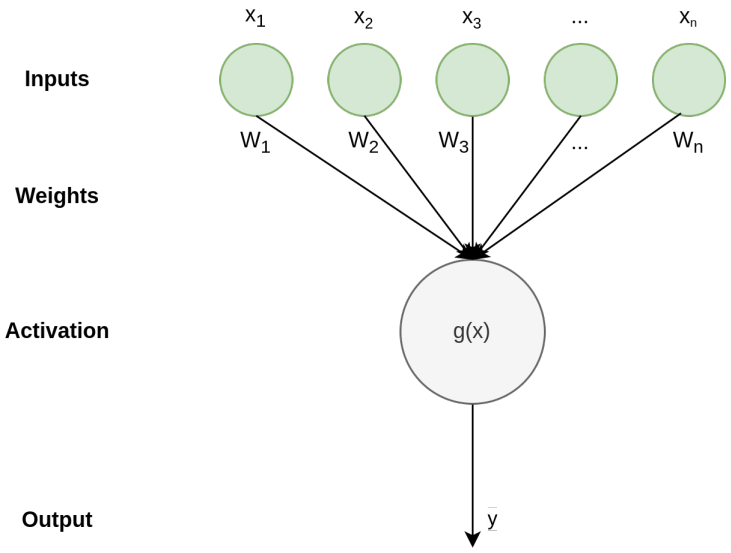


Figure 2.13: An example of a neuron from an ANN

The neuron is a single computation unit within an ANN with scalar inputs and outputs (Goldberg, 2015) as illustrated in Figure 2.13. Each neuron input x_n has a corresponding weight, W_n. During forward propagation, the neuron multiplies each input with its corresponding weight and sums them together. A non-linear activation function $g(x)$ (section 2.6.3.3) is then applied and the resulting value is passed as the output of the neuron. Neurons also typically learn to apply a bias term b which functions as a y intercept, allowing the neuron to shift its outputs by a constant if required (for example a minimum may be learned because all training outputs are greater than some threshold.). This operation is described formally in equation 2.34 (where $\cdot$ denotes the dot product between vectors **w** and **x**)

$$\begin{aligned} y' &= g(\sum_{i=1}^{n}(\mathbf{W}_i\mathbf{x}_i) + b) \\ &= g(\mathbf{w}\cdot\mathbf{x} + b) \end{aligned} \tag{2.34}$$

$$\mathbf{x} \in \mathbb{R}^n, \mathbf{W} \in \mathbb{R}^n, b \in \mathbb{R}^n$$

2.6.3.2 Feed Forward Neural Network

A feed forward neural network (FFN) is an ANN in which consists of an n-dimensional input layer that accepts an n-dimensional input, an output layer with a shape that depends on the type of problem you're trying to solve (section 2.6.3.4) and 1-to-many hidden layers of neurons (section 2.6.3.1 below) that moderate and propagate information flowing through the network. The number of hidden neurons and number of hidden layers in the network are hyperparameters that may be optimised through systematic exploration and testing as part of the ML workflow (section 2.1.3). The inputs of each layer in an FFN are directly connected to the outputs from the previous layer as illustrated in Figure 2.14. These layers are known as fully connected or 'dense' layers.

ANNs are trained by randomly initializing each of the model weights and parameters and then iteratively following the process of back propagation (section 2.6.3.6)

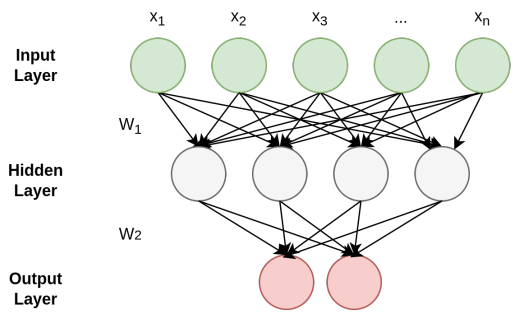


Figure 2.14: An example of a Feed Forward Artificial Neural Network with One Hidden Layer

2.6.3.3 Activation Functions

There are a number of common activation functions used within ANNs including:

- Sigmoid (as discussed in section 2.6.1.2 above)
- Hyperbolic Tangent (tanh) which transforms input values x into the range [-1,1] (Goldberg, 2015) and
- ReLU function (Glorot et al., 2011) which clips values for $x < 0$, returning values in range [0,∞].

There are no general theories about which activation functions work well in which scenarios. Activation function selection is therefore, typically part of hyperparameter tuning during the course of the ML workflow (section 2.1.3) based on empirical model evaluation.

2.6.3.4 Problem Types & ANN Output Shapes

For binary classification ($y \in [0, 1]$) and regression tasks it is common for the output layer to consist of a single neuron with a sigmoid activation function. For regression problems, the output value may be interpreted directly as a regression output or by scaled in post-processing. For binary classification the output corresponds to a model's confidence that the input belongs to class $y = 1$ and can be combined with a threshold to produced a class label.

For multi-label classification problems, it is common to transform the final model output using the *softmax* function. *Softmax* generates a probability distribution over k possible outcomes (Goldberg, 2015). For $\mathbf{x} = \{x_1, ..., x_k\}$ where k corresponds to the number of classes in a classification task:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{2.35}$$

The output is a vector of real valued, non-negative numbers. Each element in vector x is normalised in turn such that the resulting numbers all sum to 1 and can be treated as a probability distribution over possible class labels. *Softmax* is typically applied in conjunction with cross entropy loss training objective (section 2.6.3.6 below).

2.6.3.5 Forward Propagation

Information in an ANN is propagated through each layer of the network from input to output. This process is known as forward propagation and is used to infer model outputs for new unseen data and is also the first step in the process of training an ANN. Here we illustrate this process using the the FFN illustrated in Figure 2.14 with an input layer, a single hidden layer consisting of h neurons and an output layer with dimensionality corresponding to the number of classes depicted in y (in this case 2 classes).

Neurons in each FFN layer have the same input dimensionality and activation functions which means that parameters for each layer can be efficiently grouped together into single tensors and forward propagation can be efficiently calculated using a single linear algebra equation per layer. A sample input, encoded as n-dimensional vector $\mathbf{x} = x_1, x_2, x_3...x_n$, is passed into the network. The $n \times h$ dimensional matrix W_1 controls the strength of the connections between each input dimension n and each corresponding neuron in the ANN's hidden layer. Likewise, a vector $\mathbf{b}_1$ also contains the learned bias terms for each neuron. To begin with $\mathbf{W}_1$ and $\mathbf{b}_1$ are randomly initialized but these parameters are tuned and learned via back propagation (see below).

In the first step, intermediate output from hidden layer 1 ($\mathbf{h}_{out}$) is generated by applying the neurons in the hidden layer to input $\mathbf{x}$ as follows. We generalise and extend the equation 2.34 such that it can be applied in parallel to all neurons:

$$\mathbf{h}_{\text{out}} = g(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) \tag{2.36}$$

We next propagate the intermediate output $\mathbf{h}_{out}$ through the output layer. The output of this model is a 2-dimensional vector representing the probability distribution $p(y|x)$[7]. We multiply the intermediate output $\mathbf{h}_{out}$ with weights W_2 and apply the *softmax* transformation function to the final output of the model:

$$y' = softmax(\mathbf{h}_{\text{out}}\mathbf{W}_2) \tag{2.37}$$

We can describe the whole FNN by combining equations 2.36 and 2.37 together:

$$y' = softmax(g(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2) \tag{2.38}$$

2.6.3.6 Training & Backward Propagation

Training an ANN involves propagating information forward through the network (section 2.6.3.5), comparing the output of the network to the true output and then propagating the amount of error caused at each stag in the process back to each layer in the network and incrementally updating the associated weights and biases (section 2.6.3.1). This process relies heavily on differential calculus, allowing us to find the gradient of the output with respect to each model parameter.

Here we illustrate a single back propagation step for input x using the FFN described in section 2.6.3.2. First we forward propagate the input x through the network and generate y' as described in equation 2.38. Next we calculate the loss or error generated by the network. We use a loss function (section 2.6.3.7) to describe the difference between y, the true label according to the dataset and y', the predicted output from the model. We

[7] astute readers will have noticed that a more efficient way to represent this problem would be to use a single sigmoid output and thresholding as described above

use Categorical Cross-Entropy loss (sometimes referred to as the negative log likelihood (Goldberg, 2015)) which provides a probabilistic interpretation of the softmax output from 2.6.3.5 which is a probability distribution over two possible classes.

$$L_{CE}(y', y) = \sum_{y} y_i log(y'_i) \tag{2.39}$$

We next use Stochastic Gradient Descent (SGD) to make a small update to the weights in the model based on the error $E = L_{CE}(y', y)$ and its gradients with respect to each weight parameter $\frac{\delta E}{\delta \mathbf{W}}$ in the model (LeCun et al., 1998b):

$$\mathbf{W} := \mathbf{W} - \frac{\delta E}{\delta \mathbf{W}} \quad \forall \mathbf{W} \in [\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}] \tag{2.40}$$

Parameter $\eta \in [0, 1]$ is a learning rate hyperparameter limiting the extent of each model weight update per iteration that is typically decreased with every iteration t of back propagation e.g:

$$\eta_t = \eta_0(1 + \eta_0 \lambda t)^{-1} \tag{2.41}$$

Where η_0 is the initial value identified with a small training sample and λ is an additional hyperparameter (Goldberg, 2015; Bottou, 2012).

We repeat this process for all values x, y in the training set, gradually updating all weight parameters in the network and reducing the error E with each iteration. Often, training is done on mini-batches containing multiple training pairs $(x_0, y_0, x_1, y_1..x_n, y_n)$ [8] and the training error is averaged per mini-batch. This approach is computationally efficient and by randomly shuffling the training set into batches we introduce noise into batches which may allow the iterative gradient descent process to escape local minima (LeCun et al., 1998b).

A number of extensions and improvements to SGD have been proposed in recent years. Current SOTA training regimes often use RMSProp (Hinton et al., 2012) and ADAM (Kingma and Ba, 2017) which provide automated strategies for adjusting the learning rate per mini-batch which often simplifies the need for manual learning rate tuning (Goldberg, 2015).

A single complete iteration over all samples in the training dataset is called an epoch. The number of epochs that the model is trained for and when to stop training is another hyperparameter that can be set through empirical experimentation. A common strategy is to use the validation set (see section 2.4.1) to evaluate the network's current performance at the target task after every epoch to see if model is adequately fit. Stagnation or reduction of a model's performance after training is a sign that the model is beginning to overfit on a task and that training should stop. Often, practitioners will keep snapshots of

[8] n is often a small power of 2 e.g. $2, 4, 8, 16, 32$ because this facilitates efficient memory allocation on the GPU

model weights after each epoch and save those that correpond to the highest model performance on the validation set as the 'best' model.

2.6.3.7 Loss Functions

Common ***Classification*** loss functions for training neural networks are:

- **Categorical Cross Entropy** as outlined above in equation 2.39, provides a measurement of how different two probability distributions are from each other which is particularly useful when working with softmax outputs.
- **Binary Cross Entropy** or log loss measures the similarity between two binary distributions and is described in equation 2.32 for training logistic regression models. Logistic regression models can be considered a special case of FNN with a single hidden layer and a sigmoid activation function.
- **Hinge Loss** is the same loss function as used in SVM classifiers (section 2.6.2). Hinge loss assumes that $y \in \{-1, 1\}$ and is defined as $L_{hinge} = max(0, 1 - y \cdot y')$. Loss is 0 when y and y' share the same sign and $|y'| \geq 1$ (Goldberg, 2015).

A common ***Regression*** loss function for training neural networks is Mean Squared Error (MSE) where:

$$L_{mse} = \frac{1}{|\mathbf{y}|} \sum_{i=1}^{|y|} (y_i - y'_i)^2 \tag{2.42}$$

2.6.4 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a special class of ANN commonly applied in image processing. Whilst we do not directly make use of CNNs in our work, convolutional layers are used within other models that we discuss in this thesis. CNNs were first introduced by Fukushima (1980) and LeCun et al. (1998a) and more recently popularised by AlexNet (Krizhevsky et al., 2012) in the era of deep learning.

CNNs are inspired by the way that visual cortexes work in the brains of vertebrate animals (Fukushima, 1980). As opposed to FFNs, neurons in a convolutional layer are connected only to small region of the input (known as the receptive fields), allowing partial processing of the input without the need to learn to compensate for unrelated noise. All regions of the input are covered by overlapping receptive fields to maximise the chance that one or more fields capture relevant information. CNNs typically connect many convolutional layers together to produce a hierarchy of intermediate representations with increasing levels of abstraction and complexity (for example an image classifier's first layers may contain neurons that map raw pixels into features like eyes, nose, seceding layers may see eyes and nose in the same receptive field and learn to identify faces).

With each layer in the network, CNNs progressively combine information from neighbouring receptive fields into increasingly dense representations with some data necessarily lost at each stage. This makes CNNs particularly well suited to identifying relationships across neighbouring regions of inputs more likely to struggle with long distance relationships between non-neighbouring regions.

In NLP, CNNs can be used to process text by assigning receptive fields that correspond to words or characters with subsequent layers learning to combine these features into phrases and sentences. They have been shown to be useful for generating word representations for previously unseen words based on character-level inputs (Peters et al., 2018a) and have been used in popular NLP library spaCy (Honnibal and Montani, 2017) to complete a variety of downstream tasks. CNNs often struggle to learn complex and long-distance dependencies between words, necessitating the use of recurrent architectures (section 2.6.5) and Transformer-based models (section 2.6.8.2). However, CNNs can be used to augment these latter models to improve overall model performance (Hassan and Mahmood, 2018; Peters et al., 2018a).

2.6.5 Recurrent Neural Models

Traditional formulations of ANNs such as the FFN described above support only limited representations of sequential data. Textual data, which is inherently sequential, can be represented in an FFN as a variation of BoW (section 2.5.1) or as a mean vector (section 2.5.6.3) but word order cannot be preserved in either case. Recurrent neural networks (RNNs) are a *general* class of neural architecture that solve this problem by allowing the network to process a sequence of inputs in order. To facilitate this a single RNN layer observes a sequence of input vectors $\mathbf{x}_0, ..., \mathbf{x}_n$ of length n one after another, and combines them with an internal variable s which is used by the model to store state information relating the sequential inputs together as illustrated in Figure 2.15.

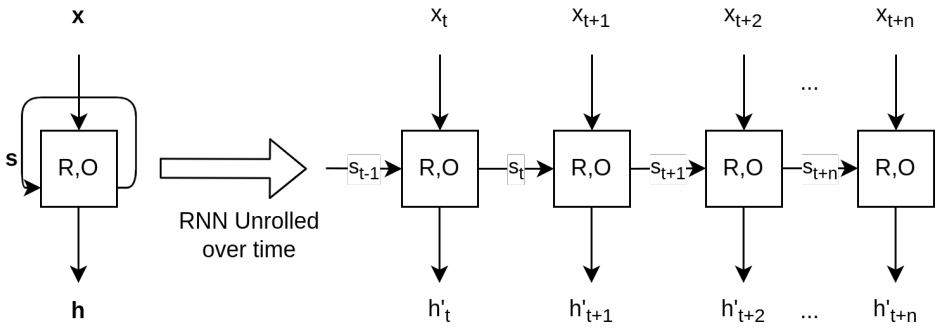


Figure 2.15: An example of a recurrent neural network (RNN) with unrolled representation on the right. At timestep t the input x_t and previous state s_{t-1} are processed together by the cell and produce intermediate hidden output h_t.

For each element **x**, the RNN layer produces a corresponding intermediate output

$\mathbf{h}'_0, ...\mathbf{h}'_n$ which may then be passed through to subsequent layers (e.g. dense layers as described in section 2.6.3.2) for use in downstream tasks. Commonly the final state from the layer $\mathbf{h}'_n$, which is the result of processing all preceding inputs $\mathbf{x}_0, ...\mathbf{x}_{n-1}$ is used as the input to a fully connected layer that carries out the classification or regression task of interest e.g. $y' = softmax(\mathbf{W}\mathbf{h}' + \mathbf{b})$

The model requires an initial state $\mathbf{s}_0$ to combine with the initial input $\mathbf{x}_0$ in the RNN layer. In practice, is usually initialized randomly or with a constant value. Subsequently, for each time-step t in the input sequence, $\mathbf{x}_t$ is passed into the RNN along with state $\mathbf{s}_{t-1}$ and outputs $\mathbf{s}_t$ and $\mathbf{h}'_t$ are generated.

The general RNN function is defined recursively in equation 2.43 and makes use of two internal functions: R which maps the previous state and the current input step onto a new state and O which maps the current input state (as generated by R) onto an intermediate output.

$$\begin{aligned} RNN(\mathbf{s}_0, \mathbf{x}_{1:n}) &= \mathbf{s}_{1:n}, \mathbf{h}'_{1:n} \\ \mathbf{s}_t &= R(\mathbf{s}_{t-1}, \mathbf{x}_t) \\ \mathbf{h}'_t &= O(\mathbf{s}_t) \end{aligned} \tag{2.43}$$

Below we discuss two RNN implementations: the Simple RNN (section 2.6.5.1) which provides a simple baseline implementation of the R and O functions above and LSTM (section 2.6.5.2), a widely used, powerful, sequential model for encoding long distance dependencies e.g. when the first and final words in a sentence are related.

2.6.5.1 Simple RNN (SRNN)

The seminal Simple RNN (SRNN) formulation proposed by Elman (1990) and modified for text processing by Mikolov (2012) takes the form:

$$\begin{aligned} \mathbf{s}_t = R_{SRNN}(\mathbf{s}_{t-1}, \mathbf{x}_t) &= g(\mathbf{x}_t\mathbf{W}^x + \mathbf{s}_{i-1}\mathbf{W}^s + \mathbf{b}) \\ \mathbf{h}'_t = O_{SRNN}(\mathbf{s}_t) &= \mathbf{s}_t \\ \mathbf{s}_t, \mathbf{h}'_t \in \mathbb{R}^{d_n}, \mathbf{x}_t \in \mathbb{R}^{d_x}, \mathbf{W}^x \in \mathbb{R}^{d_x \times d_n}, \mathbf{W}^x &\in \mathbb{R}^{d_s \times d_s} \end{aligned} \tag{2.44}$$

Weights $\mathbf{W}^x$, $\mathbf{W}^s$ and bias $\mathbf{b}$ are randomly initialised and learned during training. s_0 is randomly initialized. tanh and ReLu are commonly used as the activation function $g(x)$ (see section 2.6.3.3).

Despite this model's simplicity, it provides good results for sequence tagging and language modelling tasks (Goldberg, 2015; Mikolov, 2012). However, one of the main drawbacks of SRNNs is that their performance quickly degrades for long input sequences for two reasons (Bengio et al., 1994):

1. RNNs are more capable of encoding information about recent inputs in their state and may *'forget'* about inputs much earlier in the sequence meaning that long term depen-

dencies between words may be lost - e.g. for a next word prediction task "sky" would be easy to predict in the context "not a cloud in the" but much harder in the context "Commercial jets have wings and are normally powered by jet engines which... The wings help it to stay in the";

2. The vanishing gradient problem, which is caused by back-propagation through a large number of time steps since the gradient gets smaller after every round of differentiation with respect to the previous layer. This would also be a problem for exceptionally deep FFNs (which is what an RNN looks like when you 'unroll' it);

2.6.5.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a type of recurrent neural network which aim to solve the problem of long term dependencies outlined in section 2.6.5.1 above. We do not directly use LSTMs in this thesis. However, we briefly discuss them here as an example of a more powerful RNN architecture which has significantly influenced how NLP practitioners model sequences of text.

LSTM cells are an extended and slightly more complex version of the neuron used in FFNs (section 2.13). Like neurons in SRNNs, LSTM cells recurrently pass state information C_t to themselves as they operate on a sequential input $\mathbf{x}_{1,\ldots,t}$[9]. However, LSTM cells also contain separate functions or 'gates' that allow them learn to conditionally store and retrieve information over a large number of time-steps. A forget gate governs the extent to which historical information is retained. An input gate governs the extent to which current input $\mathbf{x}_t$ affects or modifies the cell's state and an output gate governs the extent to which the current state changes the intermediate output $\mathbf{h}_t$. These gate mechanisms allow LSTM cells to preserve long term dependencies between time steps by controlling when information is added or removed from the state variables and reducing the diluting effect of simply multiplying each input with the previous state as in SRNN.

For a more thorough mathematical definition of LSTM architectures, we direct the reader to Chris Olah's LSTM blog post [10].

2.6.6 Multi-Layer RNNs

As discussed in section 2.6.5.1 above, RNNs are less effective at encoding relationships between distant inputs, even despite improvements offered by the LSTM architecture (above). One way to overcome this problem is to stack and concatenate a forward RNN layer and an RNN that processes the input sequence in reverse order such that each intermediate output contains information relating to recently encoded inputs at the beginning and end of the se-

[9] the dimensionality of which is a hyper-parameter configured as part of the model development workflow discussed in section 2.6.2

[10] `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

quence. This approach has been shown to provide improved model performance (Goldberg, 2015) for a number of text processing tasks.

A common configuration in NLP is the bi-directional LSTM or bi-LSTM which serves as the basis for ELMo contextual embeddings (as discussed in section 2.5.7.1). In this configuration two LSTM cells are stacked on top of each other, one that processes sequence $\mathbf{x}_{t:t+n}$ forwards through time (i.e. examining each element in sequence $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{n-1}, \mathbf{x}_n$), producing intermediate outputs $\overrightarrow{\mathbf{h}'}_{t:t+n}$, and another that examines each step in the input in reverse order (i.e. $\mathbf{x}_n, \mathbf{x}_{n-1}, ..., \mathbf{x}_1, \mathbf{x}_0$) producing intermediate outputs $\overleftarrow{\mathbf{h}'}_{t:t+n}$. Intermediate outputs from the two layers are then concatenated together to produce a joint intermediate output which can be used for downstream tasks $\mathbf{h}'_t = [\overrightarrow{\mathbf{h}'}_t; \overleftarrow{\mathbf{h}'}_t]$.

2.6.7 Encoder-Decoder

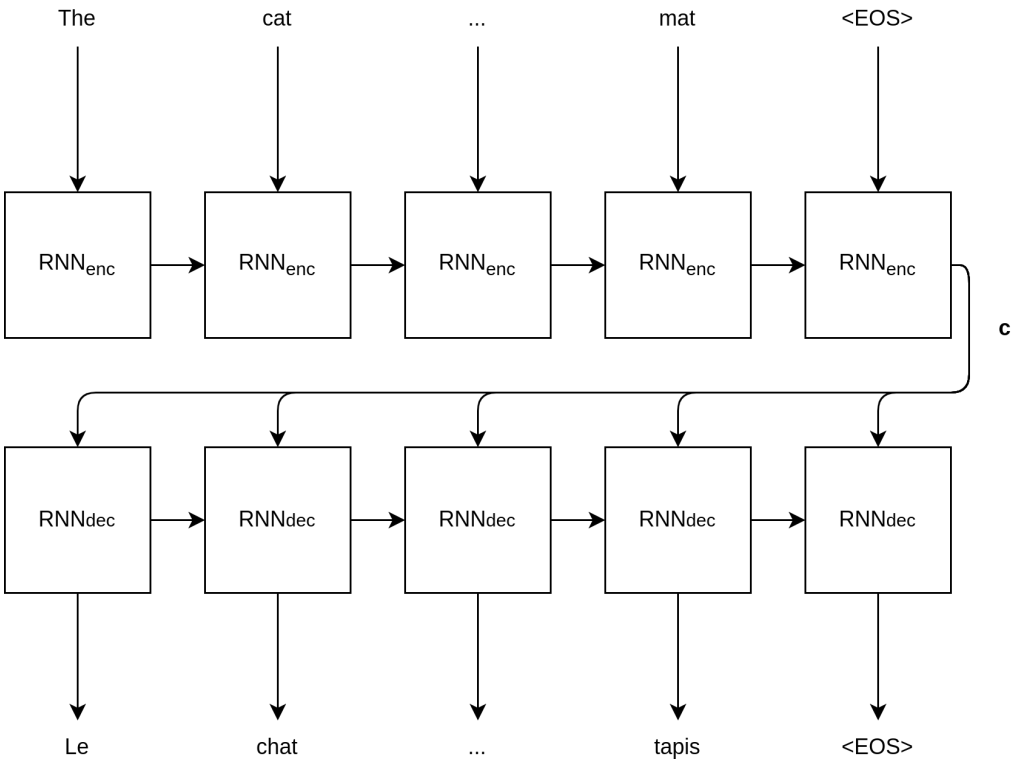


Figure 2.16: An Encoder Decoder model which encodes an English phrase into **c** intermediate representation which is decoded into French.

The Encoder-Decoder architecture provides a framework for building RNN models that transform or translate an input into a new output via a latent intermediate encoding. The Encoder-Decoder architecture was originally proposed to facilitate neural machine translation (Cho et al., 2014; Sutskever et al., 2014; Goldberg, 2015) but also offers a useful general architecture for a number of other NLP and non-NLP taskss (for example, Image

Captioning in which a model encodes an image into an intermediate representation which is decoded into a description of what was in the image Parikh et al. (2020)).

The general Encoder-Decoder architecture consists of two components that are connected together: the encoder receives the raw input **x** and encodes it into an intermediate context vector $\mathbf{c}'$. Then, the decoder receives the intermediate vector $\mathbf{c}'$ and translates it into an intermediate output **h** which is usually consumed by subsequent output layers and mapped to a final classification or regression output **y'**. Encoder-Decoder models are usually conditioned on sequences of inputs and outputs and the encoder and decoder are therefore usually RNN components. However, $\mathbf{c}'$ is normally a single vector requiring the encoder component

The model works by comparing the decoder-generated output sequence with the true output sequence and back propagating error through the entire network and updating all weights including those for the encoder. Some language modelling tasks such as machine translation, the length of the generated sequence may differ from that of the input sequence e.g. "how are you?" $\rightarrow$ "Ça va?". Therefore, language model decoders are often conditioned to predict a special end of input token, $<EOS>$ at the end of a sequence. Then, if during network inference $<EOS>$ is predicted as the most likely next token, the generation process may be halted.

2.6.8 Attention & Transformers

Even bi-LSTM layers can struggle to model relationships between inputs in very long sequences, particularly if those two inputs are very far apart (for example, there is a useful grammatical relationship between words at the beginning and end of a long sentence or short paragraph). The Attention mechanism (Bahdanau et al., 2014) was originally formulated as a way for decoders (section 2.6.7) to take into account the full input sequence **x** and learn to *focus* on the most important parts of the input during generation of outputs.

In an Encoder-Decoder network, the entire sequence of outputs $\mathbf{h}'_0, ..., \mathbf{h}'_n$ is generated by conditioning the decoder on **c**, often the final state of an LSTM layer as in the example in Figure 2.16. This can provide an information bottleneck since the model must learn to encode all the information it needs to produce the output sequence in a single vector and, as stated above, this is exacerbated for longer sequences as the model must compress more information into the context vector.

Attention mechanisms remove this information bottleneck by conditioning the decoder on the whole input sequence at each time step as illustrated in Figure 2.17.

Context vector $\mathbf{c}_i$ is derived by 'attending' to the entire input sequence as a weighted sum of encoder hidden states (or 'annotations' as described in Bahdanau et al. (2014)). For

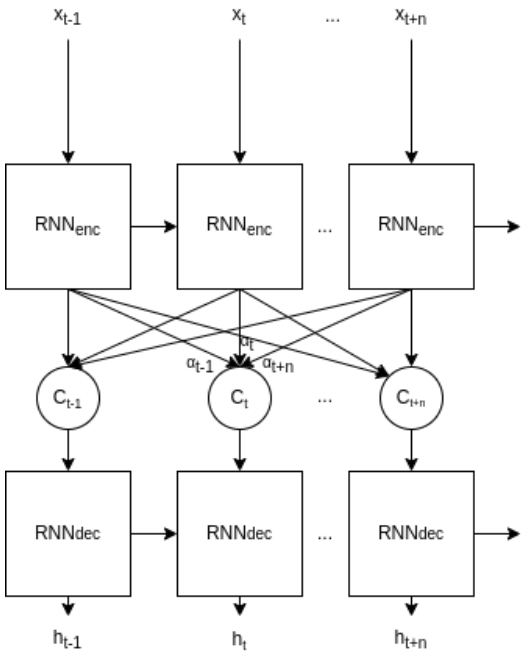


Figure 2.17: An Encoder Decoder with an Attention mechanism which allows the decoder to 'attend' over all inputs during generation of each output.

an encoded sequence $\mathbf{h}'$ with length n:

$$\mathbf{c}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{h}'_j \tag{2.45}$$

The weight α_{ij} of the annotation/hidden state is computed by applying an FFN with softmax output (section 2.6.3.4) to the intermediate output **h'**$_j$ and the previous hidden state of the *decoder*, s_{i-1}:

$$\alpha_{ij} = softmax(FFN(s_{i-1}, \mathbf{h}'_j)) \tag{2.46}$$

The parameters for this FFN are jointly learned through back propagation of the encoder/decoder model.

2.6.8.1 Comparing Attention and Convolutional Neural Networks

In section 2.6.4 we briefly discussed CNNs which learn to combine low level feature data from small, overlapping regions of the input data into more abstract complex features. Attention mechanisms serve a similar purpose, aggregating information from different regions of an input sequence in order to learn complex relationships and dependencies between them. However, the two mechanisms are quite different in practice.

CNNs are most effective at identifying relationships between neighbouring inputs e.g. pixels in the outline of a car, phrases in a sentence or characters in a word. They do

this by progressively pooling or combining information from receptive fields within each subsequent CNN layer which also has the effect of reducing the amount of information available. Long distance relationships between inputs are unlikely to be identified by CNNs unless it is deep enough that they appear in the same receptive field which may come with the trade off of significant information loss.

On the other hand, attention mechanisms allow models to carry out pairwise comparisons of all inputs without any data loss allowing them to learn to characterise both short and long distance dependencies between input elements. However, this flexibility comes at the cost of significantly increased computational complexity which increased quadratically with the length of the input sequence.

2.6.8.2 Transformers

Building on the success of the encoder-decoder-with-attention architecture, Vaswani et al. (2017) proposed the Transformer. This architecture removes RNN layers, instead directly passing in the full input sequence and applying attention mechanisms directly to densely connected layers. The primary motivation for this is improved computational throughput; GPUs facilitate parallel computation of large-scale tensor operations (e.g. matrix multiplication) but the calculation of an RNN's current state s_t depends upon each previous state s_{t-1} which limits the parallel processing of a given input sequence. By removing the RNN and the recursive dependencies between inputs, network propagation can be accelerated through parallelisation.

Transformer Architecture The original formulation of the Transformer follows the general encoder-decoder pattern (as described in section 2.6.7) and they consist of an encoder block and a decoder block (Figure 2.18.a). Encoder blocks consist of 6 identical stacks of multi-head attention mechanisms that feed into fully connected FFNs and then a normalisation layer. Decoder blocks have the same architecture except for the addition of a third layer that performs attention over the output of the encoder stack. Transformers can also be used in an encoder-only configuration as shown in (Figure 2.18.a) - this is the form used by BERT (Devlin et al., 2019).

Positional Encoding In lieu of inferred positional context learned by an RNN, information about the position of each input in the sequence, designated Positional Encoding (PE), is provided by adding sine and cosine waves to the inputs before they are fed into the attention model, depending on their offset in the sequence:

$$\begin{aligned} PE_{(t,2i)} &= sin(t/10000^{2i/d_{model}}) \\ PE_{(t,2i+1)} &= cos(t/10000^{2i/d_{model}}) \end{aligned} \tag{2.47}$$

Where t is the input time offset, i is the dimension in the embedding and d_{model} is the number of dimensions output by the transformer block. Vaswani et al. (2017) use 10,000 as the denominator to support a large number of unique PE values by cycling slowly through

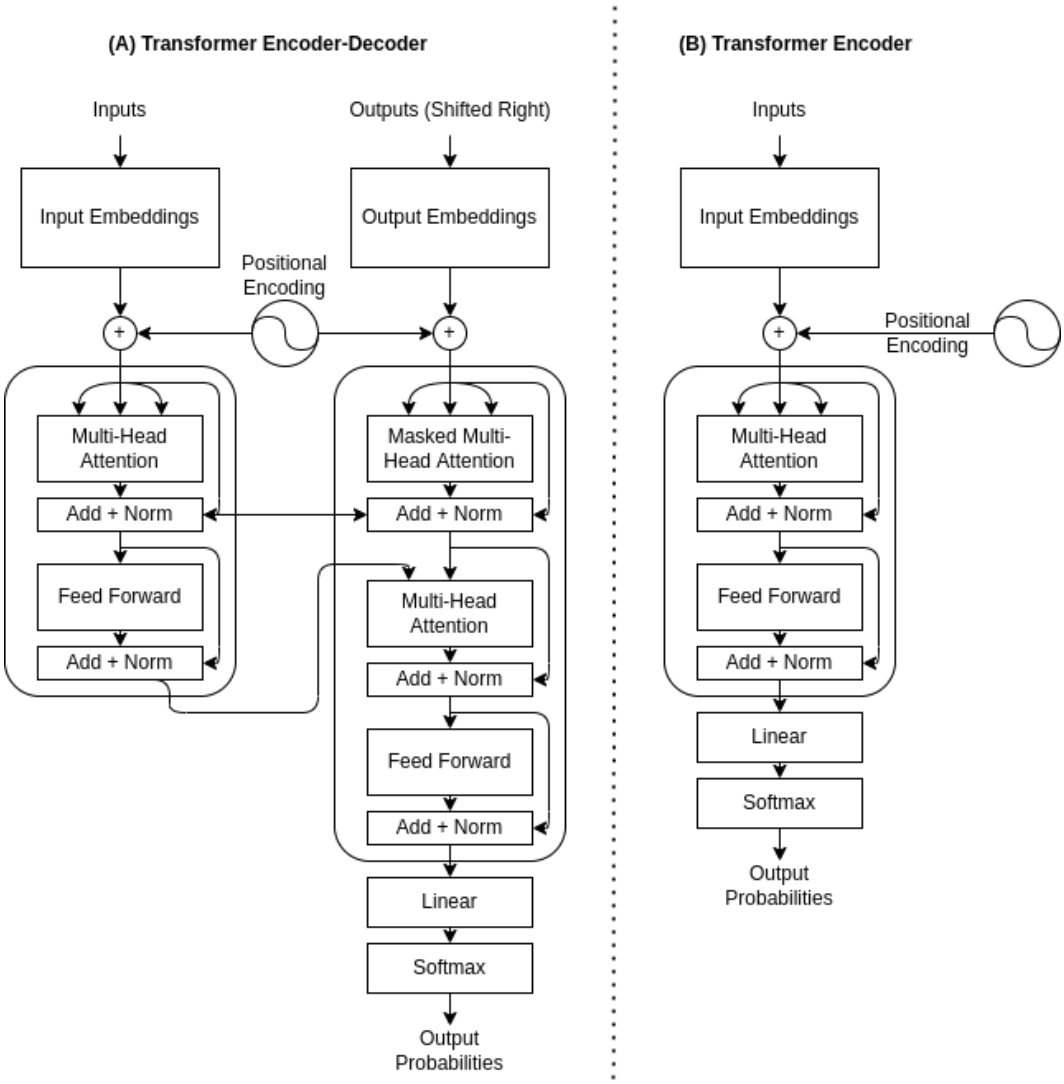


Figure 2.18: A transformer block

the , supporting long input sequences. The cycle over the sinusoidal wave forms. A more detailed and intuitive explanation of how this works is offered by Kazemnejad (2019).

Scaled Dot-Product & Multi-Head Attention Vaswani et al. (2017) generalise the definition of attention as mapping a query and a set of key-value pairs to an output where the query, **Q**, keys **K** and values **V** are all vectors with respective dimensionalities d_q, d_k, d_v for sequence with length T:

$$\begin{gathered} Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \\ \mathbf{Q} \in \mathbb{R}^{T \times d_q}, \mathbf{K} \in \mathbb{R}^{T \times d_k}, \mathbf{V} \in \mathbb{R}^{T \times d_v}, \end{gathered} \tag{2.48}$$

This formulation of attention, known as Dot-Product attention is shown to be faster

and more space-efficient in practice (Vaswani et al., 2017) than the additive attention approach proposed by Bahdanau et al. (2014) (see above). The variables **Q**, **K** and **V** contain different data depending on the task at hand. In the Encoder block, the input is used for all three. This is known as self-attention (Vaswani et al., 2017) and it allows the model to learn how best to attend to items in an input sequence with respect to other items in the sequence (for example, learning grammatical relationships between words). In the decoder block, Q comes from the output of the encoder block and K and V are the outputs generated by the decoder so far.

The model's statistical performance can be further improved by allowing the transformer to learn to apply multiple attention operations to different subsets of the input in parallel. This is known as Multi-Head Attention. Each attention head learns to attend complementary subsets or 'views' of the input sequence and the resulting attention outputs are concatenated together for downstream processing. For a model with h attention heads:

$$\begin{gathered} MultiHeadAttention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [head_1; ...; head_h]\mathbf{W}^O \\ head_i = Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \\ \mathbf{W}^O \in \mathbb{R}^{hd_v \times d}, \mathbf{W}_i^Q \in \mathbb{R}^{d_{model} \times d_Q}, \mathbf{W}_i^K \in \mathbb{R}^{d_{model} \times d_K}, \mathbf{W}_i^V \in \mathbb{R}^{d_{model} \times d_V} \end{gathered} \tag{2.49}$$

Feed Forward The third part of each transformer block is a standard FFN with two fully connected linear transforms. These layers are applied to the full sequence of attention outputs which typically use the *ReLu* activation function (section 2.6.3.3).

Add & Norm Finally, the add and norm blocks in the transformer provide model normalisation and stabilise training. Additive Residual connections proposed by He et al. (2015) reduce the risk of vanishing gradients and Layer Normalisation proposed by Ba et al. (2016) normalises activities of the neurons in each layer, helping to reduce training time and encourage faster convergence.

2.7 Transfer Learning & Neural Language Models

Transfer learning, the ability to apply a model trained on one problem to a different but related problem is not a new concept. Early transfer learning work dates back to the mid 1970s (Bozinovski, 2020). Techniques such as domain adaptation (Farahani et al., 2020) can be applied to a range of classical machine learning algorithms (e.g. SVM, section 2.6.2). However, the advent of deep learning has made neural transfer learning much more practical and accessible. As a result, transfer learning has become ubiquitous in modern Natural Language Processing with authors adapting general purpose language models such as BERT (Devlin et al., 2019) (section 2.5.7.2) and BERT-like derivatives to achieve a plethora of state-of-the-art results across numerous NLP tasks (Cattan et al., 2020; August et al., 2020; Wright and Augenstein, 2021; Joshi et al., 2020; Brack et al., 2021).

A typical neural transfer learning workflow involves taking a Pre-trained Language Model (PLM), and extending it by adding additional, task-specific parameters such as new randomly initialised dense layers (section 2.6.3.2). The whole model is then trained on the new task and both new and existing weight parameters are adjusted accordingly via back-propagation. Models are usually trained with a reduced learning rate (section 2.6.3.6) to avoid changing the PLM weights so that they no longer generalise well to new datasets, known as catastrophic forgetting.

Another strategy employed to avoid catastrophic forgetting is the use of Neural Adapter layers. Neural Adapters are additional intermediate layers of parameters that are added *between* layers in pre-trained models and fine-tuned on a specific task while the pre-trained layers are kept frozen (Houlsby et al., 2019). This allows the model to learn task specific weights in the additional layers whilst removing the risk of catastrophic forgetting in the original PLM layers.

2.7.1 Few-Shot & Neural Transfer Learning

Few-shot learning is a special case of supervised machine learning (section 2.1.1) where only a limited number of labelled training and testing instances (x, y) are available. Given the expensive and time-consuming nature of labelling training data, the ability to create accurate machine learning models from a small amount of data is particularly useful. Like transfer learning in general, state-of-the-art performance in few-shot learning settings has improved significantly in recent years thanks to the advent of deep learning and PLMs (Wang et al., 2021).

There are a number of strategies and techniques for few-shot learning which are described in detail by Wang et al. (2021). However, in this thesis, we primarily make use of few-shot learning via transfer learning, pre-training PLMs to complete related tasks and then fine-tuning them using smaller datasets to complete a primary task of interest. In some cases we make use of multiple stages of transfer-learning to this end. For example, In chapter 8 we start with a pre-trained RoBERTa model (a BERT variant by Liu et al. (2019a), see section 2.5.7.3), fine-tuning it on a large co-reference resolution (section 2.8) dataset and then fine-tune it again on our novel task.

2.8 NLP Tasks Relevant To This Work

In this section we describe a number of common NLP tasks that we make use of in the course of this thesis.

2.8.1 Text Classification

Text classification is a special case of general machine learning classification (section 2.1). Given a string of text x (which depending on the setting may be a document, a paragraph, a sentence or even shorter) we aim assign to it a label y from a set of possible labels L. A common case of text classification is email spam detection in which messages are labelled as spam or not spam.

2.8.1.1 Sequence Classification

Sequence Classification is a special case of classification in which we aim to assign every element in a list of inputs $x_0, x_1, ..., x_n$ a corresponding label $y_0, y_1, ..., y_n$. It is used in cases where relationships exist between inputs that provide useful context for the prediction of successive steps instead of modelling each output as an independent classification problem.

2.8.1.2 Named Entity Recognition

Named Entity Recognition (NER) is a sequence classification task in which each word in a text is labelled according to whether or not it belongs to a noun-phrase that relates to a Named Entity (anything that can be referred to with a proper name such as a person, a geographical place or an organisation (Jurafsky and Martin, 2009)). Named Entities can serve as useful contextual anchors for other NLP tasks - for example, we might want to find all news articles that mention a particular person or organisation, or we may wish to find all film reviews that mention a particular actor in a negative light.

2.8.2 Co-reference Resolution

Co-Reference Resolution is the task of recognising whether mentions of a named entity, noun or noun phrase in a document refer to the same entity (i.e. whether they are co-referent). It can resolve inconsistent surface forms of an entity (e.g. "Joe Biden made a speech... Later, President Biden said...") and anaphoric references ("Bob wished he had his umbrella. He was soaked through"). Co-Reference Resolution can be expressed as a pairwise operation on two named entities x_1, x_2 and their corresponding document contexts c_1, c_2:

$$y = p(x_1 = x_2 | c_1, c_2) \tag{2.50}$$

We describe the two flavours of co-reference resolution explored in this thesis below.

2.8.2.1 Intra-Document Co-reference Resolution

Intra-Document Co-Reference Resolution (IDCR) is used to identify co-referring mentions of entities within the same document. Intra-document co-reference resolution is a well understood task with mature training data sets (Weischedel et al., 2013) and academic tasks (Recasens et al., 2010). The current state-of-the-art model by Joshi et al. (2020) is based on work by Lee et al. (2017, 2018) and uses a modern BERT-based architecture. Comparatively, CDCR, which involves co-reference resolution across multiple documents, has received less attention in recent years (Bagga and Baldwin, 1998; Rao et al., 2010; Dutta and Weikum, 2015; Barhom et al., 2019). The model constructed by Cattan et al. (2020) jointly learns both entity and event co-reference tasks, achieving current state-of-the-art performance for CDCR, and as such provides a strong baseline for experiments in CD^2CR. The models of both Cattan et al. (2020) and Barhom et al. (2019) are trained and evaluated using the ECB+ corpus (Cybulska and Vossen, 2014) which contains news articles annotated with both entity and event mentions.

2.8.2.2 Cross-Document Co-reference Resolution

Cross-document co-reference resolution (CDCR) is the task of recognising co-referring mentions of the same entity across multiple documents. CCDCR is harder than IDCR and the difficulty scales with the number of documents.

CDCR is a useful NLP process that has many downstream applications. For example, CDCR carried out on separate news articles that refer to the same politician can facilitate inter-document sentence alignment required for stance detection and natural language inference models. Furthermore, CDCR can improve information retrieval and multi-document summarisation by grouping documents based on the entities that are mentioned within them.

2.8.3 Information Retrieval

Information Retrieval (IR) is the task of identifying and retrieving relevant information from a collection of documents based on a given query. An example of a widely-known information retrieval system is the search engine which uses a set of keywords entered by the user as a query and surfaces relevant web pages that ideally contain the information that the user is are interested in. Information Retrieval depends upon many of the document representation techniques discussed in section 2.5.1. Readers are directed to Manning et al. (2009) for an in depth introduction to and discussion of information retrieval concepts.

2.8.4 Semantic Textual Similarity

Semantic Textual Similarity (STS) deals with determining how similar in meaning two texts are (as opposed to other types of similarity such as thematic similarity). STS is a well-

defined task with popular annual workshops (Agirre et al., 2012; Marelli et al., 2014b). Current state-of-the-art models successfully measure semantic similarity within the context of these purpose-built corpora (Subramanian et al., 2018). STS often takes the form of a scoring/regression task where pairs of texts or documents are assigned a score from 1 to 5 depending on the degree of similarity between them and is usually carried out at sentence or phrase level.

2.9 Conclusion & Summary of NLP & ML Activities in this Thesis

This chapter has outlined a range of working patterns, tasks and computational techniques that can be applied to natural language and provided detailed descriptions of a range of machine learning models and neural architectures commonly used in NLP. The material covered in this chapter should have prepared the reader for the following NLP and Machine Learning activities:

- We use Linear and Support Vector Regression for predicting different types of scientific impact in Chapter 4.
- We make use of Support Vector Classifiers when we predict whether news articles are relevant or spammy and we apply information retrieval techniques to find scientific papers that are potentially related to news articles in Chapter 5.
- We explore a range of text and document representations and corresponding similarity metrics to carry out information retrieval experiments that identify corresponding pairs of news articles and scientific papers in Chapter 6.
- We use bag-of-words, word2vec, GLoVe and BERT embeddings to identify which parts of scientific papers are discussed prominently in news articles in Chapter 7.
- We train a RoBERTa-based CDCR algorithm to identify co-referring mentions of entities between news articles and scientific papers that they discuss in Chapter 8
- We train a series of few-shot learning regression models to identify and quantify scientific impact in news articles and scientific papers in Chapter 9

3

A Background in Scientometrics & Scientific Impact

"Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."

Jules Verne

Scientometrics is the quantitative study of science, communication in science and science policy (Hess and Hess, 1997; Leydesdorff and Milojević, 2013). Modern scientometrics was pioneered in the 1950s by Eugene Garfield who devised the Science Citation Index (SCI), an interdisciplinary index of citations which Garfield used to calculate a number of scientific impact metrics based on citations accrued (Goodwin and Garfield, 1980). The field has now evolved to encompass a range of methods and metrics for quantifying scientific outputs, many of which are used ubiquitously for evaluation purposes within academia. Scientometrics relates specifically to texts (i.e. scientific papers and related documents) as empirical units of analysis as opposed to fields like "sociology of science" which focuses on the individual behaviours of scientists in their laboratories (Leydesdorff and Milojević, 2013).

The recent exponential growth of scientific publications and focus by academics on collaborative work (Bornmann and Mutz, 2015) have motivated the need to better measure scientific impact. Scientists wish to better understand their own outputs and how they work can benefit the academic community and wider society and to find new opportunities for collaboration. Likewise, funding bodies, both private and public, increasingly want to understand where their support is likely to yield the biggest returns and now expect research scientists to plan for and demonstrate the impact of their work as part of their grant application process.

There are numerous ways that scientific work can have impact on the world around us and likewise, many ways for us to measure this impact. In this chapter we focus on research questions **RQ1.1** and **RQ1.2** (section 1.2) as we explore current, widely used metrics for measuring scientific impact. We categorise scientific impact into two broad classes which we describe in detail along with metrics for quantifying impact below.

3.1 Academic Impact

Academic Impact refers to the influence that that scientific research has within the academic sphere. For example, the contribution of novel theories, methodology, data set or models that are widely used or that enable or inspire new works. Academic Impact metrics are often used to rank scientific journals in terms of which ones are most likely to allow a scientific

work to reach a broad audience and become impactful in its own right. Over recent decades, starting with Garfield's experiments on his SCI data set (Goodwin and Garfield, 1980), academic impact has most frequently been quantified in terms of citations received. However, the rise and prevalence of internet publishing and academic use of social media has also led to the development of alternative metrics (or 'altmetrics') for quantifying academic impact. Altmetrics are typically data driven, using data from social media interactions and natural language processing to quantify academic impact in new, more granular ways. In this section we describe and compare some of the most commonly used of these academic impact metrics.

3.1.1 Citation-based Metrics

Citation-based impact metrics are built upon the intuitive assumption that people whose work receives more citations have had more impact (and likewise journals that receive more citations are more impactful and desirable to publish in). Citation-based metrics can be aggregated and used to quantify research success at individual, departmental and institutional levels. Citation-based impact metrics are the most prolifically used impact metrics used in the academic community today, often used by university departments to rank academic outputs as a deciding factor in career progression and tenure applications (Leydesdorff and Milojević, 2013).

3.1.1.1 Journal Impact Factor (JIF)

Journal Impact Factor is one of the original citation-based impact metrics defined by Garfield (2006). The metric ranks scientific journals in terms of the mean number of citations of articles published in the last two years in a given journal. Garfield originally calculated JIF annually using his SCI data set but this is now done by a private company Clarivate [1]. There are a number of other providers of JIF and similar, journal-level metrics. Most charge for their data but some, such as Elsevier, provide free access to JIF data[2].

Whilst journal level metrics like JIF do provide some insight into the relative success of academic journals, they cannot be used to meaningfully understand the individual contributions of institutions or scientists. JIF-style metrics are also highly susceptible to skew from successful or unsuccessful outlying papers and since data is usually reported at a high level, it is often impossible for an observer to rationalise or indeed reproduce scores independently (Rossner et al., 2007). Additionally, research has found JIF and h-Index can be manipulated through self-citation (Meho, 2007; Bartneck and Kokkelmans, 2011).

[1] https://clarivate.com/

[2] https://www.scopus.com/sources

3.1.1.2 Citation Count & h-index

A common way to evaluate an individual scientist's academic contributions is to simply sum the number of citations that they have received or to take the mean of their citations-per-publication. However, these are crude measures that are often skewed by outlying data points and may be more representative of individual papers than an academic's full back-catalogue of work. For example, a researcher who has a large number of lesser known publications but who has a single breakthrough project which is widely publicised and cited could end up with a high mean citations-per-publications count.

Hirsch's h-Index is an author-level metric that is able to distinguish between frequent strong publishers and publishers with a few exceptionally popular papers. h-index is defined by Hirsch so that "a scientist has index h if h of [their] N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each." where N_p denotes the number of papers that they have published (Hirsch, 2005). This means that an author can only achieve an h-Index of 50 if they have published at least 50 times and 50 or more of their publications have at least 50 citations each. A number of improvements and variations to h-index have been proposed including g-index (Egghe, 2006) which allows highly cited papers to bolster low-cited papers as part of an author's overall score and i10-index, used by Google Scholar[3] which simply counts how many of an author's papers have more than 10 citations.

3.1.1.3 Mean Normalised Citation Score (MNCS)

The accumulation of citation and conventions for *when* and *why* a work should be cited vary significantly across different scientific fields (Mcallister et al., 1983; Waltman, 2015). More recent works can also expect to receive more citations as the number of new publications, and thus the probability of being cited by one of them, increases exponentially (Bornmann and Mutz, 2015). Therefore, whilst the simplistic strategy of counting a paper or author's citations and comparing this sum with the total number of citations that their peers received may seem attractive, it does not yield a fair comparison across different fields of study or indeed over different periods of time.

Mean Normalised Citation Score (MNCS) is a metric designed to facilitate cross-discipline comparison of citation outputs by normalising for year of publication and scientific sub-field. The "expected" number of citations for a given paper is determined by taking the mean of citations for papers in the same field published in the same year. The MNCS value is the ratio of the actual number of citations a paper received in comparison to the "expected" citation count (Waltman, 2015). An author's overall contribution could be calculated by taking the average of their MNCS scores for all of their papers. However, this approach is still prone to skew from papers that are exceptionally popular (with respect to

[3] `https://scholar.google.co.uk/`

their field and year of publication) and thus it is very difficult to differentiate between authors who have many successful publications and authors who have one or two particularly well known works.

3.1.2 Alternative Metrics (altmetrics)

Alternative metrics or "altmetrics" attempt to measure academic impact based on evidence from the internet including social media and other relevant data sources (Priem et al., 2010; Piwowar, 2013). In recent years, a number of systems that serve this purpose have been proposed:

- Altmetric (`https://altmetric.com`) is an organisation which uses a publication's online footprint (Twitter mentions, Facebook posts and shares etc.) to award it an impact score Adie and Roe (2013).
- Impact Story (`https://www.impactstory.org/`) is an online service that provides a combined view of academics' citations and social media footprint in order to try to provide meaningful context around a person or institution's academic impact. The service generates profiles for researchers automatically using their unique ORCID ID (Haak et al., 2012) and "gamifies" scientific impact by awarding authors with badges that represent milestones in impact.
- Semantic Scholar (`http://www.semanticscholar.org/`), whose primary function is as a research search engine, also offers some novel features to enable academic impact monitoring, such as citation importance classification (Valenzuela et al., 2015a) and graphs of citation velocity (how many citations a work receives per month) and acceleration (change in citation velocity).
- McKeown et al. (2016) explored using NLP technologies to extract information from the full text of academic papers in order to track the prevalance of new technical terms in the community, such as 'microRNA'. Similarly Prabhakaran et al. (2016) use topic models (section 2.5.5) to track which contexts topics are used in over time.

Compared to the citation-based metrics discussed above, the data-driven approaches provided by these systems can produce much richer contextual insight into how the impact of scientific work propogates through the academic community and the framing of impact generated. However, they rely upon up-to-date data concerning the works they are measuring and recent studies have shown that altmetric data coverage varies significantly by scientific discipline (Banshal et al., 2019) and the systems can be gamed through the generation of false likes and mentions using 'bots' (Bornmann, 2014).

3.1.3 Criticism of Academic Impact Metrics

The widespread use of citation-based impact metrics by the academic community for assessing individual performance has faced growing criticism in recent years. The European Association of Science Editors (2007) observed that academic institutions routinely and inappropriately use JIF to measure the productivity of individual researchers and their quality of their work even though it is a journal level metric and cannot be meaningfully applied to individuals. Likewise, Jorge Hirsch, creator of the h-index, has criticised the over-use of his metric for assessing academic performance (e.g. use of h-index by academic recruiters to screen job applicants) (Hirsch, 2020). Edwards and Roy (2016) argue that increased pressure to publish and reduced funding can lead to unethical behaviour from scientists who want to keep up appearance. Many of these issues are also applicable to altmetrics (Bornmann and Haunschild, 2016) which (as discussed above) can also be gamed, potentially leading to some of the concerning behaviours discussed by Edwards and Roy (2016).

In response to the debate on when, where and how to use academic impact metrics, a number of declarations such as DORA[4], the Leiden Manifesto (Hicks et al., 2015) and Hong Kong Manifesto (Moher et al., 2020) have been written, advocating for the development and use of broader, more inclusive methods for assessing research and reduced dependence on traditional citation-based approaches (Overlaet, 2022).

3.2 Comprehensive Impact

In contrast to *academic* impact metrics, we assign the name Comprehensive Impact to the broad impact of scientific research upon society, culture, the economy, policy and the natural environment. Interest in these types of impact has increased in recent years leading to the development of many new metrics (HEFCE, 2012, 2019; Lane and Bertuzzi, 2010; Steingard et al., 2022; Sørensen et al., 2022). However, comprehensive impact can manifest in many forms, often dependent upon the scientific discipline that it originated within. Medical researchers may generate impact by saving lives with new treatments. Chemists and engineers may generate impact by inventing new, efficient manufacturing processes that save millions of dollars and reduce pollution. Understanding the extent to which such works impact society would not be feasible without investments in large centralised data collection processes.

The broad and fragmented nature of comprehensive impact presents a further challenge: impact generated within different scientific disciplines may not be directly comparable. How can we directly compare lives saved directly by new medicines with money saved and pollution reduced by improved manufacturing processes? Many recently-introduced comprehensive impact metrics are specialised to a particular scientific discipline such that cross-discipline comparison is not possible. For example, Steingard et al. (2022) propose

[4] https://sfdora.org/

measuring academic publications based on how well they align with the United Nations Sustainable Development Goals which is a helpful metric but only provides a partial view of scientific work through the lens of sustainability. Likewise, Sørensen et al. (2022) propose a questionnaire-based metric with dimensions that are specific to the field of Occupational Health which may be inappropriate or irrelevant in other disciplines. Whilst building a questionnaire for measuring performance in other disciplines is possible, questionnaires from different disciplines would not be comparable.

There have also been some attempts at creating broad, cross-discipline comprehensive impact metrics. However, these typically rely on centralised efforts made at government level (Lane and Bertuzzi, 2010) and require the investment of significant human effort (HEFCE, 2012), limiting where and by whom they can be used as we discuss below.

3.2.1 Notable Comprehensive Impact Metrics

3.2.1.1 STAR METRICS

STAR METRICS (Lane and Bertuzzi, 2010) is a United States Government project aiming to platform and tools that records where federal funds are invested in research and "offers the scientific community the opportunity to be proactive and to augment anecdotes about the value of science to the nation's health, security, and economic vitality" (Largent and Lane, 2012). The program was run by a consortium of federal government agencies including the White House Office of Science and Technology Policy, the National Institutes of Health, the National Science Foundation, the US Department of Agriculture and the US Environmental Protection Agency (Topousis et al., 2010). It was primarily concerned with understanding scientific impact in 4 key areas:

- Economic Growth - e.g. the number of patents files and spin-out businesses started as a result of scientific innovation
- Workforce outcomes - e.g. how many students are hired into jobs relating to government funded research programmes and spin-out companies
- Scientific Knowledge - e.g. academic impact measured through publication and citation counts
- Social outcomes - e.g. health and environmental outcomes that can be tied to scientific funding.

At time of writing, STAR METRICS has been retired. However, a number of spin-off projects have been created including USASpending[5], an open data platform that enumerates United States government spending on scientific grants.

[5] `https://www.usaspending.gov/`

3.2.1.2 Research Excellence Framework (REF)

The Research Excellence Framework (REF) is a United Kingdom Government-run assessment system for evaluating the quality of research conducted at UK institutions designed to highlight examples of good scientific research and to demonstrate examples of a variety of different impact types through the publication of qualitative impact case studies (REF 2014, 2011; HEFCE, 2019). REF is a centrally managed, labour intensive process which executed once every 4 years. Research Disciplines are approximately divided up into 34 ***Units of Assessment (UoAs)*** and each university may prepare one submission per UoA per round of assessment.

The framework measures three distinct elements:

- the quality of research **outputs** such as publications performances and exhibitions, this metric aggregates academic impact and was recently found to correlate strongly with citation-based metrics (Pride and Knoth, 2018).
- the **impact** of research beyond academia - i.e. its comprehensive impact
- the **environment** that supports the research i.e. the quality of the facilities at the university and the people involved.

In this thesis, we specifically focus on the REF **impact** scoring mechanism. REF defines impact as "effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia (REF 2014, 2011, p. 26)." This definition is very close to our definition for comprehensive impact. Each REF submission includes an impact case study for assessment purposes, typically providing details of the type of impact obtained as well as the names of external sources that can corroborate the impact, for example industrial partners who have directly benefited as a result of the research (REF 2014, 2011, pp. 27-30). Impact case studies are then evaluated by a UOA-specific expert assessment panel who assign it a score between 0 (no impact) and 4 (high impact).

Despite representing a step in the right direction for understanding a diverse range of impact types, the REF still suffers from a number of shortfalls. The assessment process is resource intensive, requiring a committee of academics to evaluate each submission individually, necessitating brief 3-5 page submissions (REF 2014, 2011, p. 51). Thus reports tend to focus on a few high impact works from institutions, penalising academics who contribute in small amounts to many projects.

Whilst REF does contain provisions for interdisciplinary researchers, submissions are only assessed by one UoA sub-panel (a panel that judges work deemed to be in a similar academic discipline e.g. Physics, Mathematics, Biology and so forth) (REF 2014, 2011, p. 15) placing great importance on employing diverse assessment panels who can fairly judge interdisciplinary work. The subjective nature of human assessors, combined with flexible

guidelines and varying opinions on what makes a good case study is also a potential weakness of this system. The recent REF 2021 assessment acknowledged the importance of interdisciplinary research by including an interdisciplinary research specialist in each UoA assessment panel, a specific interdisciplinary output marker to be applied to such work and a section on institutional support for interdisciplinary research in the environment assessment (HEFCE, 2017). However, the effect of these revisions on REF outcomes relating to interdisciplinary work has not yet been studied in depth.

Additionally, REF has a strict 10-year assessment window within which all supporting research must have been conducted."Sleeping beauties in science" are publications that received very little attention at the time of publication and then suddenly become popular overnight, often due to an enabling breakthrough in another area (van Raan, 2004). For example, the Long-short Term Memory architecture for neural networks which was developed in 1997 (Hochreiter and Schmidhuber, 1997) but only recently became popular due to advances in parallel processing in Graphical Processing Units (see section 2.6.3.2). It is likely that these papers would therefore be excluded from REF submissions, both at initial time of publishing when they seem unimportant and after their true value is revealed, many years after the assessment period.

Finally, the expensive and labour intensive nature of the REF process means that the process can only be run once every few years and must be centrally funded and coordinated by the UK Government. Therefore, REF impact score is of limited use for short-term measurement or even regular self-assessment of the comprehensive impact of one's work.

3.3 Conclusion

In this chapter I introduced the field of scientometrics and a scheme for classifying scientific impact into two broad groups: academic and comprehensive.

I discussed academic impact, the impact that scientific work has on the rest of the scientific community and how such impact is usually measured using citation-based metrics like JIF and h-Index. I described how the rise of internet publishing and the ubiquity of social media have also paved the way for data-driven 'altmetrics' which provide additional context about how impact may be propagated across the academic community. I also outlined how academic institutions have faced a recent wave of criticism for the ways that they apply citation-based metrics and altmetrics to measure individual performance of their academics. Finally I talk about how this backlash has led to the publication of a number of manifestos and position papers arguing for the user of broader, more representative and inclusive metrics that measure scientific outputs beyond papers and citations.

I have also defined comprehensive impact, the impact of scientific work outside of academia on society, policy, the economy and the environment. I explained how comprehensive Impact is particularly difficult to measure because it covers a very diverse set of

outputs that may be different for different scientific disciplines. I described the advantages and disadvantages of two prominent comprehensive impact metrics: STAR METRICS and the United Kingdom's Research Excellence Framework (REF).

In this thesis I use REF Impact Score as a baseline for measuring comprehensive impact at a national level. In the following chapter I explore the relationship between REF Impact score and the most popular academic impact metrics. In successive chapters I use data collected from the REF 2014 assessment to explore how discussion of scientific work in news articles relates to its associated REF impact score.

4

A Statistical Comparison of Academic & Comprehensive Scientific Impact Metrics

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind."

Lord Kelvin, 1883

4.1 Introduction

The academic community is well known for its regard of publications and citations thereof as proxies for scientific impact, giving rise to the widely-used aphorism "publish or perish". Such impact is commonly measured using instruments discussed in section 3.1.1; h-index (Hirsch, 2005) for individuals and Journal Impact Factor (JIF) Garfield (2006) for journals & publication venues. Widespread online dissemination and discussion of scientific work has also led to the creation of alternative metrics (altmetrics) (Priem et al., 2010; Piwowar, 2013) like those discussed in section 3.1.2 which track web activity corresponding to scientific works, including tweets, blog posts, social shares and 'likes'. Despite their practical differences, citation-based metrics and altmetrics both serve the purpose of indicating of how much attention publications receive from other academics and internet users more broadly. However, questions about what these metrics really tell us about scientific impact have been raised (see section 3.1.3) leading to the creation and adoption of comprehensive impact instruments such as STAR Metrics (Lane and Bertuzzi, 2010) and UK REF Impact Score (HEFCE, 2019, 2012) (section 3.2) which measure the comprehensive scientific impact of academic works by taking into account their real-world outcomes such as the incorporation of startups and creation of jobs, the allocation of new patents, novel health outcomes and treatments, the creation or amendment of legislation or increased public awareness through media coverage. However, these metrics rely on time-consuming human-centric processes or data that is only collected within certain geographies or disciplines. Thus, at time of writing, no such metrics have yet been adopted at large by the global academic community.

In this chapter we study **RQ1.3** (section 1.2) by investigating the statistical relationships between academic and comprehensive impact. If academic, citation-based and altmetric-based scientific impact is predictive of comprehensive impact then real world outcomes of scientific works could be forecast or approximated without the need for expensive processes and data collection infrastructure. However, it is also possible that comprehensive and academic impact metrics are statistically independent or that the relationship is

discipline dependent e.g. a breakthrough in mathematics may generate academic interest but may not have practical applications without further applied research. We carry out a correlation analysis of widely-used citation-based metrics, altmetrics and UK REF Impact Score, a comprehensive impact metric used nationally in the UK, to see whether academic and comprehensive impact scores correlate and to what extent academic citation-based and altmetrics can be used to predict the comprehensive impact of a scientific work.

This chapter is based on our publication *'Measuring Scientific Impact beyond Academia: An Assessment of Existing Impact Metrics and Proposed Improvements'*. published in PLOS ONE 12, no. 3 (9 March 2017) and accessible at `https://doi.org/10.1371/journal.pone.0173152`.

The work in this chapter was carried out between 2016 and 2017 and focused on the REF 2014 study. It was subsequently published as a journal article in PLoS One which has since accrued 150+ citations [1]. Some of our original criticisms of REF 2014 study were addressed in the more recent REF 2021 assessment. However, as we discuss in this chapter, there are still a number of areas that could be improved. We also note a more recent study (Wooldridge and King, 2019) which arrives at a different conclusion to us about correlation between altmetric score and REF Impact score. We discuss this study and its implications in detail in section 4.7.

Our primary contributions are:

- A method for practical combination of heterogeneous citation metadata from open access repositories into citation networks
- An information retrieval approach for identifying and linking to scientific publications from unstructured and inconsistent bibliographies in REF case studies.
- An open-access dataset linking UK REF case studies to scientific publications derived from 7.4 Million scientific papers
- A robust statistical correlation analysis of REF Impact Score against citation-based metrics and altmetrics

4.2 Method

To facilitate the analysis of existing academic impact metrics in relation to the REF impact score we perform the following steps:

- We collect REF impact case studies submitted to the 2014 REF assessment and extract structured information such as Institution and Unit of Assessment.

[1] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0173152

- We extract paper metadata, including titles, authors and citation relationships between papers from three popular open-access scientific publication aggregators to build up a large collection of publication metadata.
- This information is used to form three citation networks, data structures that describe citation relationships between a large collection of papers which can be used to generate citation-based metric scores for any paper in the network.
- We generate one citation network for each of the open-access resources and we call these the Experimental Citation Networks (ECNs).
- We develop and run a matching pipeline which we use to programmatically search for links between REF impact case studies and papers in the ECNs.
- We use our ECNs to calculate citation-based metrics for REF-related papers and use an external sources to provide altmetric data.
- We visualise the calculated impact metrics and REF impact scores for linked case studies and ECN papers in a graph and report associated correlation scores.
- Finally we construct and evaluate a machine learning regression model using the calculated impact metrics as input and REF impact score as output.

Our data collection process, text processing and matching pipeline and our data models are described in more detail below.

4.2.1 Data Collection - REF Impact Case Studies

REF make available their entire back catalogue of REF2014 impact case studies (6637 case studies), via their website[2]. These were downloaded via an automated process and stored in a relational database (the structure of which is described in more detail in section 4.2.3 below).

REF submissions are grouped by Unit of Assessment (UoA) for assessment purposes. UoAs broadly correspond to scientific disciplines, for example, “Computer Science and Information Technology" or “Clinical Medicine." Although the more recent REF 2021 assessment includes special provisions for evaluation of interdisciplinary research (HEFCE, 2019), the REF 2014 process required that multi-disciplinary research must be submitted under a single UoA. Assessment panels were permitted to consult each other in the case of multi-disciplinary work. However, the final 2014 REF impact score was allocated with respect to the UoA that the work was submitted under.

REF impact case study scores are released on a unit-of-assessment-per-institution basis (where a single department/UoA at an institution may submit multiple studies depending on their size in accordance with the REF 2014 guidelines (REF 2014, 2011, p. 28)).

[2] `http://impact.ref.ac.uk/CaseStudies/`

Impact scores allocated to these studies are on a scale from 4* (excellent) down to 1* (poor) or unclassified if the work was not deemed to have any impact. To avoid institutions taking unfair action against academics with low quality REF results, per-case study results are not made available. Instead, for each UoA and institution, the percentage of case studies that have been deemed to be in each of these 5 classes of impact is provided. This makes reporting results on a scale more granular than "per-UoA-per-Institution" (e.g. "Computer Science at University of Warwick") impossible since we cannot know the impact score of any individual case study. Therefore for each of our experiments, we calculate scores for the metric under examination for each REF case study. Then we work out the mean score for said metric per-UoA-per-Institution. For example, the average score for all REF case studies submitted from Computer Science in Warwick would count as one data point and all REF case studies for Computer Science in Aberystwyth as another and so forth. Since case studies can only be submitted to one UoA, these per-UoA-per-Institution result groupings can be considered disjoint.

4.2.2 Data Collection - Scientific Papers

There are a large number of online sources for academic papers and related metadata that could be processed into citation networks. Some sources such as Web of Science[3] even provide their own citation networks that can be consumed automatically. For our study, we deliberately use open access sources where possible to maximise reproducibility. We also aimed to generate citation networks that are as large and diverse as possible rather than rely on data dumps from individual open-access journals, many of which address very specific areas of research (such as yeast cultures in Biology or deep neural networks within Computer Science). Large, diverse citation networks facilitate more accurate calculation of citation metrics for the REF studies under examination since there is a high probability that citing papers are also included in the citation networks. Research aggregators such as CiteSeerX[4], PubMed Central[5] and arXiv[6] collect open access publications and pre-prints from across broad scientific domains (Computer Science, Biology/Medicine and Mathematics/Physics respectively) are ideal sources for building and collecting large citation networks that cover these scientific domains comprehensively.

Snapshots of citation networks from RefSeer (Huang et al., 2014) (which uses data from CiteSeerX) and Paperscape (George and Knegjens, 2014) were both downloaded and integrated into the data model. RefSeer primarily contains papers relating to computer science and information technology, Paperscape is a citation network built from arXiv, an open access research aggregation service that hosts mostly papers and pre-prints concerned with mathematics, physics and some computer science. These citation networks contain

[3] http://ipscience.thomsonreuters.com/product/web-of-science
[4] https://citeseerx.ist.psu.edu/i
[5] https://www.ncbi.nlm.nih.gov/pmc/
[6] https://arxiv.org/

approximately 5.3M and 903K papers respectively. A citation network was also generated for the PubMed Central open-access collection[7] which containing a further approximately 1.2M biology and medical papers.

RefSeer provides rich citation data, including self-citations but does not provide author information which initially prevented us from calculating per-author h-index for the RefSeer citation network. However, we were able to use Sickle[8] to access the CiteSeerX Open Archives Initiative (OAI) repository, containing all CiteSeerX and therefore RefSeer paper metadata. We searched the OAI repository using the IDs of papers from the RefSeer data dump and stored it alongside the initial citation network in the relational database system.

We henceforth refer to the citation networks assimilated and enriched from RefSeer, arXiv and PubMed Central for the purpose of this study as the Experimental Citation Networks (ECNs).

4.2.3 Data Model

The relational data model used to store our ECN data is shown in Fig 4.1. Papers can have many authors, a title and a year of publication. Many-to-many relationships between the citee and citer fields of the citations table and the papers table facilitate the construction of directional citation graphs. Links between REF Impact Case Studies and individual papers are defined via the study_papers table and our methodology for discovering these links is described in Section 4.2.4 below.

4.2.4 Linking REF Case Studies and Other Data Sources

REF Impact Case Studies are 4-5 page documents which are primarily unstructured free text aside from basic metadata such as institution name and UoA. This makes them particularly challenging to process and link to the ECNs. Each study contains a bibliography section which is typically populated with formal, structured, references to underpinning scientific publications that contributed to the impact described in the study. The processing pipeline outlined in Fig 4.2 was developed to extract these references and use them to query our ECNs to formalise links between REF impact case studies and scientific publications and store them in the relational database.

The first stage in the pipeline uses regular expressions to identify strings that appear to be citations within the case study bibliography. Formatting of bibliography entries can vary based on standard practices within an institution, discipline or even based on the personal preferences of of the author. Therefore, we use a greedy regular expression pattern to extract all possible reference strings and further parse them using Freecite[9], a pre-trained

[7] https://europepmc.org/downloads/openaccess
[8] http://sickle.readthedocs.io/en/latest/
[9] http://freecite.library.brown.edu/

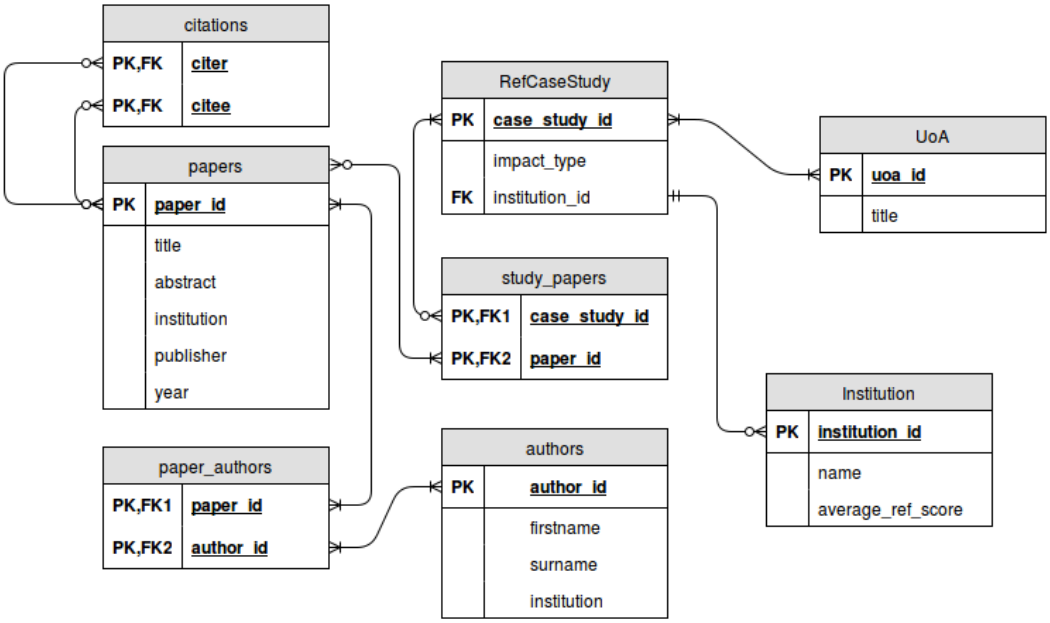


Figure 4.1: Entity relationship diagram describing how relationships between publications (papers), REF studies, institutions, authors and inter-paper citations

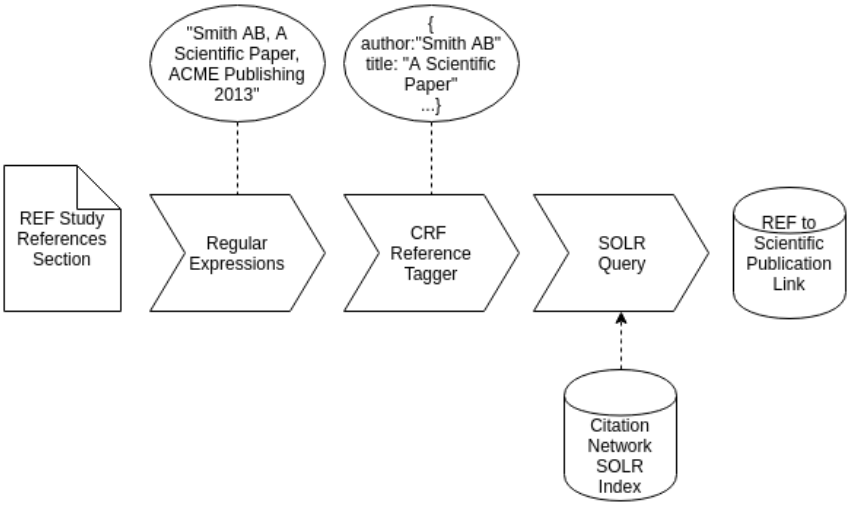


Figure 4.2: Pipeline process for linking REF studies to publications

Conditional Random Field (CRF) model that identifies possible boundaries for the title, author, journal name and year within each reference string.

Next, we attempt to match the references against the ECNs. Paper metadata from all ECNs is indexed in an Apache SOLR search index and each of the raw reference strings, as extracted by the regular expression matcher in the first step, are used to search the SOLR index. Any candidate papers results from the SOLR query are compared field-by-field with the output from the Freecite model and if the title and author are a close enough match, a link between the REF case study and the publication is created in the database. Matching

between the SOLR results and extracted case study reference is very strict, requiring exact match between the publication titles (normalising for case and punctuation) as well as a match for at least one author. Publication titles provide a high degree of uniqueness within our dataset and we estimate that there are very few false positives in our dataset at the expense of reduced recall for citation matches. An alternative matching strategy taking into account author name and year of publication was also tested. However, we found a large number of authors with the same surname and initial who published in the same year which yielded an unacceptable number of false positives. Matching based in title substrings also yielded a large number of false positives, particularly in cases where the extracted reference publications had short titles.

Table 4.1 lists the number of links from REF studies made for each ECN. Since REF bibliographies are free text fields and the exact number of references is unknown, we are not able to calculate recall or precision. However our regular expression and CRF pipeline returns a total of 6627 references and of these we are able to match 1052 papers from our ECNs.

Table 4.1: Number of links between scientific papers and REF studies identified for each experimental citation network using the process outlined in secton 4.2.4

ECN Source	**REF Studies**	**Papers**
arXiv	68	91
CiteSeerX	370	639
PubMed Central	273	322
Total	711*	1052

*There are 647 unique studies linked across the three ECNs, some of which appear multiple times giving a total of 711. Paper duplication is explored in section 4.3.1 below.

4.3 ECN and REF Matching Results

Our search pipeline was able to identify links between papers in the available ECNs and 647 unique REF case studies. Since smaller granularity is not possible (as discussed in section 4.2.1 above), results are grouped by UoA-per-Institution yielding 235 result groups (with an average case study population of 2.65 and Standard Deviation of 3.2) for further experimentation and visualisation.

Figure 4.3 below shows the number of institution level submissions per UoA for the top 10 UoAs. The composition of these most frequently identified links is largely as one might expect since the three main ECNs at our disposal are arXiv, which mainly contains papers pertaining to Physics, Mathematical Sciences and Computer Science and Informatics, CiteSeerX which contains mostly works related to Computer Science and Informatics and PubMed Central, which mainly contains works in the Allied Health Professions, Dentistry

Nursing and Pharmacy, Clinical Medicine and Psychology, Psychiatry and Neurscience fields.

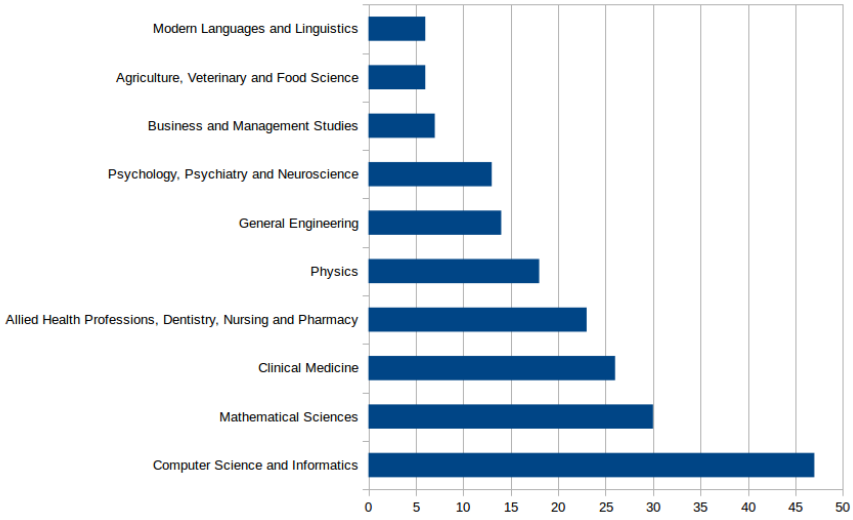


Figure 4.3: Breakdown of Institution Submissions found to have links to paper in the ECNs per Unit of Assessment. We show here the 10 UoAs with the largest number of ECN-linked submissions only.

The predominance of Computer Science papers as the largest UoA is due to the fact that the CiteSeerX ECN, which only focuses on Computer Science research, is the most homogeneous network among the ECNs we consider. By contrast, papers from PubMed are further distinguished into several UoA: Clinical Medicine, Allied Health Professions, Psychology etc. Papers from arXiv can be distinguished into papers from Physics, Mathematical Sciences etc.

4.3.1 ECN Overlap and Duplicate Papers

Internally the ECNs contain no duplicate papers, however since authors are free to publish their work in more than one journal or repository, there is potential for overlap between ECNS. ECN metadata is quite sparse and there are no computationally efficient means for identifying duplicate papers between networks other than to do a normalised title comparison for every paper title as discussed in section 4.2.4 above.

Since there are of the order of 7 million papers in our ECNs, we restricted the search and comparison to the 1052 papers that have explicit links to REF case studies, as discovered above, and those that cite them. This is sufficient in explaining the extent of ECN duplication within the scope of our study. We found that of the 1052 papers linked to REF case studies, 63 were duplicate entries. This duplication mainly stems from the PubMed and RefSeer ECNs which both have significant coverage of Bioinformatics publications. We confirmed that the papers within the ECNs that cite the 63 duplicate papers do not

themselves overlap. This is the ideal case and allows the duplicated entries to be treated as single entries with complementary, inbound citations from both ECNs.

4.4 Incorporating Altmetric Data

There are a large number of companies who provide altmetric data (as discussed in Section 3.1.2). For the pruposes of our experiments in this chapter, we selected Altmetric.com as our source for altmetric data owing to their simple API which takes a paper's PubMed or DOI identifier as input and produces a continuous 'Altmetric.com score' which could be directly compared with REF Impact Score or used in regression without any further manipulation.

Altmetric.com "score" is a weighted count of online attention that a publication receives. Mentions of scientific works in different sources online such as social media sites (Facebook, Twitter etc) and news publishers (e.g. BBC or The Times) increase an article's score by a predetermined amount. The full listing of sources and weightings and an explanation of how these are aggregated can be obtained at the Altmetric Website [10]. We used Altmetric API to retrieve scores for as many of the papers linked to REF studies as we could. This search was carried out on 26 August 2016. For PubMed and arXiv ECNs this was relatively easy because Altmetric provides a REST API for retrieving scores for papers with PubMed and arXiv publication IDs. For the CiteSeer ECN data, we used an online scientific paper metadata aggregation service, CrossRef[11] to identify DOIs for each of the papers which were then passed to the Altmetric.com API to obtain scores for these papers. Unfortunately a large number of the papers under investigation (approximately 40%) had no Altmetric.com score at all (the API endpoint returned a 404 indicating missing paper profile).

4.5 Comparisons of Academic Impact Metrics vs REF Impact Score

Below we plot REF Impact against Mean Normalised Citation Score, h-Index calculated wrt. author and wrt. impact case study and Altmetric.com score against REF Impact score for the top 5 UoAs as identified in Section 4.3. These UoAs are 'Computer Science and Informatics', 'Mathematical Sciences', 'Clinical Medicine', 'Allied Health Professions, Dentistry, Nursing and Pharmacy' and 'Physics'. We provide the full data set for experimentation and further analysis in digital format via Figshare[12]. In each case we also calculate Pearson correlation coefficient between the two metrics to assess statistical correlation. We deliberately omit Journal Impact Factor(JIF) from our study because it is only available at

[10] https://help.altmetric.com/
[11] http://www.crossref.org/
[12] https://figshare.com/s/751679e8993a7fe2c5d8

journal level and cannot be meaningfully calculated at granularities comparable with REF Impact score.

4.5.1 Mean Normalised Citation Score versus REF Impact Score

Our first experiment examines the relationship between MNCS and REF case study impact score. MNCS allows us to compare how many citations papers attached to REF impact studies receive whilst normalising for the discipline-specific citation behaviours discussed above. MNCS is usually normalised by year and by scientific discipline (Waltman, 2015). The metadata within the ECNs typically included year of publication but scientific discipline was not typically available. Instead, the UoAs from the associated REF case studies were used as scientific disciplines for the purpose of normalisation although it assumes that all papers associated with a REF case study are from within the same UoA as that case study. MNCS scores are then further averaged across all papers attached to a given case study to facilitate reporting in terms of UoA-per-Institution as explained above.

MNCS was plotted against average REF impact score and the chart can be seen in Fig 4.4 below. Visual inspection of the graph appears to show a weak positive correlation between the two axes which is most obvious when 'All UoAs' are plotted together (bottom right). However, the Pearson coefficient of Impact Score vs MNCS for All UoAs is $r = 0.035$, suggesting that the relationship is very weak.

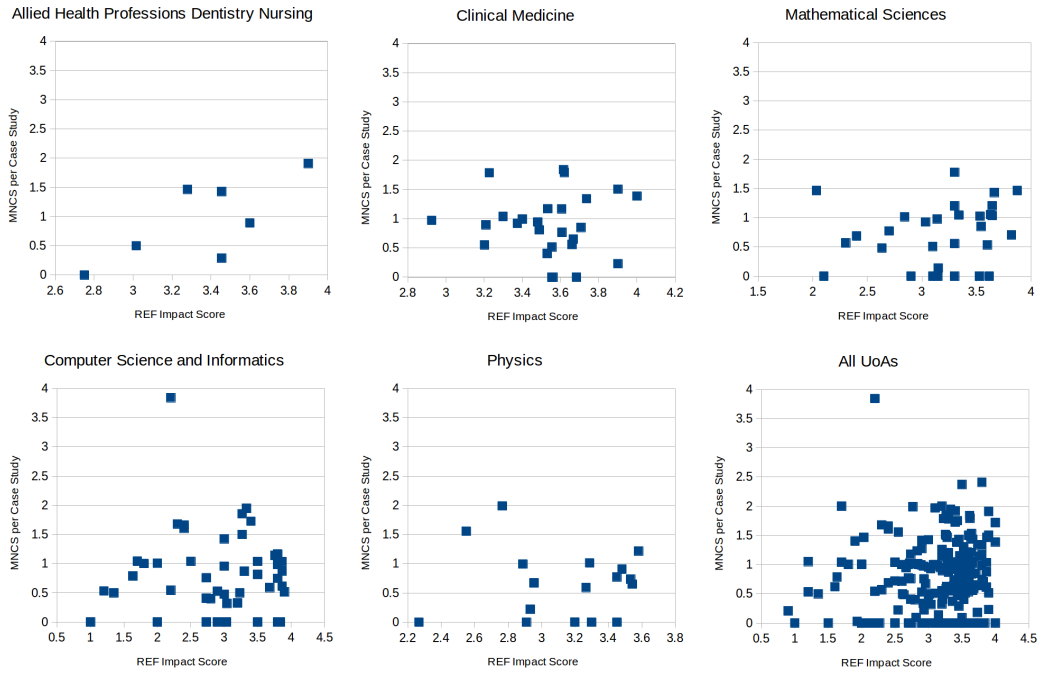


Figure 4.4: Average MNCS per REF case study vs REF case study score.

4.5.2 h-index versus REF Impact Score

h-index, as defined in section 3.1, is typically calculated with respect to an individual in order to give an indication of their publication record (limitation on interdisciplinary comparisons notwithstanding). However, h-index can also be calculated for groups of authors to facilitate aggregate comparisons at department or institution level. In this experiment, we evaluate REF impact score against per-author h-index and also calculate per-case-study h-index by aggregating all linked articles found within a given case study's bibliography and any citations thereof.

Comparison of h-index within ECN versus Commercial Citation Networks

Since the h-index values used in this study are generated from the open access data within our ECNs, it is likely that some citations of papers within the ECN from papers published in closed, commercial journals, are missing. This would artificially lower observed h-index values within our ECNs. In order to understand the extent of the missing data, we calculate the h-index of the most prolific authors in our study and compare these against their respective Google h-index as a probable upper limit in Table 4.2.

Author	**Google (Since 2011)**	**Google (All Time)**	**ECNs**
Ellis R. (UCL)	87	144	94
Filippenko A.(University of California, Berkeley)	91	144	72
Jennings N. R. (Imperial College London)	62	107	52
Gächter, S. (Nottingham)	45	49	51
Griffiths T. L. (University of California, Berkeley)	53	62	44
Wooldridge M. (Oxford)	47	82	39
Shawe-Taylor J. (UCL)	40	59	29
Papaloizou, J. (Cambridge)	41	71	20
Merrifield M. (University of Nottingham)	27	43	17
Pourtsidou A. (ICG Portsmouth)	8	8	5

Table 4.2: Comparison of H-indices for Authors: Google vs ECNs

Google Scholar's author profile page provides two values for h-index, the 'all time' value, which is calculated with all papers known to be authored by the person, and a 5-year rolling window, which at the time this data was collected (2016) used the same h-index calculation but on the subset of papers that the author has published since 2011.

Although there is some variation between our h-index scores and Google's, the ECN h-index scores are fairly close to the Google 'since 2011' scores and the deltas between each author also scale down respectively for both Google h-index values. We were satisfied

that despite the absolute differences between our h-index values and those of Google, the information from the citation network provides enough context to perform relative ranking and measurement within our dataset.

Author h-index vs REF Impact Score

We calculate h-index for all authors of each paper linked to a REF impact study. We then take the mean of h-indices for authors associated with these studies. We call this Average Author h-index. All authors are considered with the same weighting and importance across the corpus and no significance is given to the order of author listings on publications in this study.

Duplication of citations could artificially boost an author's h-index if it leads to citations of their work are counted multiple times (e.g. a single paper published by Jane Smith appears under Smith J. and is not merged correctly). However, in order for duplicate papers to significantly impact an author's h-index, any paper contributing to h, the author's current h-index, would have to have at least $h + 1$ citations (contributed through novel and duplicate links). Since duplication is very limited within our ECNs (as discussed in section 4.3.1), we believe that the probability of this happening at a scale significant enough to noticeably alter the correlation between Author h-index and REF impact study is very low.

The plot of Average Author h-index vs REF impact score is shown in Fig 4.5. There is no visible correlation between h-index and REF Impact Score and this is further validated by a Pearson coefficient of $r = -0.005$ on this dataset.

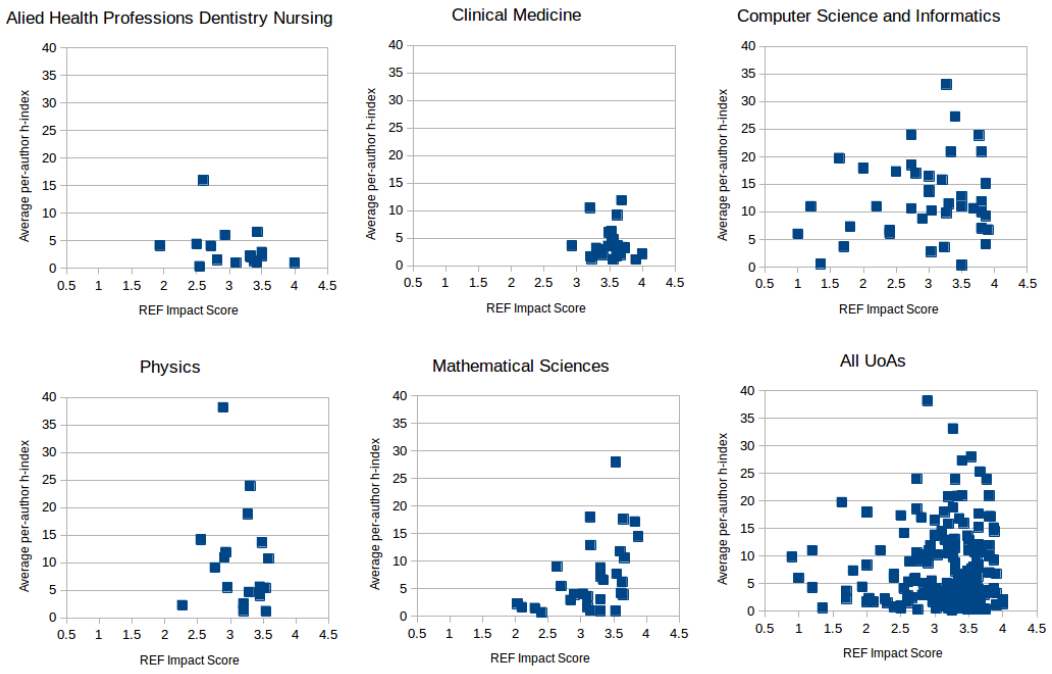


Figure 4.5: Average per-author h-index vs Average REF Impact Score

Unlike MNCS above which measures the citations of works directly attached to each case study, Average Author h-index is more reflective of each contributing author's historical publication record and citations. Therefore, it is unsurprising that the association between average author h-index and REF Impact Score is weak. For example, academics at the start of their careers with limited publication histories could collaborate with prestigious teams on highly impactful REF case studies or vice versa. This comparison also breaks down for interdisciplinary authors whose h-index is normal for their primary field of interest but an outlier in the UoA that the impact case study was submitted to.

Per-Case Study h-index versus REF Impact Score

Here, we calculate h-index per REF impact case study using papers linked to each study that are found in the ECNs. This metric more succinctly encapsulates the academic impact of the REF case study than the per-author metric used above by taking into account only papers that are directly linked to the case study rather than providing an average of contributing authors' historical works. The duplication issue outlined above is even less likely here since publication to case study linking is so strict and requires exact matching of title text.

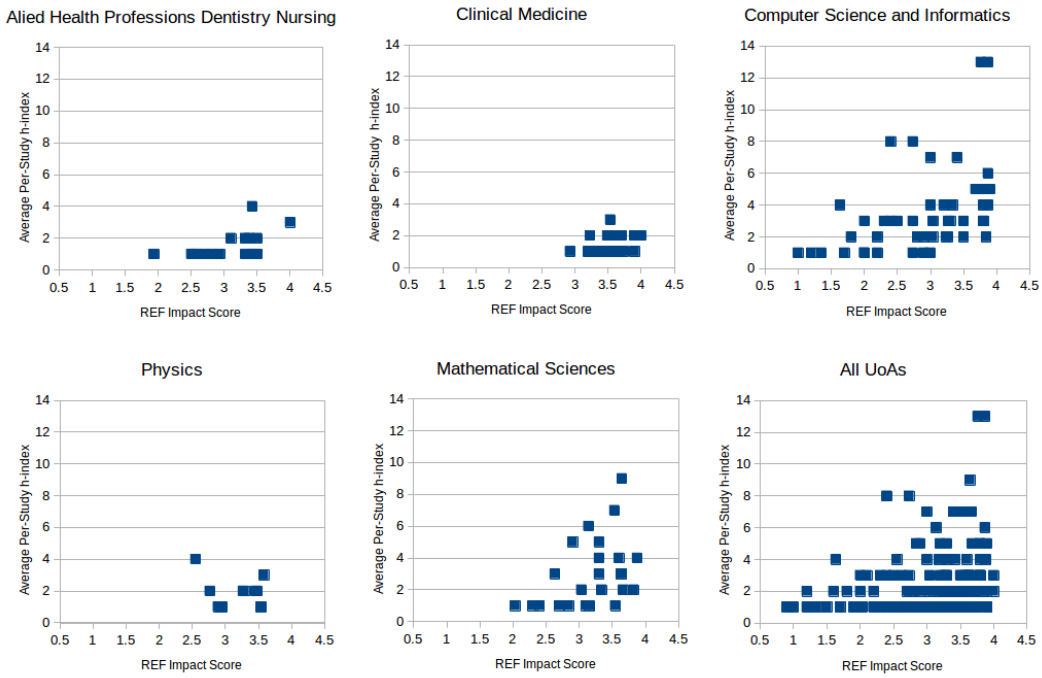


Figure 4.6: Average per-study h-index vs Average REF Impact Score

Figure 4.6 shows a plot of mean average case study h-index versus REF impact score per UoA per Institution for the top 5 UoA categories. The graph shows a more obvious relationship between the two metrics which is reflected by a stronger Pearson coefficient of $r = 0.141$. The per-case-study h-index behaves acts almost like an averaging filter over the

MNCS, reducing the overall effect of a single paper with a large number of citations on any given study.

4.5.3 Altmetric.com Score versus REF Impact Score

An average Altmetric.com score is calculated with respect to each REF Impact Case Study and plotted against REF Impact Score per UOA per Institution in Figure 4.7. One might consider Altmetric.com score closer to REF Impact Score than the citation-based metrics based on its coverage of online news and social media interactions which overlap with REF's coverage of news articles. However, surprisingly the results show that there is little in the way of correlation between Altmetric score and REF impact. The Pearson coefficient for the overall relationship is $r = -0.080$. This finding seems to support the experience of Thelwall et al. (2013). Their work investigates the correlation between social data, used by altmetric providers like Altmetric.com to calculate scores, and citation data. They found that social data coverage was inadequate for drawing any conclusions.

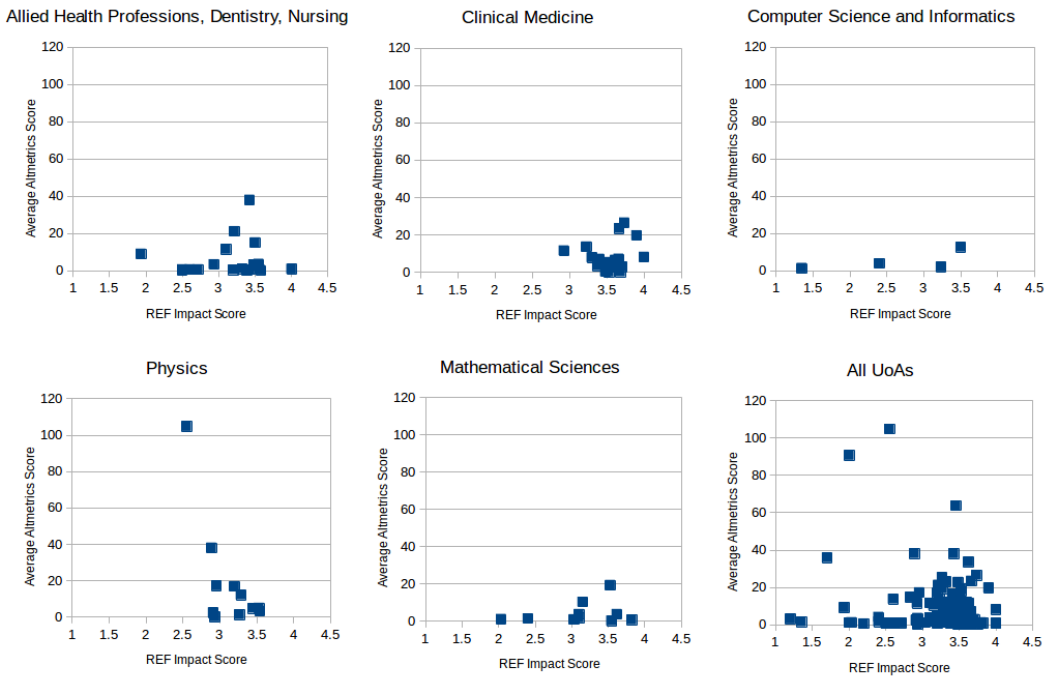


Figure 4.7: Average Altmetric Score vs Average REF Impact Score. For All UOAs, one outlier is not displayed at (2.75,218).

We found that many of the 40% of papers with no Altmetric.com score came from Computer Science, leaving only 4 samples displayed in Figure 4.7 (top right). This could suggest that although Computer Science papers are cited many times by academics (as previous graphs imply), they are not discussed as frequently on social media. A study by Costas et al. also found that Altmetrics coverage of Mathematical and Computer Science

papers tends to be much lower than disciplines like Biomedicine and Health or Natural Sciences and engineering (Costas et al., 2015). Haustein et al. (2016) suggest that papers are more likely to be tweeted if they are "...curious or funny, have potential health applications or refer to a catastrophe..." Computer science papers are often abstract or focus purely on a technique or algorithm rather than application. This could explain why computer science papers receive less online attention than biomedical and health papers that have applications that can easily be understood.

The inverse appears to be true for Physics and Allied Health Professions papers, which have some of the lower h-index and citation counts in previous graphs but have the highest Altmetric scores. Perhaps these papers address concepts that are more tangible to the public on social media. However, neither of these UoAs demonstrate any kind of correlation between Altmetric scores and REF scores.

4.5.4 Correlation Scores Summary

Pearson r coefficients for each of the UoAs in the experiments discussed in Section 4.5 above are shown in Table 4.3. We note that r coefficient values vary significantly across each UoA which is likely reflective of the discipline-specific citation and web interaction behaviors discussed above.

UoA	**Allied Health Professions, Dentistry, Nursing and Pharmacy**	**Clinical Medicine**	**Computer Science and Informatics**	**Mathematical Sciences**	**Physics**
# of Papers Linked to UoA	65	72	280	136	52
MNCS (r)	0.229	-0.172	-0.003	0.182	-0.06
Data Points	16	23	38	23	12
Author h-Index (r)	-0.094	-0.21	0.168	0.461	-0.182
Data Points	23	25	42	30	18
Case Study h-index (r)	0.178	0.139	0.418	0.347	0.023
Data Points	16	23	46	25	12
Altmetric Score (r)	0.081	0.210	0.058	0.102	-0.426
Data Points	18	20	4	10	11

Table 4.3: Pearson r coefficient scores for metrics evaluated against REF Impact score

4.6 Regression Modelling

The Pearson r coefficients calculated for academic metrics and REF Impact Scores above generally point to weak or no correlation. However, we explore whether a combination of academic metrics could provide a complementary set of features for training a model to estimate REF impact case study.

We implemented a linear regression Ordinary-Least-Squares (OLS) baseline model for predicting REF impact score for a case study using per-study h-index, per-author h-index and mean altmetrics score as features using SciKit-Learn (Pedregosa et al., 2011). We also pass in UoA as a categorical variable. The model was trained using Leave-One-Out cross-validation since the data set was too small to meaningfully divide into folds. We also used Leave-Out-One-Feature to understand which of the metrics are most influential in the model. We evaluated the performance of the model using R^2 metric as defined in Section 2.4.4 and results of this analysis can be seen in Table 4.4.

We found that, when all features are included, the predictive capability of the model was very poor ($R^2 = -0.113$). Holding out UoA features improves the score but all scores < 0 indicating poor model fit. We also tried training a Support Vector Regression (SVR) with RBF kernel to see if the aforementioned features could be separated on a hyperplane. However, the R^2 scores for this model were also poor.

Features	**R2 Score**
All Features	-0.113
Without Author H-Index	-0.102
Without Per-Case-Study H-Index	-0.062
Without Altmetric Score	-0.067
Without UoA	-0.033

Table 4.4: **Regression Model Prediction Results**

4.7 Discussion

We have shown that for our ECNs which cover an extensive proportion of online, open-access scientific publications as of 2017, there is negligible correlation between REF impact score and commonly used academic impact metrics. We note that our findings do not conflict with those of Pride and Knoth (2018) who found that REF *output* score (as opposed to REF *impact* score) strongly correlates with citation-based metric scores. These findings do support our earlier supposition that comprehensive and academic impact metrics are independent and measure different things and offers further merit to the suggestion that the academic community should broaden their measurement of impact beyond citation metrics (European Association of Science Editors, 2007; League of European Research Universities, 2015; Edwards and Roy, 2016).

A recent study by Wooldridge and King (2019) concludes that Altmetrics *do* correlate with REF Impact Score although like us, they conclude that REF Impact Score does not correlate well with citation-based measures. Whilst their results appear to partially conflict with our own, the authors posit that the difference may lie in their use of a more complex model for per-institution UOA scores as opposed to use of per-UOA mean. We also note their use of a much larger, commercial citation network to which we did not have access during our work and the recency of their work which may have allowed time for more Altmetric data to have been accumulated since the 2014 REF assessment. Furthermore, at the time of original publication, Altmetric.com coverage of the papers in our ECNs was insufficient to allow us to draw strong conclusions about statistical interactions between Altmetric.com score and REF impact score. We also remain mindful of the concerns of Banshal et al. (2019) and Ortega (2018); even if Altmetric scores correlate with or are predictive of REF Impact Score, inconsistent data coverage between scientific disciplines remains a concern.

On balance, we believe that Wooldridge and King (2019)'s findings are encouraging and provide an opportunity for future work to explore multi-modal methods that combine Altmetric data with other heterogeneous data sources that provide evidence of comprehensive impact. We discuss this further in section 10.2.2.

Our novel application of h-index to papers linked to REF case studies yields scores that correlate more strongly with REF Impact Score than other academic metrics. However, as shown in Table 4.3, the extent of this correlation strongly depends each specific upon UoA/scientific discipline. Per-case-study h-index does not appear to be predictive of comprehensive impact.

Significant time was invested into data collection and sanitation and ensuring that the ECNs have comprehensive coverage of STEM disciplines. However, only a small number of REF case studies could be associated with scientific papers in our ECNs. The free-text format of the REF impact case study bibliographies presented a significant barrier to link extraction and a large amount of text had to be discarded. It is also likely that papers referenced by REF case studies were missing from the ECNS due to having been published in closed-access journals or simply not made available via arXiv, CiteSeerX or PubMed Central.

Papers need time to accumulate citations. The REF guidelines stipulate that all supporting works must have been published in the ten year window starting in December 2003 and ending in December 2013. The ECN data dumps for Paperscape and RefSeer (CiteSeerX) were taken from 2013 and 2014 respectively. The PubMed ECN was generated in August 2016. It is likely that papers published close to the time that these dumps were generated have disproportionately low citation-based metric scores due to having had less time to accrue citations and altmetric data. However, given that we normalised for year of publication in our comparison of MNCS against REF Score, we would expect a stronger correlation of MNCS with REF Score if accrual of citations was a major limiting factor.

4.8 Conclusion

In this chapter we have explored the relationship between widely-used academic impact metrics that focus on citations and web interactions and REF Impact Score, a comprehensive impact metric that focuses on specific case studies that detail the ways in which the scientific works under evaluation affect society, the economy and policy. Our aim has been to understand to what extent academic and comprehensive scientific impact are related and, given the high complexity and cost of current generation comprehensive impact metrics, whether academic impact metrics can be used to approximate comprehensive impact.

We have shown that for a large collection of 7.4 million open-access scientific papers and 6600 REF Impact Case studies, comprehensive and academic scientific impact metrics correlate weakly or not at all, nor can comprehensive impact be approximated using academic impact data. Further work could look to increase coverage of the experimental citation networks, including newly published articles as well as closed-source publications and to improve matching of published articles with REF impact case studies which could also be expanded to include REF 2021 impact case studies. However, such work would require significant, open-ended investment into numerous commercial access to closed-source citation networks and paper metadata as well as the resolution of tackling challenges like de-duplication of metadata and improved linking between scientific papers and REF impact case studies. We leave these challenges as possible future work and instead, focus on finding new ways to measure and understand comprehensive scientific impact more directly.

REF Impact Score provides a reasonable baseline for the development of new comprehensive scientific impact metrics. Its qualitative and broad scope allow the fair and holistic consideration of a variety of evidence that scientific works under evaluation are impacting the world beyond academia. However, REF is also an expensive, centrally coordinated process, calculated infrequently and with pseudo-anonymity. Such limitations mean that REF Impact Score and other similar initiatives cannot, in their current form, be used proactively by scientists to understand and expand their real world impact, nor is it likely to be adopted widely by the academic community.

Thus, we conclude that there is a clear need for the development of efficient, cost-effective tools and metrics to help scientists to understand the the comprehensive scientific impact that their work generates and to serve as more appropriate instruments for measuring the broad impact of scientific work as advocated for by European Association of Science Editors (2007), League of European Research Universities (2015) and Edwards and Roy (2016). In the next chapter, using REF as our starting point, we begin to explore ways to combine and link real-world datasets with related scientific papers using big-data processing, machine learning and natural language processing techniques. We then use this data to characterise the comprehensive impact of scientific works automatically.

5

Empirical Methods for Linking Scientific Papers to Evidence of their Comprehensive Impact

"Extraordinary claims require extraordinary evidence"

Carl Sagan

5.1 Introduction

There are as many ways to measure comprehensive scientific impact as there are to produce it, from the number of lives saved by a groundbreaking treatment to the number of laws and policies influenced by a meta-analysis on passive smoking to the number of novel high-density batteries sold by its inventor's spin-out company. This makes comprehensive impact paradoxically easy to measure and yet very hard to compare broadly. There are many outputs that can be *measured* but very few that are generalisable or freely and widely available across disciplines, institutions or countries; even if the private company selling high density batteries agrees to share their sales figures, we can't directly quantify that in terms of lives saved or government policies influenced.

As discussed in Chapter 3, existing comprehensive impact metrics typically avoid this set of problems by limiting their scope to a particular discipline (Sørensen et al., 2022), type of impact (e.g. financial (Lane and Bertuzzi, 2010)) or geography (REF 2014, 2011). In the latter case, the focus on qualitative rather than quantitative outcomes in REF impact case studies enables broader comparison of distinct outcomes at the cost of an expensive and labour-intensive process and even then, comparison is still siloed into units of impact as discussed in Chapter 4.

To successfully measure comprehensive impact broadly and at scale across disciplines, geographies and types of impact requires supporting evidence that is ubiquitous across these categories and a method for comparing and measuring such evidence. This is likely a significant reason for the continued popularity of citation-based academic impact metrics within academia; citations and publications are a commonality shared by all academic disciplines internationally and counting them is simple and intuitive despite the limited scope of such methods.

In this chapter tackle research question **RQ2** (section 1.2) and Challenge **CH3** (section 1.3.3). We briefly discuss different types of comprehensive impact and datasets that provide evidence of such impact. After assessing the pros and cons of each, we focus specifically on news articles that describe scientific work and assemble a corpus of news ar-

ticles linked to scientific publications (further discussion of other types of impact and their possible uses can be found in our future work section in Chapter 10). Finally, we explore the statistical relationship between scientific news and comprehensive impact by linking pairs of news articles and scientific publications to REF Impact Case Studies and the ECNs from Chapter 4.

This chapter is partially based on our publication *'HarriGT: Linking News Articles to Scientific Literature'*. which was published in Proceedings Of the 56th Annual Meeting Of the Association for Computational Linguistics-System Demonstrations, 19–24, 2018 and can be accessed at `http://aclweb.org/anthology/P18-4004`.

Our specific contributions are:

- A web-based tool for semi-supervised matching of news articles to scientific papers
- A novel algorithm for ranking candidate scientific papers to facilitate efficient manual matching to news articles that discuss them
- The HarriGT corpus: A set of 5903 news articles linked to 9891 linked scientific abstracts and 1086 full-text scientific papers
- A subset of the HarriGT corpus linking 140 news articles and 108 scientific papers to 103 REF Impact case studies.
- We show that the REF impact case studies that have one or more linked news articles are likely to have been awarded a higher impact score than case studies that are not linked to news articles.

5.2 Evidence of Comprehensive Impact

Evidence of comprehensive impact can take many forms. The subset of such evidence that is widely and freely available and common across a large number of disciplines and geographies is small. Here we briefly discuss a number of such sources.

5.2.1 Commercial & Financial Impact

Many academic works created within universities lead to the creation of spin-off companies and startups which go on to achieve commercial success. Likewise, scientific works published by private research & development departments often lead to improved commercial outcomes for their parent company. Although evidence that pressure to commercialise research may have adverse impacts on research environments (Caulfield and Ogbogu, 2015), both Lane and Bertuzzi (2010) and REF (Parks et al., 2018; HEFCE, 2019) propose that jobs created, money earned and patents filed as a result of scientific work could serve as proxy measures for the commercial impact of science. However there are a number of challenges regarding the availability, coverage and interpretation of these sorts of data.

Companies' financial information may only be available in limited quantities and geographies. All companies in the UK, privately or publicly owned, are obliged to file public notices of ownership upon incorporation as well as financial statements on an annual basis in which they disclose their profits, losses and any outstanding debts. These documents are then made public record and available for free download from the UK Government [1]. However, the UK's policy here is an exception and such data does not exist for most other countries. Therefore, in the majority of territories where this data does not exist as a public record it would need to be volunteered by relevant commercial entities which prevents the widespread use of financials as an impact metric.

Even where financial data is available, statutory financial records are usually high level documents that are unlikely to provide a detailed breakdown of revenue and understanding the impact that a scientific work has on a company's financial or employment figures may also be difficult or impossible. The filings of a purpose-build spin-out company may present a clear picture but in larger companies, a significant number of factors can affect financial performance; the commercialisation of a piece of research may be one of a number of strategies being employed by the company at any given time. Furthermore, authors within such companies are unlikely to be directly responsible for its commercialization; most often, ownership of privately developed intellectual property is retained by the employer, even if the original author leaves the organisation.

Patents are documents that disclose the technical details behind a novel product which are made public record in exchange for a time-limited period of legal exclusivity to make, use or sell the disclosed product granted to their author. Patents often directly cite supporting research allowing automatic detection of links between patents and the research that led to them being granted(Gerrero-Bote et al., 2019). Unfortunately, the significant geographical and disciplinary variation in the numbers of patent filed (Sung et al., 2014; European Commission. Joint Research Centre., 2017) limit the utility of patent data as a universal indicator of commercial impact. Furthermore patents are expensive and can take many years to file and are therefore seen as a poor investment in industries, like computer science and machine learning, where technologies become obsolete very quickly. The cost of filing patents also serves as a barrier to entry, further excluding researchers who are unable to afford to patent their work even if they want to.

In summary, commercial financial and patent data is sparse, fragmented and paints a limited picture of the impact that science has on the global economy. Available commercial data may provide partial information about the economic impact of research but needs to be used in combination with other data to provide holistic coverage of comprehensive impact.

[1] `https://www.api.gov.uk/ch/companies-house/`

5.2.2 Impact on Government Policy

Government policy is another important area that scientific research can impact. Governments rely on research publications to help them to make sensible decisions about the health and productivity of their citizens, directly affecting large numbers of people in the process. Notable examples in the UK are the introduction of a national minimum wage in 1998, the implementation of a ban on smoking in public places in 2007 and the codification into law of The Climate Change Act in 2008 (Rutter et al., 2012). The UK Government was also particularly reliant upon the quick dissemination and processing of scientific findings for ministers during the recent COVID-19 pandemic (Cairney, 2021). We, like REF (HEFCE, 2012, 2019), propose that scientific works which influence policy either directly or indirectly, during the course of its implementation, have produced impact. We considered a number of data sources for extraction of such impact.

A number of governments are beginning to adopt open data and open governance policies, making available information about how they operate for free online[2]. In the UK, data sources like the Hansard Records, an archive of all parliamentary debates for the last 200+ years[3] and POSTnotes [4], impartial lay-briefings on scientific topics given to ministers before debates, provide ample opportunity for text mining to identify evidence of scientific impact on policy. Commercial resources like Overton [5] which provides a structured database enumerating citations of scientific works made from a variety of policy documents (Szomszor and Adie, 2022) may also provide insight into research's political impact.

The major limitation of using government policy as a universal indicator of comprehensive impact is its exclusivity. A large number of scientific works lead to tangible commercial and other applied outcomes. However, most do not end up affecting policy unless they affect or bring new insight or understanding to societal issues. For example, a new method for manufacturing more efficient solar panels is unlikely to lead to new legislation; however the new solar panels could still impactful in terms of commercial success and popularity. Like commercial and financial data (section 5.2.1 above), open policy data is also fragmented with information spread across multiple data repositories for each country. Therefore evidence of impact on policy would also need to be used in combination with other types of evidence to provide a fair and holistic view of comprehensive impact.

5.2.3 Scientific Journalism and Comprehensive Impact

Scientific journalism is one of the primary ways that the general public interact with science. News about scientific work serves an important form of impact in its own right in its ability to raise awareness of scientific issues and is also recognised in the REF guidelines (HEFCE,

[2] UK: `https://data.gov.uk/`, US: `https://data.gov/`, EU: `https://data.europa.eu/`
[3] `https://api.parliament.uk/historic-hansard/api`
[4] `https://post.parliament.uk/`
[5] `https://www.overton.io/`

2012, 2019) as a contributing form of impact for case study submissions.

As well as raising awareness of scientific work, scientific journalism provides evidence of many types of comprehensive impact including those discussed in sections 5.2.2 and 5.2.1 as well as in many other areas such as health, the environment and culture. For example "Scientists Create Tomatoes Genetically Edited to Bolster Vitamin D Levels" discusses specific work involving genetically enhanced tomato plants and links them to a new law being proposed in the UK permitting the widespread use of genetically engineered food crops (Geddes, 2022). "R2-D-Chew: Robot Chef Imitates Human Eating Process to Create Tastier Food" (Abdul, 2022) outlines advances in automated sensing of food which it links to advances in automated food preparation and a domestic appliances manufacturing company, Beko. "Robot-Assisted Surgery Can Cut Blood Clot Risk and Speed Recovery, Study Finds" outlines a specific study which showed that robotic surgery for certain conditions can lead to better outcomes than human surgery (Gregory, 2022). In each case, these articles mention specific scientific works and directly summarise the impact that those works have generated. Thus, it is usually possible to understand and appreciate the impact that a scientific work discussed in a news article has had without needing to rely upon third-party proprietary datasets.

Scientific news articles are also plentiful and ubiquitous; most countries have a national press who publish science and technology articles. The majority of these articles are made available online and free access to them via news websites is usually provided. Such articles can then be gathered using web scraping technologies which can extract the full text content from a given news article for downstream analysis. There are also a number of web archives such as CommonCrawl[6] and the UK Web Archive (JISC and the Internet Archive, 2013) that provide free, historic access to web content that has already been collected. Although web scraping was previously considered controversial, recent lawsuits in the United States have concluded that web scraping is legal on publically-accessible websites[7]. Furthermore, processing of scraped, copyrighted material, such as news articles, is usually permitted in countries like the United Kingdom[8] and countries within the European Union[9] which provide legal copyright exceptions for application of text and data mining technologies.

Journalists tend to publish articles that discuss scientific works that they consider newsworthy based on the work's scope, scale and novelty, timeliness, cultural relevance and how relatable the work is (Molek-Kozakowska, 2017). Press Releases are written briefings provided by scientists, often with help from university press officers, that outline the newsworthiness of their work, often by using some of the aspects such as timeliness and cultural

[6] https://commoncrawl.org/

[7] https://cdn.ca9.uscourts.gov/datastore/opinions/2022/04/18/17-16783.pdf

[8] https://www.gov.uk/guidance/exceptions-to-copyright

[9] https://eur-lex.europa.eu/eli/dir/2019/790/oj

relevance as a hook to interest journalists (Fuoco, 2021). For example, by providing a timely link between a study showing the need to frequently wash hands due to contamination from harmful flame retardants found at participants' homes and the COVID pandemic during which early advice was also to wash your hands to avoid infection *(ibid.)* Recent work by MacLaughlin et al. (2018) found that scientific journalists are much more likely to publish news about scientific work that has an associated press release.

Clearly not all scientific works can feature in the news. However, many breakthrough studies that are not mature enough to directly change policy or become a billion pound startup idea but which lay the foundations for further-reaching works, are considered interesting and newsworthy enough to report on. Additionally, the policy-changing, money-making works also tend to be newsworthy and do often feature in publications, making scientific news a rich, broad resource for understanding comprehensive impact. For the remainder of this work, we primarily focus on the link between scientific news articles and comprehensive impact. Discussion of other sources of evidence for comprehensive impact explored in this section and possible integrations thereof can be found in Chapter 10.

5.3 NLP and IR Approaches for Linking News Articles and the Scientific Papers They Discuss

We aim to evaluate the effect that links to news articles, or a lack thereof, can have on a REF impact case study's score. However, before we are able to do this, we require a practical way to find news articles that discuss the scientific works that we previously linked to REF case studies in Chapter 4.

Within our data model, visualised in Figure 5.1, news articles may have a direct outbound link to a scientific paper via a DOI or by referring to the work within the body of the work. Alternatively, links between scientific works and news articles may be inferred by parsing REF case studies which provide both citations to supporting scientific works and references to news articles that discuss the work.

5.3.1 Process Overview

We develop a multi-step article linking process visualised in Figure 5.2 that allows us to explicitly link corresponding News Articles and Scientific Papers to each other and to REF Case studies that mention them using the data model outlined above.

The process takes as inputs three datasets. The first is a large multi-terabyte historical web archive containing a large number of news articles which we describe in more detail in section 5.3.2). We additionally re-use both the REF case studies and citation networks identified in Chapter 4.

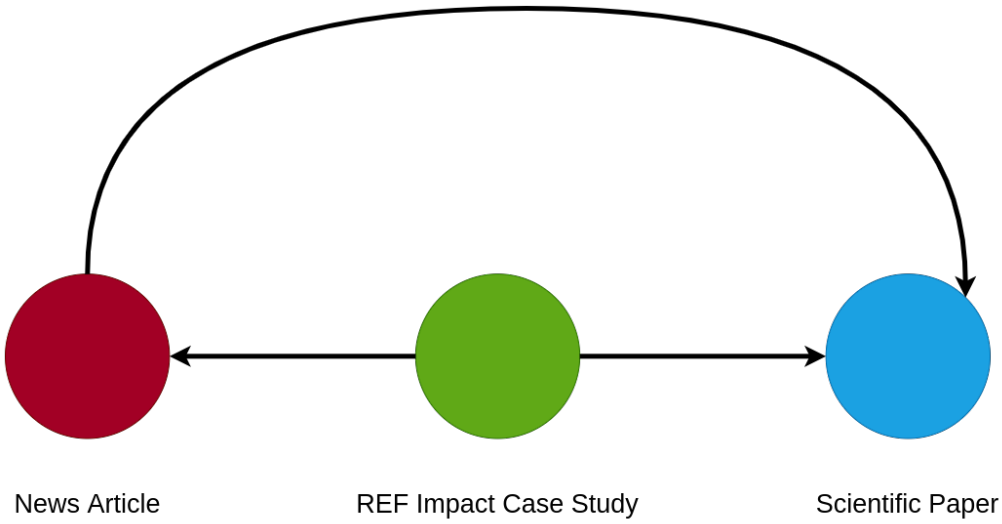


Figure 5.1: A visual representation of possible relationships between news articles, REF impact case studies and scientific papers.

The process begins with the identification of news articles within the web archive (section 5.3.2). We then carry out automated processes that identify explicit references to news articles from within REF case studies (section 5.3.3) and explicit references to scientific papers from within the news articles (section 5.3.5). Scientific news articles that cannot be automatically linked to REF case studies or scientific papers are submitted for human verification in our HarriGT annotation tool (section 5.3.6).

HarriGT, uses semi-automated processes to filter out irrelevant news articles that have a low probability of containing a link to scientific works (section 5.3.6.1) and automatically propose scientific papers that are likely to be linked to each article (section 5.3.9). Annotators have two manual touchpoints with the process. Firstly, they may override the article filter, marking news articles as irrelevant if they do not discuss scientific work or vice versa. Secondly, they review relevant news articles and the automatically proposed candidate scientific papers, formalising links between them where appropriate.

Finally, news articles, scientific papers and REF case studies are assembled into a single linked data set and missing links are inferred and made explicit where appropriate.

5.3.2 Identifying and Extracting News Articles From the UK Web Archive

In order to build a comprehensive corpus of news articles, we worked with the UK Web Archive (JISC and the Internet Archive, 2013), a comprehensive collection of the *.uk* top-level domain between 1996 and 2013. Content is stored in ARC and Web Archive (WARC) compressed format (Technical Committee ISO/TC 46, 2017) and indexed separately in CDX files which contain metadata about every URL that was scraped and a pointer to the related content within the WARC structure. The JISC Web Archive is approximately

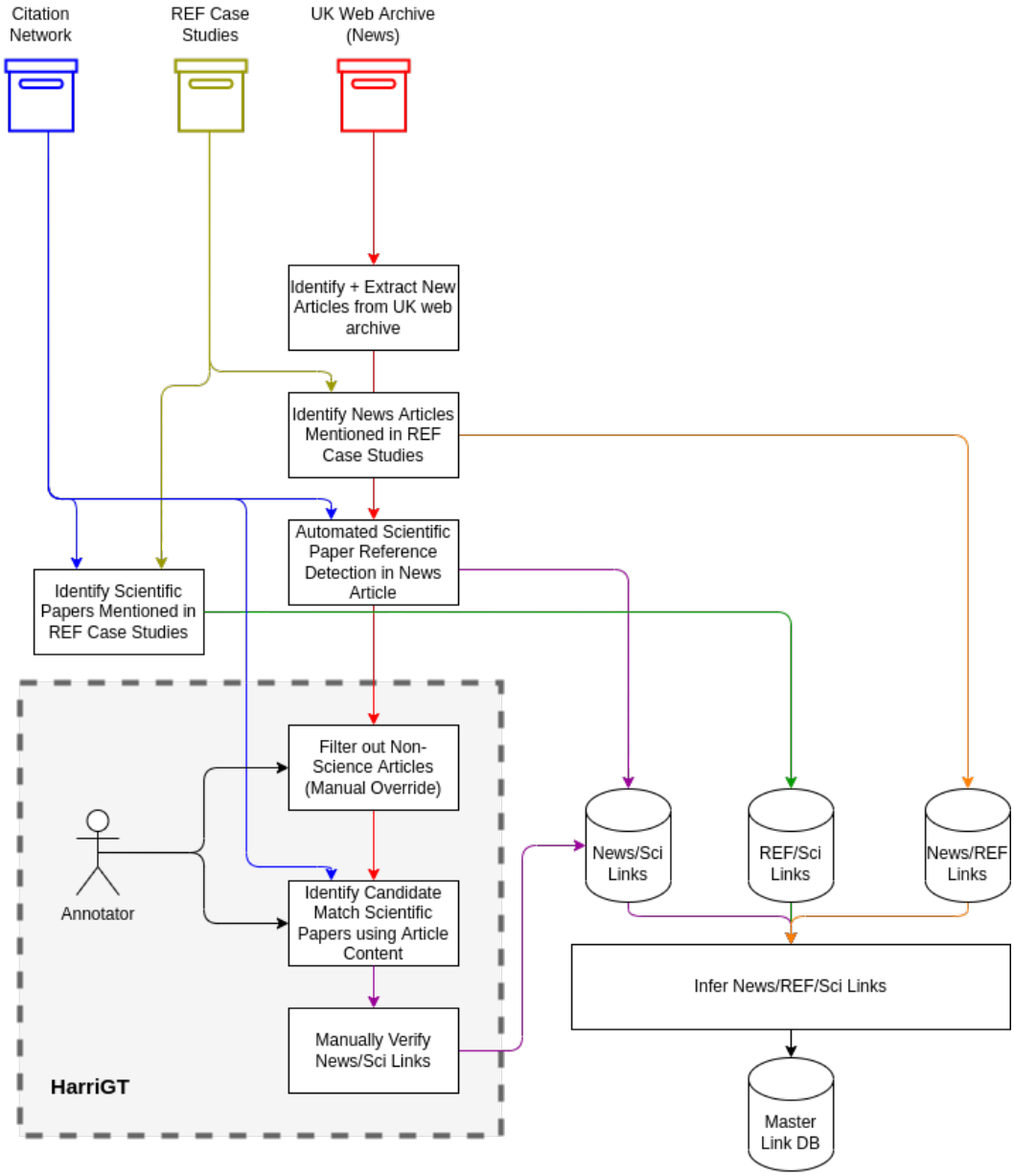


Figure 5.2: A flowchart outlining the process we have developed for linking news articles, scientific papers and REF Case Studies. Each sub-process (rectangular box) is described in depth in corresponding sections in this chapter. The shaded area shows which parts of the process reside within our HarriGT annotation tool.

62 Terabytes in size, and many websites with a *.uk* web address are irrelevant. Therefore, identifying and filtering relevant content was a primary concern.

Content collected before 2011 is stored in ARC files and following a software change at the UK Web Archive, content after 2011 is stored in WARC files. The latter format is easier to work with since it additionally stores HTTP protocol metadata and header information as well as the HTML payload itself, allowing users to more efficiently scan

for and quickly and reliably discard records that returned HTTP errors (e.g. 404 content not found) and non-text data (e.g. images, PDFs, audio clips). In order to make use of the easier-to-use WARC formatted dataset, we initially decided to restrict our investigation to WARC archives retrieved between 2011 and 2013. It should be noted that this date range corresponds to when the web content was *collected* rather than when it was originally *published*. Since most mainstream news publishers do not rescind old content, a large number of web pages published before 2011 are included in these collections archives, including content dating back as far as 10 years prior to REF submissions in 2013.

We compiled a list of web addresses for local and national UK news outlets via a Wikipedia article[10] in order to reduce the number of hostnames that our tool should inspect down to 205. The archive index files also provided metadata about the type of each WARC entry and whether the original scrape was successful or not (e.g. whether the URL was invalid). This brought down the total number of WARC entries to be examined to approximately 11.5 million. Requests to the BLOB store hosting the web archive were optimised through a script that identified batches of URLS archived in the same BLOB.

The contents of the archives were typically HTML and thus we needed to extract the title and body of each news story. HTML layouts can vary significantly between sites but news articles follow a typical layout and thus extraction of content fields can be carried out using rules and patterns rather than a machine learning approach. For our purposes we found that the open source library newspaper[11] was highly effective and gave us access to an article's title, authors, publication date and other metadata.

During the process we realised that some news articles had been duplicated in the archive. This can occur when a web crawler retrieves a URL that has been generated erroneously by the scraper script or the website being scraped. This can lead to multiple links to the same content. Examples include incorrectly appending search keywords, pagination information and other parameters into URLs that do not require these parameters. To get around this problem, we introduced a hashing system, taking the SHA256 (U.S. Department of Commerce and National Institute of Standards and Technology, 2012) hash of the title body text from each article and only accepting new content if its hash is not already known.

Most online news publishers categorise their content by subject matter. However, we found that limiting our analysis to just science and technology sections of the newspapers led to exclusion of relevant material. A second approach was to only accept articles that pass two high-level keyword filters. The first, simpler check is to see whether or not an article contains one or more keywords: *science, scientist, professor, doctor, academic, journal, research, publish, report*. We deliberately chose these keywords as a simplistic filter to reduce the amount of current affairs/celebrity gossip news that was initially accepted into

[10] `https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_Kingdom`
[11] `http://newspaper.readthedocs.io/en/latest/`

our system.

For the second of our filters, we ran a Named Entity Recognition (NER) algorithm (Honnibal and Montani, 2017) that provided multi-word expression identification and classification for names, locations and geo-political entities. From the results of the NER execution, we only accepted articles with at least one organisation containing *University, College* or *Institute*.

The final step in the pre-processing pipeline was the identification of each article's publication date. Publication date is one of the most salient features in our paper candidate scoring algorithm discussed below. Recent digital news articles give their date of publication in their HTML metadata. However, for many of the old articles in the web archive, this information was not present. For articles with no known publication date, we first attempted to retrieve the same URL from the live internet where much of the original content is still available but with updated layouts and metadata. If the content can't be found, we used a set of regular expressions (found within the newspaper library mentioned above) to try and find the date in the article HTML.

We used Apache Spark (Zaharia et al., 2016) to process the the articles in batches on a large-scale cluster. Our process yielded a collection of 1.3 million quality press and tabloid digital newspaper articles from UK outlets including BBC News, The Guardian, The Telegraph, The Daily Express, The Independent and The Daily Mail.

5.3.3 Identifying News Articles mentioned in REF Case Studies

The linking process between REF case studies and news articles is summarised in Figure 5.3. As discussed in Chapter 4, REF impact case studies contain a free-text "Supporting Works" section which we previously parsed to identify links between case studies and scientific works. Case studies also have a free-text "Details of the Impact" section in which academics enumerate the comprehensive impact that their work has had and a "Sources to Corroborate The Impact" section on which authors provide references and links to external sources, including news articles providing corroborating evidence.

We built SOLR[12] indices of these fields for all 2014 REF impact case studies. We also indexed all news article URLs captured during the processing of the web archive (as discussed above in section 5.3.2). We search our impact case study index for URLs containing the domain name of any of the popular newspaper websites from our list of news outlets (also described in section 5.3.2). Any URLs with matching domains are then used to query the index of news articles to see if they match.

Of the 6640 REF case studies, 633 (9.5%) case studies contain links to one or more news article hosted at one of the included newspaper websites. Most of these case studies (445) link to one news article with 128 case studies linking to two news articles and 50 linking to more than two. This meant that, in order to build a larger data set, it was also

[12] `https://solr.apache.org/`

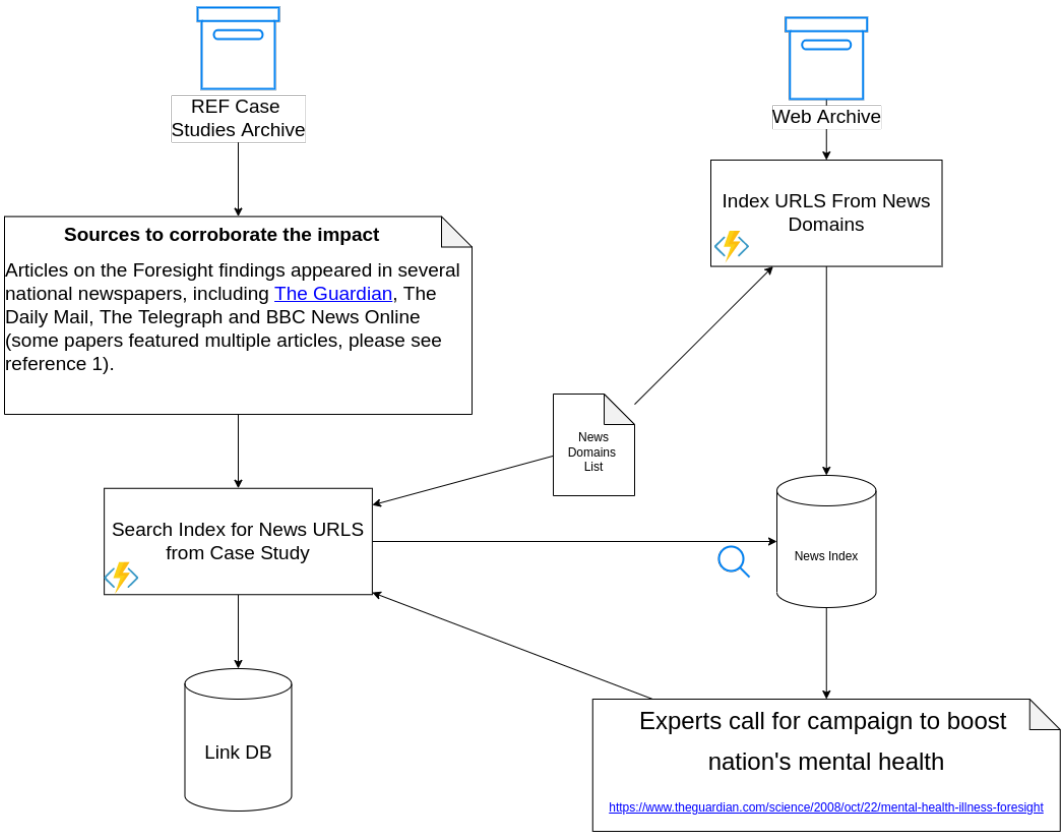


Figure 5.3: A visual summary of the REF impact case study to news article matching process. Historical news articles extracted from the news archive discussed in section 5.3.2 are indexed and then case studies are systematically searched for hyperlinks to popular news sites. News links within case studies are used to query the news index and exact matches with indexed articles are stored as links in the Link DB.

important to be able to find links from news articles back to REF case studies via scientific papers (as shown in Fig 5.1).

5.3.4 An Introduction to Extracting Citations from News Articles

Citation extraction from news articles reporting on scientific topics remains a challenging and relatively unexplored task. There are no conventions, formal or informal, for citing a scientific work in a news article. On the other hand, parsing and understanding citations between scientific works is a domain that has seen a lot of attention from academia in recent years and provides partial solutions that can be applied to news articles.

Citations in scientific papers are relatively well structured and formulaic. As such, pattern-based extraction mechanisms have been found to yield good citation extraction results (Councill et al., 2008). Disambiguation of the scientific work and authors to which a citation refers can be a much more challenging task. This especially applies in cases where

authors have ambiguous names (e.g. J. Smith). One approach is to assign scientific works and authors unique identifiers such that there is no ambiguity in cited works (DOI and ORCID respectively) (Paskin, 2015; Butler, 2012). A more pragmatic approach is needed to disambiguate publications and authors for which no DOI or ORCID ID have been assigned. Huang and Ertekin (2006) present a method for disambiguation of authors using a learned distance metric that takes into account author's known names, affiliations and venues that they typically publish at. Similar approaches have led to the creation of citation networks that store relationships between huge volumes of scientific works. Networks such as CiteSeerX (Wu et al., 2015), Microsoft Academic Knowledge Graph[13] and Scopus[14] provide external access via APIs for research and application development purposes.

Drawing inspiration from these approaches, we automate the generation of links in cases where journalists do provide structured citations and/or DOIs and URLs for the documents that they discuss (Section 5.3.5) and when such information is unavailable, we use a fuzzy matching and scoring strategy, like that of Huang and Ertekin (2006), to propose links for human approval via our HarriGT annotation tool (Section 5.3.6).

5.3.5 Automated Scientific Paper Reference Detection in News Articles

Most scientific journalists will directly cite scientific work either using a traditional citation strings that can be parsed using methods like Freecite (as discussed in Chapter 4). Alternatively, they may providing a hyperlink leading to the digital location of the work which can be followed or give an exact DOI which can be used as a unique identifier for the work. These helpful behaviours are common in quality press papers such as The Guardian and higher quality online news portals like the BBC.

In these cases, we were able to use an automated approach to record links between documents. Regular Expressions were used to match DOI strings in newspaper text. For extracting hyperlink references, we developed a web scraping script to follow the hyperlinks and extract DOIs from the HTML metadata tags embedded in the target web pages. These metadata tags are mostly standardised across academic journal publishers' websites allowing full automation of our scraper scripts.

5.3.6 HarriGT: Semi-Automated Linking of Ambiguous Links to Scientific Papers from News Articles

Some scientific journalists tend to make passing or implicit references to scientific work, normally omitting DOIs and URLS and usually in the format "Researchers at $< Institution >$ have published a study in $< journal >$". Occasionally, journalists completely omit all key information about who funded or even carried out a given study from their reports making

[13] `https://makg.org/`
[14] `https://www.scopus.com/`

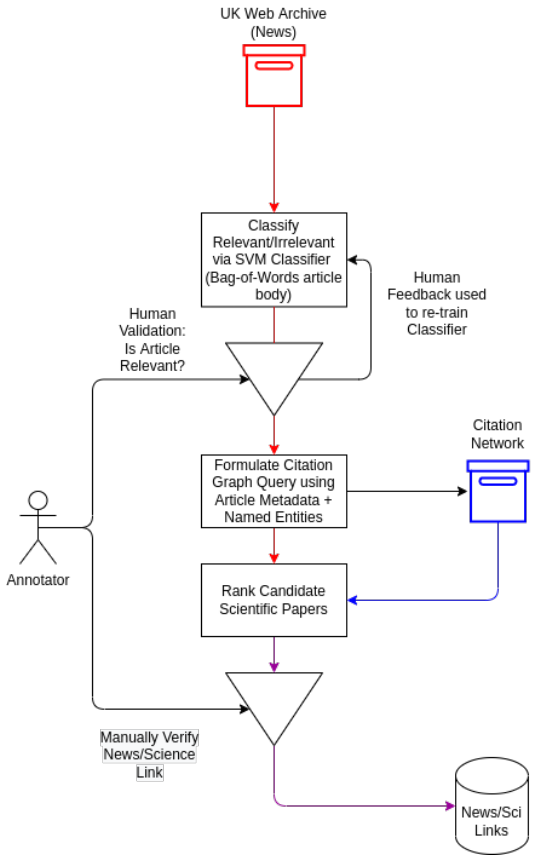


Figure 5.4: An expanded view of the annotation process diagram for HarriGT.

identification of the work very difficult (Bubela et al., 2009). Furthermore, journalists often solicit the opinions of academics who work in the same field as but were not directly involved with the subject of the article which can further confuse attempts to automate citation extraction (Conrad, 1999). These behaviours are more common in tabloid format papers such as The Daily Express and The Daily Mail.

To help us to identify and confirm links for these ambiguous matches, we developed HarriGT, a tool that combines an automated NLP-based matching pipeline with human-in-the-loop supervision in order to quickly match news articles to scientific papers where only partial information is available. Figure 5.4 outlines the process that HarriGT uses to generate matches between news articles. Articles from the web archive collection are stored in a database and labelled using a spam classifier. Information from the non-spam news articles is then used to search external academic APIs and citation networks for candidate scientific papers that could be linked to the news article.

A human annotator can then view the news article and candidate scientific papers via a web interface (shown in Figure 5.5) and decide whether to link them. The annotator can also review the spam model results and update them accordingly to correct misclassification. The spam classification model and scientific paper matcher components as

well as the citation graph integrations shown in the diagram are discussed in more detail below.

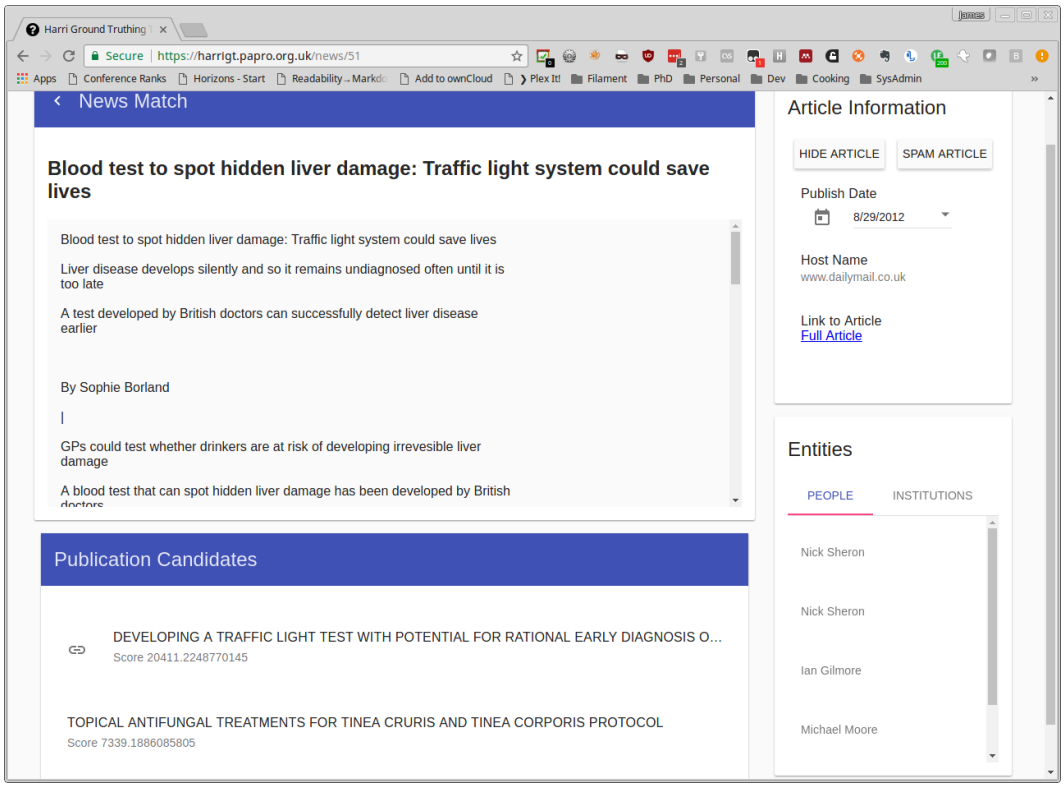


Figure 5.5: HarriGT Web UI shows a news article in the top pane and a selection of candidate matches to scientific papers in the bottom pane that the annotator can review and formally link to the news article. Helpful metadata about the news article that help the annotator identify authors, institutions and publication dates for the research are shown in a side-bar on the right.

5.3.6.1 HarriGT Relevant/Irrelevant Classification Model

Although our keyword filter during pre-processing removes a large number of general interest articles that do not discuss scientific work, there are still a number of articles that pass this initial screening that are off topic. For example, a celebrity article may mention that they went to a particular university or a crime article may involve a doctor or scientist. On the other hand, a **relevant** article should focus on one or multiple scientific advancements e.g. discovery of a new fossil, invention of an improved industrial process, successful trial of a medical treatment. Whilst such an article may include a celebrity interest story, e.g. a quote from an actor who is personally affected by a medical advancement, coverage of the scientific advancement and the investigators involve should remain its primary purpose.

Identifying whether an article is relevant or not can be very time-consuming for annotators, who may spend several minutes reading significant parts of an article and even

Model Type	Accuracy	F1-Score
SVM	0.94	0.94
Naive Bayes	0.82	0.86

Table 5.1: Micro-averaged Results from Relevant/Irrelevant Models. Irrelevant Articles: 2085, Relevant Articles: 840

start looking through paper matches before realising that there are no relevant academic links to the story. Hiding irrelevant articles accelerates annotation by reducing the number of articles that do not yield any links to scientific work that the human annotator must review. However, hiding these articles is also a non-trivial goal since we must also avoid hiding too many relevant articles which could hinder the quality of our annotated dataset. We address this issue by including a machine learned "relevant/irrelevant" classifier model into HarriGT. Within the user interface, news articles can be marked as **irrelevant** by the human annotator if they contain little relevant scientific content.

Articles are placed in separate tabs in the user interface depending upon which class they are assigned. Having been briefed about the definitions of relevant and irrelevant as discussed above, users of the HarriGT tool are encouraged to re-label articles that are incorrectly classified and periodically review the irrelevant tab. A set of 50 relevant and 50 irrelevant news articles were collected using the HarriGT interface in order to provide an initial labelled dataset for the model. Subsequently, the model was re-trained using new examples from the **irrelevant** and **link** categories periodically as the users continued to correctly tag articles. This ongoing review process ensures that model performance is repeated over time and helps annotators to minimise time spent reviewing irrelevant articles.

We initially trained two machine learning models to address the problem, a Naive Bayes classifier and a Support Vector Machine. We used Grid Search to identify the best training hyper-parameters for feature extraction and the models. The optimal feature hyper-parameters were found to be unigram and bigram bag-of-words features with TF-IDF weighting, maximum document frequency of 75% and a maximum vocabulary size of 10,000. We found that an SVM with a linear kernel and $C = 1$ produced the best results and used this model in the live system. Table 5.1 shows our model results after 4 iterations of training and use.

Given the size of the corpus, the hardware environment that the model was required to support and the positive results from the SVM model, we decided not to explore deep learning approaches to relevant/not-relevant filtering.

This model facilitates the general classification of news articles as relevant or irrelevant in the context of science in general. However, news articles are not yet linked to specific papers at this stage.

5.3.7 Citation Graph Integrations and Candidate Retrieval

In order to provide candidate scientific works for each newspaper article, we required integration with rich sources of metadata for as many scientific disciplines as possible. We decided to integrate HarriGT with the Microsoft Academic Knowledge[15] , Scopus[16] and Springer[17] APIs. These APIs all provide broad, up to date coverage of known academic works. Each API had a different search endpoint with differing query languages and syntax that had to be catered for.

Each of the APIs returns metadata such as title, names and affiliations of authors, name of publishing venue and date of publication. In most cases each API returned a DOI so that each work could be uniquely identified and hyperlinked via the HarriGT interface. This allowed us to de-duplicate items returned by more than one API.

Articles typically talk about the institution that a scientific work was carried out at and independently the name of the author e.g. "Cambridge Researchers have found that... Dr Smith who led the study said..." making automatic extraction of reference information very difficult. Therefore, we use the NER matches generated for each news article in Section 5.3.2 to identify all names and institutions in the article and run citation graph queries for each permutation. For example: "A study run by Oxford and Cambridge universities found that... Dr Jones who led the study said..." would yield two queries: (Jones, Oxford), (Jones, Cambridge). Frequently, university press officers work with journalists to ensure that news articles are published on the same day or very shortly after a scientific work is published. However, in some cases, journalists may hear about new scientific works a few days after they are published via online press releases or other, earlier news articles. In some cases, metadata associated with news articles and scientific paper publication dates can be inconsistent (in particular if there are multiple publication dates for the scientific paper corresponding to an online version and a physical version. HarriGT's searches are bounded by the article's publication date plus-or-minus 90 days in order to take these variations into account,.

5.3.8 Candidate Scoring Implementation

The candidate retrieval mechanism described above in Section 5.3.7 tends to over-generate links between news articles and scientific publications, resulting in a significant number of false positives. Therefore it is important to have a mechanism for ranking these further, to avoid spurious matches and only show the user the most prominent ones for formal linking. To address this we propose a simple but effective mechanism based on the Levenshtein Ratio. Each news article is associated with a set of C candidate scientific works c_i where $i \in [0, C]$ are found using the retrieval method discussed above. News articles contain

[15] `https://makg.org/`
[16] `https://dev.elsevier.com/index.html`
[17] `https://dev.springer.com/`

two sets of entity mentions of interest: A set of N peoples' names n_j and a set of O organization names o_j. We also record the number of times each entity is mentioned M_j. For each candidate scientific work c_i, we identify a set of A_i authors' names a_i^k and their respective academic affiliations u_i^k. We also note the publication date of each news article D and the publication date of each candidate scientific work P_i.

For a given news article, we score each candidate scientific work c_i by summing over the square of Levenshtein Ratio ($L_r(x, y)$) of each pair of mentions of names and authors:

$$S_i^{per} = \sum_{j=0}^{N} M_j \sum_{k=0}^{A_i} L_r(n_j, a_k^i)^2$$

A similar calculation is carried out for organisation mentions and affiliations.

$$S_i^{org} = \sum_{j=0}^{O} M_j \sum_{k=0}^{A_i} L_r(o_j, u_k^i)^2$$

The Levenshtein Ratio is a simple, effective measure that has been used for assessing named entity similarity (Moreau et al., 2008). We also calculate Δ_D, the number of days between the publication date of the news article, D and the scientific work P_i. In cases where the candidate article has multiple publication dates (for example, online publication versus print publication), Δ_D is calculated for all publication dates and the smallest value is retained.

$$\Delta_D = \min_n(\sqrt{(D - P_i^n)^2})$$

Finally, we calculate an overall score S_i for each article by normalizing S_i^{per} and S_i^{org} by their respective numbers of distinct entity mentions and then dividing by Δ_D like so:

$$S_i = (\frac{S_i^{per}}{|N|} + \frac{S_i^{org}}{|O|}) \times \frac{1}{\Delta_D}$$

Candidates are ranked according to their S_i score in descending order so that the highest scoring candidates are presented to the user first.

5.3.9 Candidate Scoring Evaluation

To evaluate our candidate scoring technique, we use it to retrieve the N-best candidates for news articles with known links to one or more scientific papers. For each of the news articles in our ground truth collection, we retrieved all candidate scientific works from the citation graphs as described in section 5.3.7 above. We then use the scoring algorithm from section 5.3.8 above to rank the candidates then check to see whether actual linked papers

appear in the top 1,3 and 5 results (Top-K Recall).

	Top-1	**Top-3**	**Top-5**
Recall	0.59	0.83	0.90

Table 5.2: Top-K Recall for scoring algorithm for 296 news articles with 314 manually-validated scientific papers and 3964 automatically-proposed candidate scientific papers (mean 13 candidates per news article, median 11 candidates per article)

We identified a small number of reasons for sub-optimal ranking. Newspaper articles occasionally focus around candidate works published months earlier. In some cases, incorrect publication dates are being reported by the scientific paper APIs. In both cases, our system strongly penalizes candidates in terms of Δ_D. HarriGT's ranking algorithm also weakly penalizes candidates that have multiple authors in cases where only one author (often the lead) is mentioned in the newspaper text. This effect is amplified when work by the same lead author with fewer or no co-authors is also found since these candidates are preferred and filtered to the top of the list.

HarriGT's recall is not bounded by the candidate ranking algorithm but by the queries and results from our integration with Scopus, Microsoft and Springer APIs. HarriGT allows the user to **hide** news articles that are scientific but for which no relevant candidates are recommended. This action is distinct from marking an item as **spam**, which indicates that it has no scientific value and should be excluded from the corpus.

We evaluate the recall of our tool by considering items marked as **link** to be retrieved and deemed relevant and items marked as **hide** to be retrieved but for which no relevant items could be found. Thus defining recall as:

$$recall = \frac{|\{linked\}|}{|\{linked\} \cup \{hidden\}|}$$

At the time of writing, the recall of the system is 0.57. This figure may be lower than the actual figure, since papers are occasionally classified as 'hidden' by annotators if several strong candidates are presented and they are unsure which paper to link to. We expect that this figure will get stronger with more use.

5.4 Exploring the Relationship Between Comprehensive Impact & News Coverage With The HarriGT Corpus

Using the document linking strategies discussed above, we assembled a corpus of 5903 digital news articles and linked citation records (author names, title, publication date and venue) for 9891 scientific papers. The larger number of scientific papers being attributed to news articles that mention multiple scientific works. We were able to collect the full paper text for 1086 of the 9891 scientific paper citation records. We call this collection of records

and full documents the HarriGT Corpus. We were then able to identify 103 REF impact case studies with links to 140 of the news articles and 108 of the scientific paper citation records (see (F) in Figure 5.6).

We use subset F of the HarriGT corpus to explore the statistical relationships between REF impact and news coverage. As discussed in Chapter 4, REF impact scores for individual case studies are not published in order to preserve anonymity of academics. Instead, the number of 4*, 3*, 2*, 1* and unclassified case studies are reported per unit of assessment (UoA, approximately: faculty or department) at each partaking institution. REF controls for department size by mandating a minimum of two impact case studies to be submitted per UoA with an additional case study per 10 Full Time Employees (FTE) REF 2014 (2011). We found that the mean number of FTE per participating scientific UoA was 27.3 and thus the mean number of case studies submitted per UoA was 3-4.

Following the approach discussed in Section 4.2.1, we assign the mean impact score of the case studies from the associated UoA and institution as the score for individual case studies. Although some information loss is inevitable, in 96% of submissions the mean calculation involved 10 or fewer data points.

Given the national importance of REF, UK universities are heavily incentivised to

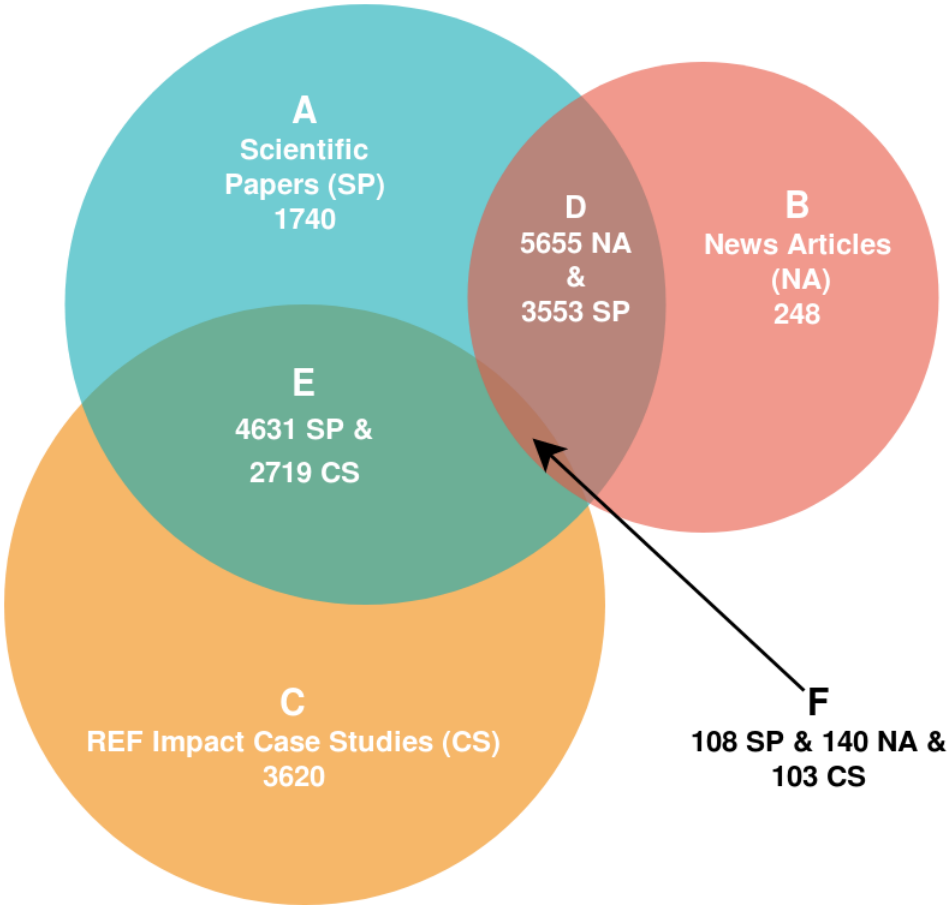


Figure 5.6: Documents in the corpus according to size and relation.

select their best scientific work for inclusion into case study submissions at UoA level. The limits on how many case studies a university may submit per UoA and the imposition of a 4 page limit per study (REF 2014, 2011) favours the inclusion of what academics consider to be their most impactful work and precludes reporting of less impactful work (in other words there is a focus on quality, not quantity). Therefore, we assume that the inclusion of a scientific paper in a REF case study is a general indicator of high comprehensive impact. The vast majority of scientific papers are not included in REF case studies either due to the small number of impact case studies submitted per institution UoA, the timing of the papers in relation to REF (i.e. they were published outside the reporting period) or the authors being external to the UK.

Using metadata from the HarriGT corpus, we split all REF impact case study results into two sets depending on whether they had news articles linked to them. Our hypothesis is that scientific papers linked to REF impact case studies with associated newspaper articles (F in Figure 5.6, referred to as "linked") have a higher comprehensive impact than those linked to cases studies without any such news articles (E, referred to as "unlinked"). Figure 5.7 shows a plot of the frequency distribution for these these two sets against average REF impact score as per the calculation above.

Using D'Agostino and Pearson's normality test (D'Agostino, 1971) we found that neither set of scores has a normal distribution ($p = 8.66 \times 10^{-17}$ and $p = 1.01 \times 10^{-220}$ for linked and unlinked subsets respectively). We therefore opted to use the non-parametric Kolmogorov-Smirnov 2-sample test Massey (1951) (KS-2 Test) to test the significance of the difference between the two distributions. The KS-2 test shows that the two samples are most likely drawn from separate populations ($p = 0.007$), supporting the hypothesis that the two sets represent distinct populations. We also test the error bounds of the two sets

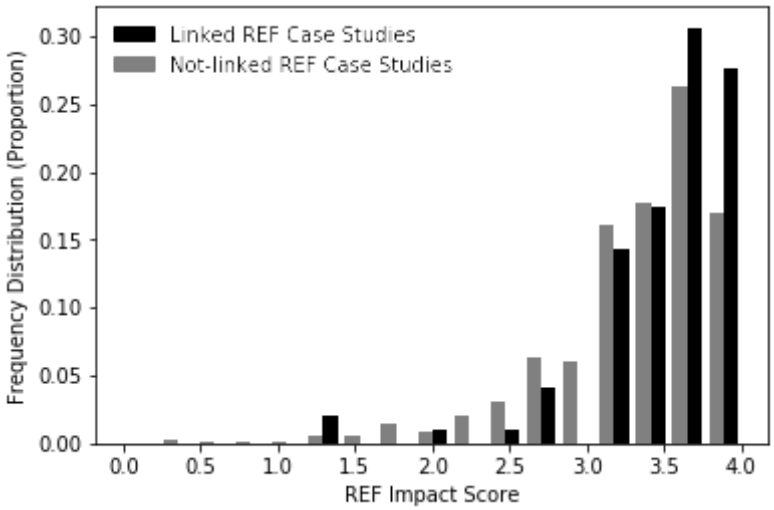


Figure 5.7: Frequency Distribution of REF Impact Scores for REF case studies that are linked ($F \cup G$, blue) or not linked ($C \cup D$,red) to one or more news article.

using a two-sample bootstrap test of mean difference (Hesterberg, 2015). This test yields a 95% bootstrap percentile confidence interval of [0.07,0.27] which suggests that *i.* the difference between the two distributions is too large to be attributed to random chance and *ii.* given that confidence interval is positive, the 'linked' set of case studies and papers tend to have a higher impact score than the 'unlinked' set.

5.5 Conclusion

In this chapter we aimed to identify a source for evidence of comprehensive impact that is easy to access, scalable and generalisable to different geographies and scientific disciplines and which has a clear statistical link to comprehensive impact performance in terms of REF Impact Score. We initially explored a number of data sources that provide evidence of scientific works' comprehensive impact, settling upon scientific news which is plentiful, covers a plethora of impact types including commercial and political impact can be easily collected. Scientific news provides broad but relatively shallow coverage of scientific disciplines. On the other hand, the remaining sources discussed in Section 5.2 provide narrower but deeper coverage of specific types of impact which may provide additional context and complement news coverage of scientific work. For example, scientific work that leads to the incorporation of a modestly successful startup or a minor policy change within a specific area of healthcare may not feature in national or international news publications but may be visible in appropriate commercial or political data. However, the biggest challenge faced by anyone aiming to build a system that integrates all of these complementary sources is their abundance and sparsity and the amount of manual engineering and analytical effort required for this integration.

We next assembled the HarriGT corpus, a dataset of news articles, scientific papers and REF impact case studies by processing over 62TB of web archive data and identifying 11 million historical news articles with potential links to scientific works. We developed a set of retrieval strategies that enabled us to search for links from news articles to scientific works and from REF impact studies to news articles where explicit links were available. For implicit links, we also built a semi-supervised web tool that uses machine learning to remove irrelevant content and facilitate ranking of candidate scientific papers, allowing humans to quickly and efficiently confirm links between news articles and scientific papers.

Finally, we carry out a statistical analysis on subsets of the HarriGT corpus with and without links to news articles, finding that scientific work with at least one associated news article is likely to achieve better a REF Impact Score than works that are not linked to the news. We find this relationship to be intuitive since more impactful works that have a larger effect on society, the economy or policy are more likely to be of interest to the general public and thus reported in the news. However, given the small size of the linked HarriGT corpus subset, future work should focus on understanding possible confounding variables such as

the presence or absence of press releases or the relative prestige of authoring academic institutions.

Having established the existence of a statistical link between news mentions and REF impact score, in the next chapters we investigate the mechanisms behind this phenomenon. In Chapter 6 we explore how information in scientific articles is transferred to newspaper articles and how we can identify and align paraphrasing excerpts of text across pairs of documents despite the differences in style and tone between the document types. Subsequently, in Chapter 9 we focus on the paraphrasing and rewording of information across the two document types, establishing a methodology for determining the 'impactfulness' of sentences.

6

Information Transfer Between News Articles & Scientific Papers

"Science is the only news. When you scan a news portal or magazine, all the human interest stuff is the same old he-said-she-said, the politics and economics the same cyclical dramas, the fashions a pathetic illusion of newness; even the technology is predictable if you know the science behind it. Human nature doesn't change much; science does, and the change accrues, altering the world irreversibly"

Stewart Brand

6.1 Introduction

In order to characterise the relationship between news coverage, scientific work and comprehensive impact that we uncovered in Chapter 5, we seek to better understand how information flows from scientific work into news articles. Although scientific papers and the news articles that discuss them share the same core subject matter, the purposes and audiences of these documents have very different purposes and audiences necessitating divergent structures and styles. The primary purpose of a scientific paper is to precisely communicate the technical details of scientific work to a deeply technical audience with appropriate scientific backgrounds in a way that allows other scientists to understand and recreate them. On the other hand, a scientific news article's primary purpose is to communicate scientific work to a broad and general readership in a way that makes them feel engaged, informed and entertained and encourages them to buy more newspapers.

As a key part of their role, scientific journalists must consider which information from a scientific paper should be included in their news articles and whether or not the content is accessible, re-writing complex scientific text in a fun and interesting way for their readership. This process usually results in writing that is significantly different from the scientific source material, some direct quotation and paraphrasing notwithstanding. Good scientific journalists will often make creative use of metaphor and simile to help make scientific work more relatable and understandable to their readers (Louis and Nenkova, 2013a). Furthermore, science news articles tend to include summaries of related work from relevant scientific disciplines in order to furnish readers with the context they need to understand the focal scientific work.

For a typical reader, the task of identifying journalistic writing that either originated in or was strongly influenced by specific excerpts of scientific writing is very challenging, direct quotation notwithstanding. Readers without a deep scientific background may have to do significant supplementary research to understand the technical scientific content enough

to be able to link it back to the simplified news writing. For example, a lay reader who encounters a new article titled "...Scientists Reverse Ageing In Old Mice Using Brain Fluid From Younger Mice...[1]" and its associated scientific paper[2] would be required to know what "oligodendrogenesis" is and that it is a useful biological process that mouse brains are less able to carry out as they age in order to make the connection that "revers[al of] ageing in old mice" refers to the resumption of oligodendrogenesis in the older mouse brain. Both the breadth and depth of this task also represents a significant challenge for automated systems; they must identify thematically aligned content across two documents that use very different vocabularies and levels of detail.

In this chapter we aim to address **RQ3** (section 1.2) and **CH4** (section 1.3.4) by developing methods for measuring semantic similarity as a proxy for information flow from scientific papers and the news articles that discuss them, despite the linguistic and structural differences between the two documents. We start with a discussion of the key differences between scientific news articles and scientific papers. We investigate several document representation methods, evaluating how well they are able to encode similarities between pairs of news articles and scientific papers via an information retrieval task. Finally, we discuss the suitability of each representation method for identifying conceptually aligned segments of document pairs and facilitating downstream analysis in later chapters

This work was carried out in Early 2018 before the widespread use of transformer-based language models like BERT (Devlin et al., 2019). Therefore, there are a number of newer models and techniques, which we did not explore at the time which could provide new directions for future work which we discuss further in section 6.6. However, many of the tasks involved in comparing and aligning information across long documents are still challenging for modern NLP models (as we briefly outline in **CH4** in Section 1.3.4).

Specifically the contributions of this chapter are as follows:

1. We carry out an analysis of document structure and vocabulary within the full-text subset of the HarriGT Corpus and show clearly the disparity between scientific papers and the news articles that discuss them.

2. We discuss and evaluate a series of state-of-the-art document representation techniques and their suitability for encoding information flow in pairs of linked news articles and full-text scientific papers HarriGT linked corpus (Chapter 5).

3. We define a new information retrieval task for determining how well the document representation techniques discussed encode information overlap between pairs of news articles and scientific papers.

4. We show that despite disjointed vocabularies, simple bag-of-words outperform more

[1] https://www.indiatimes.com/technology/science-and-future/scientists-reverse-ageing-in-old-mice-using-brain-fluid-from-younger-mice-569389.html

[2] https://www.nature.com/articles/s41586-022-04722-0

complex approaches at retrieving research papers described in corresponding scientific news articles but that they are less capable of encoding information flow between documents.

5. We make recommendations about future work and suggestions for where current state-of-the-art models could facilitate more effective encoding of information flow.

6.2 Data

The differing purposes and audiences of scientific papers and news articles lead to a number of key differences in the style, structure, length, vocabulary and readability of the two documents, all of which pose major challenges for automated linguistic analyses of the commonalities between pairs of topically aligned newspaper articles and scientific papers.

Scientific papers use very complex vocabularies to describe technical concepts and processes as precisely and specifically as possible. On the other hand, Newspaper articles aim to communicate these same concepts in a way that is accessible to as broad an audience as possible. Since larger, more complex vocabularies are negatively associated with readability (Pitler and Nenkova, 2008), it follows that news articles tend to use smaller, simpler vocabularies than scientific papers to make the content more accessible. In some cases news articles will include direct quotes from the authors of the studies that they describe and in these cases, simple exact-string matching may help to identify regions of information transfer between the two documents. However, journalists will often paraphrase the original work and use creative writing and metaphor to make complex scientific concepts more accessible to their readers (Louis and Nenkova, 2013b). For example, a recent news article reporting on COVID-19-related parosmia refers to the phenomenon instead as 'changes to smell.'[3]

6.2.1 Vocabulary and Communication Style

We ran a preliminary experiment in order to characterise differences in vocabulary across news articles and scientific papers. We tokenized all news and scientific article pairs for the top four news outlets in the full-text subset of the HarriGT corpus in order to calculate their document-wise vocabulary sizes and pairwise overlap of vocabularies. We counted all words with three or more letters that did not appear in a list of common English stop words. We calculated mean document-wise vocabulary sizes and mean pairwise-overlap by news outlet as shown in Table 6.1. Within our corpus, news articles typically have vocabularies around 14-28% of the size of associated scientific articles and between 7-12% of the words in those vocabularies are shared by both news articles and scientific papers. Furthermore,

[3] https://www.theguardian.com/world/2022/may/25/scientists-identify-trigger-molecule-for-covid-related-changes-to-smell

	News Outlet	# News Articles Linked to Science Papers	Mean News Article Vocab Size	Mean Science Paper Vocab Size	Mean Pairwise Vocab Overlap %
Quality Press	BBC	360	167	702	11.6
	The Guardian	585	266	922	12.1
Tabloid	Daily Mail	73	157	658	11.7
	The Express	57	97	685	7.7

Table 6.1: A summary of mean vocabulary size for newspaper articles and associated scientific papers in the HarriGT corpus and the size of the vocabulary overlap. The vocab size is slightly smaller for the Tabloid press than quality press, but for all news outlets, the vocabulary overlap with science papers is no more than 12.1%

the mean vocab sizes in the table also show an intuitive difference in complexity between quality press outlets which focus on research-based hard news coverage targeted at well informed, politically engaged readers and tabloid outlets which typically publish less detailed articles focusing on lifestyle and celebrity issues and target younger, less educated audiences (Bastos, 2016).

6.2.2 Document Length

We also calculated the mean lengths of all documents in the HarriGT corpus. Whilst the mean length of a scientific paper is 8151 words and median 6307, the mean length of a news article is 1012 words and median 828 words. Prior studies support our findings; scientific papers are typically around 5000-7000 words (de Araújo, 2014) dependent on scientific discipline and typical news articles tend to be around 600-700 words in length (Wobbrock et al., 2021; Menéndez Alarcón, 2012). However, Louis and Nenkova (2013b) consider only articles containing 1,000 words or more for their corpus of science news articles, suggesting that this is a reasonable length for good quality science news articles. We further break down mean lengths of the linked news articles in the HarriGT corpus in Table 6.2.

Given the respective goals of 'entertain and inform' versus 'explain precisely', the disparity in length between news articles and scientific papers is intuitive. Scientific journalists aim to summarise a story briefly without boring or scaring the reader away; it is intuitive that longer articles are harder to read than shorter articles and this was also found to be the case by Pitler and Nenkova (2008). Journalists are also likely limited to 1-2 pages of A4 paper by the newspaper editorial team and must therefore write concisely and clearly. Conversely, scientific papers must describe exactly and precisely the work carried out, requiring much more space. Scientific papers are typically limited to a set number of pages

	News Outlet	# News Articles	Mean Length (words)	Mean Length of Linked Science Papers
Quality Press	BBC	360	687	6162
	The Guardian	585	1325	9847
Tabloid	Daily Mail	73	642	5854
	The Express	57	374	6786

Table 6.2: A breakdown of mean lengths of news articles from different outlets in the full-text subset of the HarriGT corpus. Quality press articles are typically longer than Tabloid articles, and some Tabloid articles can be really very short when compared to the length of the scientific work that they describe. Linked scientific papers are of similar length except for the Guardian which links to a number of longer papers and has the longest news articles.

depending on publication venue and discipline. However, ensuring that scientific papers are easy to read is usually a much lower priority than ensuring that the work is complete, precise and rigorous. As with article vocabulary (see section 6.2.1), there is a noticeable difference in length between Quality Press and Tabloid articles. Quality Press articles are typically longer than Tabloid articles which likely reflects the preferences of the different audiences served by these outlets.

The disparity in length between document pairs makes the selection and reformulation of relevant information from scientific papers to be communicated in news articles all the more important and the omission of important information all the more detrimental to the goal of communicating scientific work.

6.2.3 Types of Scientific News & Links to Scientific Papers

Some news articles link to multiple scientific works which makes the task of identifying which part of the news article relates to which scientific paper even more difficult. Within the full-text HarriGT corpus, the majority of news articles only have one associated scientific paper. Figure 6.1 shows the distribution of articles with links to more than one scientific paper. Common reasons for articles to link to more than one work are to highlight rebuttals[4] or commentaries[5] from other academics, related works from the same author[6], or to provide reference to earlier or tangential findings[7] to help the reader to contextualise the focal work. Occasionally, writers will provide summaries of a large number of works in the same article[8]. When multiple scientific papers are referenced in a news article, attributing news content to an appropriate scientific paper becomes even more complicated.

[4] https://www.bbc.co.uk/news/health-19743584

[5] http://news.bbc.co.uk/earth/hi/earth_news/newsid_9079000/9079963.stm

[6] https://www.bbc.co.uk/news/science-environment-21958547

[7] https://www.theguardian.com/science/grrlscientist/2015/jul/30/golden-jackal-a-new-wolf-species-hiding-in-plain-sight

[8] https://www.theguardian.com/science/blog/2011/sep/29/immortality-ig-nobel-prize-winners

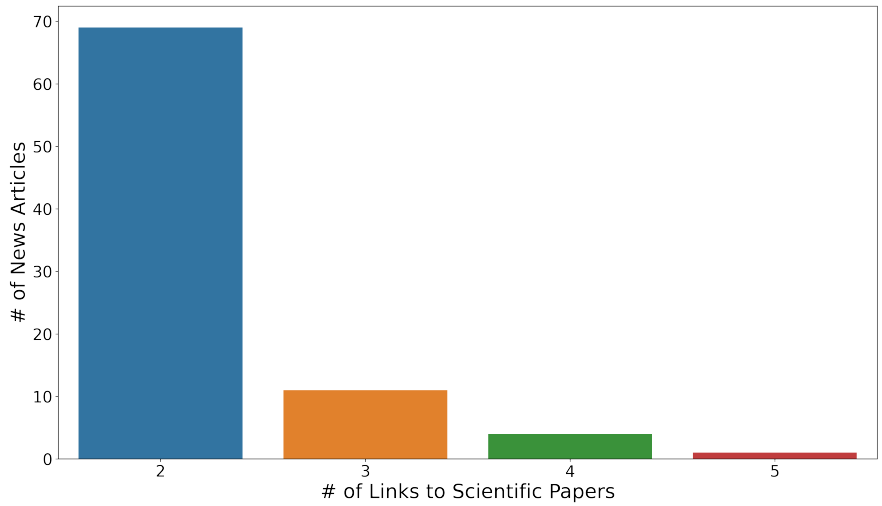


Figure 6.1: Count of news articles in the HarriGT corpus with links to more than one scientific paper. Whilst most articles (1001, not shown) only convey information about one scientific work, almost 70 articles refer to two papers and a handful of articles discuss 4 or more.

6.3 Methods

The significant differences in length and vocabulary between news articles and scientific papers highlighted above pose significant challenges to understanding how information flows from scientific papers to news articles. We next describe our approach to quantifying information flow (Section 6.3.1). We then outline a number of candidate document encoding techniques which could be used to encode news-science document pairs for comparison (section 6.3.2). Finally, we define an information retrieval task which we use to evaluate how well each of the select document encoding techniques is able to express the semantic similarity between pairs of documents (section 6.3.3).

6.3.1 Measuring Information Flow

Scientific news articles are written about, and thus informed by, scientific papers. Therefore, we hypothesise that one can approximate information flow from a scientific paper to a newspaper that discusses it via the semantic textual similarity (STS, discussed in section 2.8.4) between the two documents. Simply put: if a news article is semantically similar to its corresponding scientific paper, there has been more information transfer and vice versa.

Unfortunately, many existing approaches to STS focus on short excerpts with similar lengths and vocabulary (e.g. comparisons of single sentences or paragraphs), precluding their use in our setting without significant adaptation (Majumder et al., 2016). However,

our task is quite similar to Cross-Level Semantic Similarity (CLSS) (Jurgens et al., 2014, 2016). CLSS is a generalisation of STS, in which differently sized units of text are compared to see how well the meaning of a larger excerpt of text is captured by a shorter excerpt. For example, how well does a sentence summarize a paragraph? On the other hand, CLSS does not perfectly fit the challenges posed by our dataset. CLSS is normally performed on small contiguous blocks of text up to a paragraph length and both texts usually have similar, overlapping vocabulary. Not only are the the vocabularies of the two document being compared significantly different but the size difference between blocks of text is significantly more extreme; journalists summarise multiple pages of information into one or two pages. They also often dilute coverage of scientific work by interspersing it with background context, interviews with papers' authors and speculation about future work to make them more accessible and interesting to general readers.

Textual Entailment (TE) or Natural Language Inference is the task of detecting whether or not a premise entails a hypothesis Dagan et al. (2006). For example, the premise "the boy rode a bike to town" entails the hypothesis that "the boy went to town" but does not entail the hypothesis "the bike is red" since this latter information is missing from the first sentence. TE is normally carried out on pairs of single sentences or between a longer premise (up to a paragraph in length) and a single sentence hypothesis Giampiccolo et al.. Given that TE requires aligned premise-hypothesis pairs as an input, it may be considered a useful downstream task once information flow is determined and could be used for applications such as fact checking of news articles.

In order to find the semantic similarity between our news articles and scientific papers, they must be encoded into forms that are comparable either via a model which produces a similarity score or directly via a similarity metric (see section 2.4).

6.3.2 Representing Long Documents with Varied Vocabularies

In this section, we evaluate the suitability of several different document representation methods in our setting. We discuss how each representation method handles varied document length and vocabulary as well as their ability to encode and compare information across documents.

6.3.2.1 HarriGT Scoring Baseline

The HarriGT Scoring Baseline uses the approach described in Section 5.3.8 to match articles based on publication dates and named entity information. In our experiment, all scientific articles, including the distractor articles, are scored and ranked in descending order.

6.3.2.2 Sparse Vocabulary & Bag-of-Words Representations

Count-based and TF-IDF bag-of-words representations (BoW, see section 2.5.1) are simple and robust to document length. The dimensionality of BoW representations is governed by the size of a pre-determined joint vocabulary and is de-coupled from document length entirely. Therefore, news articles and scientific papers with different lengths can be represented using vectors with the same dimensionality and their representations can be normalised and directly compared using appropriate metrics such as Jensen-Shannon Distance (introduced in section 2.5.8.2).

BoW representations model each n-gram in their vocabulary as independent features which means that relationships between synonymous and related words are not captured. Therefore, if the vocabularies between two documents are too dissimilar due to paraphrasing (e.g. 'parosmia' and 'changes to smell'), BoW-based approaches are likely to yield poor results. However, TF-IDF representations are highly sensitive to terms that are used with a high frequency in a very small number of documents. Therefore, the tendency of journalists to directly quote excerpts of scientific papers and specific technical terms is likely to improve performance of a TF-IDF-based encoding strategy.

BoW representations can only provide a very limited explanation of information transfer in the sense that it is possible to see which n-grams are used in both documents. However, the contexts that these terms are used in cannot be further analysed without subsequent downstream processing or inspection of the source documents.

In our experiment, we make use of TF-IDF document embeddings as a baseline document encoding approach. We use Scikit-learn (Pedregosa et al., 2011) to tokenize and stem all documents and generate a bag-of-words feature vector containing all words with three or more letters and that are not on a list of English stop words that appear in the corpus. Each news article and scientific paper is represented as a sparse word count vector (as described in Section 2). For TF-IDF similarity, IDF scores are calculated with respect to the HarriGT corpus only and distractor documents are held out. Pairwise Jenson-Shannon divergence is calculated for each news/science pair and scientific papers are ranked in ascending order with respect to each news article (less divergent implies more similar).

6.3.2.3 Topic Modelling and Latent Dirichlet Allocation

Provided that a large enough training dataset is available, topic modelling approaches like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) can provide a partial solution for comparing documents of varying lengths with disjoint vocabularies by learning probabilistic associations between related words. As discussed in Section 2.5.5, LDA represents each word in a corpus as a probability distribution over learned topics based on word co-occurrence within different documents in the training data. Related and synonymous words, represented independently by LDAs underlying BoW model, have a high probability of ap-

pearing in similar contexts across a corpus, increasing the likelihood that they are assigned similar topic distributions from the model. E.g. the scientific term ‘parosmia’ and its lay definition, ‘changes to smell’ are both likely to co-occur with ‘covid-19’ and ‘symptoms’ in their respective documents, increasing the probability that the model assigns similar topic distributions to these words.

LDA also decouples representation from document length, representing documents as a probabilistic mixture of topics based on the topic distributions associated with the words that occur within them. These topic mixtures can then be compared using an appropriate similarity function as illustrated in Figure 6.2.

Information transfer between documents can also be analysed by comparing respective usage of specific topics between document representations as exemplified by Prabhakaran et al. (2016) and Lyu et al. (2022) who show topic evolution in the fields of computer science and micro-biology respectively.

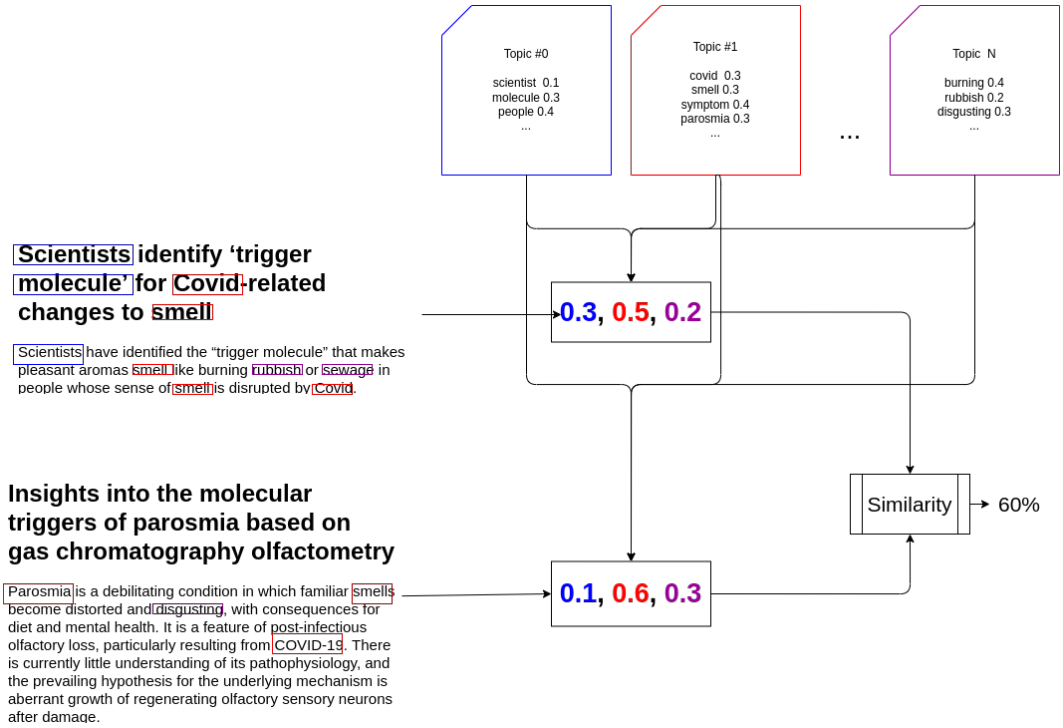


Figure 6.2: An illustration of an LDA approach to document similarity. Distinct words found with similar neighbouring words are assigned a high likelihood of belonging to the same topic.

In our experiment, we preprocess documents using the same rules for stemming, minimum characters and stop words as for Word Count above. We use the Gensim implementation of LDA (Rehurek and Sojka, 2011) to train topic models on all news articles and scientific papers in the corpus. For each document, we calculate the topic distribution and calculate the pairwise Jenson-Shannon divergence for each pair of news and science topic distributions. Scientific papers are then ranked in ascending order with respect to each news

article.

6.3.2.4 Addressing the News/Science Language Gap: PolyTM and DiaTM

Given the gap in vocabulary between news articles and science papers, there is still a risk that many important terms do not co-occur with shared or common terms frequently enough for LDA to model their semantic relationships adequately. PolyTM (Mimno et al., 2009) is an extension of LDA that addresses a similar problem to ours; how can topics be modelled across multi-lingual corpora, such as European parliamentary proceedings, given that topically aligned documents written in different languages are unlikely to have *any* vocabulary overlap? PolyTM learns a set of language-specific topic distributions from tuples of documents that are "...loosely equivalent to each other, but written in different languages..." *(ibid.)* Like LDA, PolyTM represent documents as k-dimensional vectors based on the number of topics that are modelled, facilitating the direct comparison of differently sized documents. PolyTM learns a single set of topics across all modelled languages allowing us to derive information flow across documents written in different languages by comparing topic usage between vectors. PolyTM learns a separate vocabulary per language such that terms appearing in different languages are modelled independently.

Although science papers and the news articles that discuss them may cover some of the same concepts, their distinct purpose, length and structure prevent us from considering them 'loosely equivalent'. Furthermore, unlike the polylingual task setting, news and science documents are written in two different styles or dialects of the same language which are not mutually exclusive; news articles may directly quote scientific terminology and likewise, scientific work may include simplistic language. Therefore, the possibility that a term may appear in one or the other set of documents is better characterised as a probability distribution over dialects rather than via multiple independent representations.

Crain et al. (2010) propose another extension to LDA called Dialect Topic Models (diaTM) which is positioned as an alternative to PolyTM. DiaTM learns to account for the expression of the same topic across different dialects of the same language. The model assumes that documents are written in a mixture of dialects, relaxing the constraint of PolyTM and allowing for an overlapping vocabulary, by representing the association of words with dialects as a Dirichlet distribution. Like PolyTM and LDA, documents of different lengths are be represented as k-dimensional vectors and information flow can be derived by comparing topic usage between vectors.

We re-implemented Crain et al. (2010)'s diaTM model from scratch [9] and found that it worked well with toy datasets. However, the model would not converge for our news and science documents. After a number of attempts at hyper-parameter optimisation and debugging, we decided to omit diaTM from our experiment.

[9] `https://github.com/ravenscroftj/diatm`

6.3.2.5 Doc2vec

As discussed in Section 2.5, many recent state-of-the-art works in NLP have made use of transfer-learning by through the integration of pre-trained context-independent word embeddings to provide a rich feature-set for their models[10]. As discussed in section 2.5.6.3, per-word vectors must be combined together to represent sequences of words. This can be problematic for representation of long sequences, which makes direct application of such vectors to our task inappropriate.

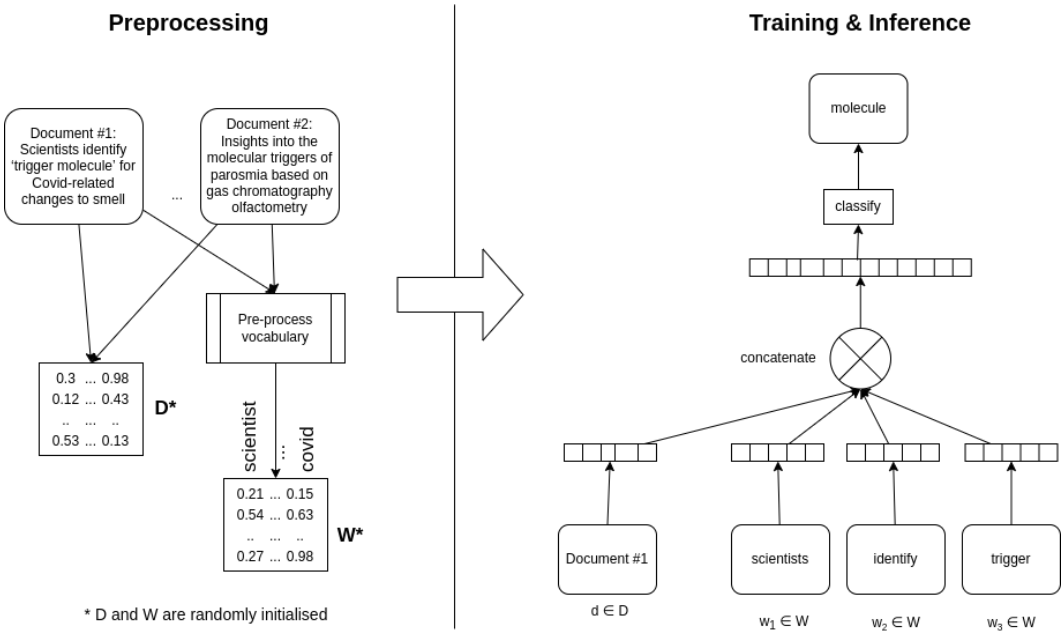


Figure 6.3: An illustration of doc2vec preprocessing (left) and training (right) processes. During preprocessing, words and documents are assigned random vectors in **D** and **W** respectively. These vectors are subsequently fine-tuned during training. New document vectors are inferred based on the vectors belonging to the words contained within the document.

As an alternative, we experiment with Doc2Vec (Le and Mikolov, 2014). Doc2Vec is an extension of the CBOW algorithm (Mikolov et al., 2013a), which learns semantic dense vector representations for full documents as opposed to single words and has been shown to facilitate good results for semantic textual similarity and question paraphrasing/duplication detection tasks (Lau and Baldwin, 2016).

We outline the Doc2Vec process in Figure 6.3. As in standard CBOW, all in-scope words are assigned vectors in a randomly initialised matrix W. Additionally, all documents in the training corpus are assigned a vector in a randomly-initialised matrix D. Training samples are generated by randomly sampling a target word and context words from the same document to be passed in as features to the classifier. The document vector and context

[10]At the time that this work was carried out, contextual embeddings were not yet widely used and models such as BERT (Devlin et al., 2019) had not yet been published. However, we discuss potential future applications of these systems in our task in 6.6

words are then concatenated together and used to predict the target word. Weights W and D are updated via gradient descent and backpropagation. As a result of this training, matrix W captures semantic representations of words and matrix D captured document semantics.

Once Doc2Vec is trained, vectors for new unseen documents can be calculated by running an inference step using words from the new document. Resulting vectors can be directly compared using measures such as cosine similarity. The dimensionality of these vectors is invariant with respect to the length of the document.

Like word2vec and LDA, Doc2Vec relies on the distributional knowledge (Harris, 1954) inherent in large text corpora to learn semantic relationships between words. When trained on a corpus of linked news articles and scientific papers, the model should be more likely to assign similar vectors to documents that have some overlapping vocabulary. Unlike LDA where dimensions in document representations correspond to meaningful topics, doc2vec dimensions are arbitrary. Therefore information transfer between doc2vec representations is difficult to track.

In our experiment, we train a Doc2Vec model as described in 6.3.2.5 using the full set of news articles and full text scientific papers from the HarriGT corpus. Distractor papers are held out during training and their document vectors are subsequently inferred. Pairwise cosine similarity is calculated for each news article/scientific paper pair including for distractor documents. Scientific papers are ranked in descending order. At time of writing, Doc2Vec has since been superseded by more recent models for long documents such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2021) which we discuss further in Section 6.6.

6.3.3 Evaluating Cross-Domain Document Representations via Information Retrieval

We evaluate the representation techniques chosen in section 6.3.2 (TF-IDF, Word Count, LDA, doc2vec) by running an information retrieval experiment using the Linked HarriGT corpus.

In order to test which representations best encode the information transfer between scientific papers and news articles, we first separate each news article from its paired scientific paper. We then take each news article in turn and rank all scientific papers in terms of the similarity of their representation to that news article. The information retrieval results act as an approximate proxy for semantic similarity and allow us to evaluate how well each representation is able to encode information transfer between documents.

For each representation, we encode all news articles and scientific papers then measure the pairwise similarity between each news/science article pair with an appropriate metric. Finally, we rank each scientific paper by most to least similar with respect to each news article. We report our findings in terms of Normalised Discounted Cumulative Gain (nDCG) which we describe below in section 6.3.4.2). We assign a relevance score of 1 to scientific

papers linked to the news article in question and a score of 0 to all other papers.

Each of the representations are tested using the full-text subset of Linked HarriGT corpus containing 1086 full news articles and 1136 full text scientific papers. To make the retrieval task more challenging, we include a further 2000 randomly sampled unrelated distractor papers from our ECNs (Chapter 4) to produce a larger set of scientific papers for ranking.

6.3.4 Metrics for Evaluation of the Ranking

6.3.4.1 nDCG vs Top-K Recall

In Chapter 5.3.9) we used Top-K Recall to show whether relevant results are ranked within the top N results. However, for worse performing ranking strategies, Top-K Recall quickly degrades to 0, yielding meaningless comparisons. For example an algorithm consistently placing relevant information in the top 10 positions in a ranked list could achieve a Top-5 Recall of 0% if the first 5 results are consistently irrelevant. For this experiment, we aim to find a ranking strategy that degrades linearly in order to provides us with a way to make like-for-like comparisons between high and low performing ranking algorithms.

6.3.4.2 Normalised Discounted Cumulative Gain (nDCG)

Normalised Discounted Cumulative Gain (nDCG) is a metric for measuring how well a ranking model works that was first described by Järvelin and Kekäläinen (2002). One of the key intuitions for NDCG is that even imperfect ranking has some value; for example, a ranking scheme that places a relevant document *near* the top of a list of 10,000 results rather than assigning it the top spot is still more valuable than a random shuffle or a scheme that places the document at the bottom of the list.

To find the nDCG for a given query, we start by assigning scores to each item that was retrieved based on what we know *a priori* about how relevant it is. For example, we might assign a relevance score of 1 to scientific papers that are linked to a news article and a score of 0 to those that are not linked. Alternatively, you may have multiple levels of relevance: in a web shop search you might assign items of the same type to the last item that the user looked at (t-shirts) a relevance score of 2 and items in the same category (clothing) a relevance score of 1. Irrelevant items (e.g. car batteries) are given a score of 0.

We next calculate an ideal order that we would like the ranking algorithm to produce based on the retrieved items by sorting them by their relevance scores in descending order such that the most relevant items are at the top of the list. Items with tied relevance scores (e.g. all the t-shirts from the above example) may be sorted arbitrarily.

We next sum the relevance scores for the ideal ranking orders to calculate the ideal cumulative gain. However, we discount the relevance score attributed to each item by an amount based on its offset from the beginning of the list (i.e. each relevant item is given

a slightly worse score for being further down the list). Thus, we calculate the ideal DCG (iDCG) for a collection of documents of length N:

$$iDCG = \sum_{i}^{N} \frac{score(i)}{log_2(i)} \tag{6.1}$$

For example, for a news article entitled "Cancer Cured In Rats" which has 2 relevant scientific papers in a collection of N papers in total, the ideal order would be:

- Relevant Paper #1 (Score=1)
- Relevant Paper #2 (Score=1)
- Irrelevant Paper #1 (Score=0)
- ...
- Irrelevant Paper #N (Score=0)

The corresponding ideal DCG (iDCG) would be:

$$\begin{aligned} iDCG &= 1 + \frac{1}{log_2(2)} + \frac{0}{log_2(3)} + \cdots + + \frac{0}{log_2(N)} \\ &= 1 + \frac{1}{1} + \frac{0}{1.58} + \cdots \\ &= 1 + 1 + 0 \cdots + 0 \\ &= 2 \end{aligned} \tag{6.2}$$

In the case that the ranking algorithm returned the following imperfect rank:

- Relevant Paper #1 (Score=1)
- Irrelevant Paper #1 (Score=0)
- Irrelevant Paper #4 (Score=0)
- Relevant Paper #2 (Score=1)
- Irrelevant Paper #5 (Score=0)

The DCG for the predicted rank would be:

$$\begin{aligned} iDCG &= 1 + \frac{0}{log_2(2)} + \frac{0}{log_2(3)} + \frac{1}{log_2(4)} + \frac{0}{log_2(5)} \\ &= 1 + \frac{0}{1} + \frac{0}{1.58} + \frac{1}{2} + \frac{0}{2.39} \\ &= 1 + 0 + 0 + 0.5 + 0 \\ &= 1.5 \end{aligned} \tag{6.3}$$

Representation Strategy	Mean nDCG by News Outlet			
	BBC	**Guardian**	**The Express**	**Daily Mail**
HarriGT Scoring (Baseline)	0.31	0.27	0.32	0.45
Doc2Vec (25 Dimensions)	0.46	0.47	0.35	0.54
Doc2Vec (50 Dimensions)	0.58	0.58	0.44	0.74
Doc2Vec (100 Dimensions)	0.61	0.63	0.44	0.82
LDA 100 Topics	0.42	0.40	0.23	0.46
LDA 300 Topics	0.64	0.50	0.47	0.65
LDA 500 Topics	0.66	0.52	0.52	0.70
Word Count Similarity	0.90	0.73	0.82	0.91
TF-IDF Similarity	**0.92**	**0.76**	**0.87**	**0.93**

Table 6.3: Results from science article retrieval experiment. Bag-of-words-based approaches yield best overall performance but perform worst on the Guardian. All other non-baseline methods perform worst on The Express.

We finally calculate our *normalised* DCG by dividing the DCG by the IDC:

$$nDCG = \frac{DCG}{iDCG} \quad = \frac{1.5}{2} = 0.75 \tag{6.4}$$

Each query may have a different maximum DCG score (e.g. in our system, each news article has one-to-many associated scientific papers) so by normalising each DCG with respect to its corresponding ideal score, we are able to make like for like comparisons between queries and meaningfully calculate the average nDCG for a given set of queries.

6.4 Results & Discussion

Our results are reported in table 6.3. They show that the Bag-of-words-based approaches yield best overall performance but perform worst on the Guardian. All other non-baseline methods perform worst on The Express. We discuss the findings for each method in the following subsections.

6.4.1 HarriGT Scoring Baseline

As expected, the HarriGT baseline scoring approach, based solely on named entities in news articles, author metadata and difference between dates of publication yields worst performance. It performs marginally better for the two tabloid outlets (The Express, Daily Mail) than the quality press outlets.

As discussed in Section 5.3.8, HarriGT Score is strongly dependent upon the delta between the publication date of the scientific paper and that of the newspaper article; our hypothesis being that news articles are more likely to focus on recently published works. We plotted the distributions of deltas between scientific paper publication date and news article publication date in Figure 6.4. The highest concentration of publications across all

news outlets comes the same day or shortly after a scientific paper is made available which likely corresponds with press embargoes being lifted when the journal is released. The two quality press outlets have more varied gaps which likely corresponds with their tendency to cover scientific work in more depth (e.g. by referencing relevant background work).

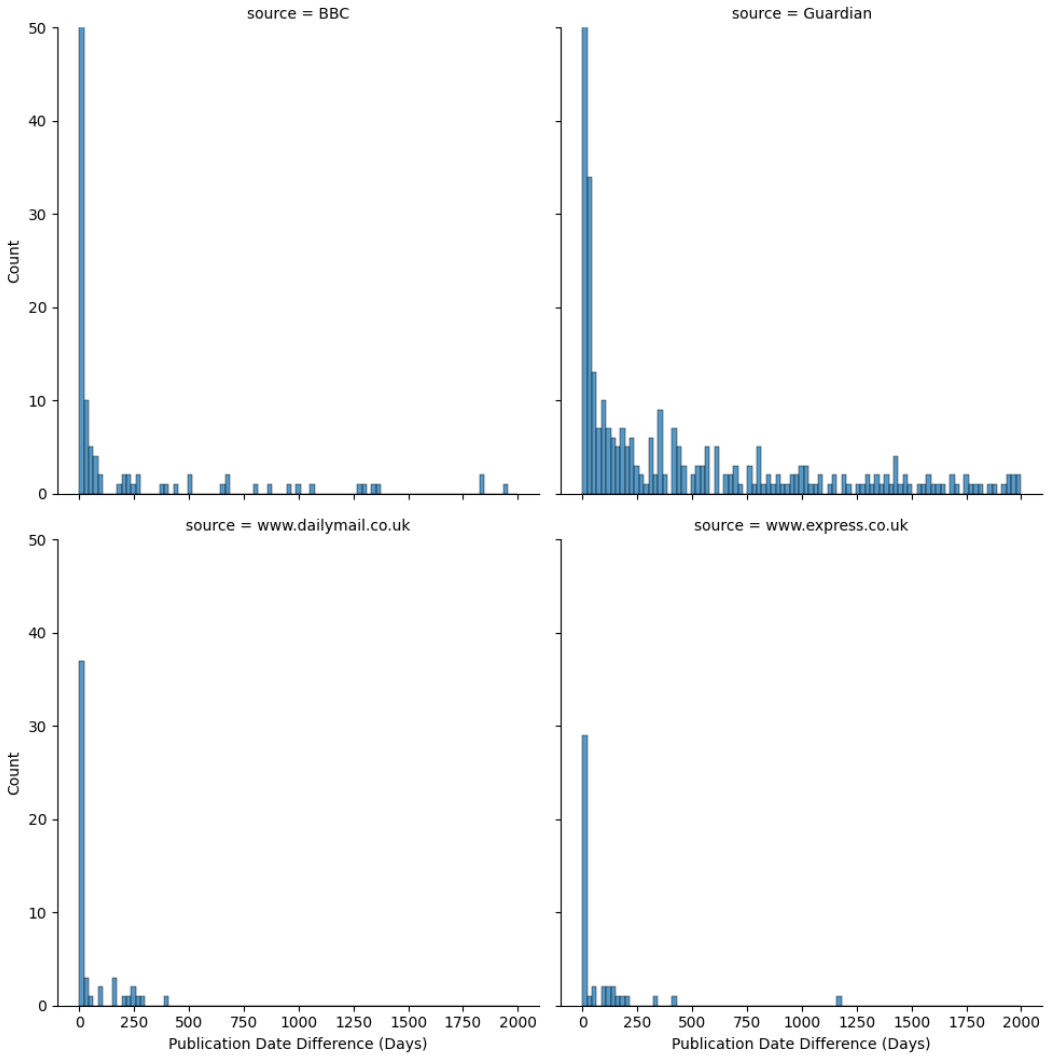


Figure 6.4: Histograms of publication date difference between newspapers and the scientific work that they are reporting on for the top 4 sources in the full-text subset of the HarriGT Linked Corpus. Graph axes limited at y=50 and x=2000 for illustrative purposes. Both quality press outlets refer to older articles to provide background context. Tabloid outlets only discuss work older than 6 months a handful of times.

Whilst the majority of the quality press articles were automatically linked to scientific papers via URLs and DOIs (see section 5.3.5), most of the tabloid article links were manually confirmed via the HarriGT web interface and scoring mechanism. Therefore, there may also be a degree of selection bias in these results since tabloid articles published closer in date to target scientific papers would be more likely to be displayed prominently

in the annotation interface.

6.4.2 Word Count & TF-IDF Similarity

#	News Article	Scientific Paper	nDCG	JSD	Overlapping Terms (Stemmed)
1	Obama-named lizard was wiped out with the dinosaurs	Mass extinction of lizards and snakes at the Cretaceous - Paleogene boundary	1.0	0.52	lizard, snake, extinct, gracili, fig, cretac, dinosaur, fi, squamat, longrich
2	Spine manipulation for neck pain 'inadvisable'	Should we abandon cervical spine manipulation for mechanical neck pain? Yes	1.0	0.34	manipul, neck, cervic, spine, pain, neurovascular, dissect, arteri, chiropract, treatment
3	Melt may explain Antarctica's sea ice expansion	Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008	0.29	0.58	ice, icesat, thick, melt, sea, antarct, water, winter, draft, knmi
4	Neanderthal Breeding Idea Doubted	Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins	0.63	0.48	neanderth, modern, popul, african, hybridis, human, genom, genet, africa, fi
5	Study sheds light on plants' 'spring switch'	PHYTOCHROME INTERACTING FACTOR4 controls the thermosensory activation of flowering	1.0	0.49	pif4, plant, flower, temperatur, wigg, warmer, 35s, crop, arabidopsi, fig

Table 6.4: A table of scientific paper and news article titles and the top 10 overlapping words by tf-idf score. A mixture of examples where the representations worked well (nDCG=1.0) and where they were less effective (nDCG < 1.0) are provided for comparison.

We observe that the two bag-of-words approaches, TF-IDF Similarity and Word Count Similarity, yield the best retrieval results overall. This likely suggests that despite the limited shared vocabulary between pairs of documents (as discussed in section 6.2.1), the terms that *are* common across articles are specialised enough that they serve as helpful identifying features within the corpus. We carried out a qualitative examination of correctly

and poorly matched document pairs for the TF-IDF Similarity approach by examining the full texts, nDCG and most important TF-IDF terms of a selection of randomly chosen news article and scientific paper pairs from the full-text subset of the Linked HarriGT corpus.

In cases where the TF-IDF retrieval approach worked well (rows # 1, 2, 5 in Table 6.4), subject-specific terms were identified as important by the TF-IDF scoring algorithm which allow them to be matched. For example in #1 stemmed versions of the lizard's species name *gracilis* and the period *cretaceous* are mentioned in both papers and act as specific reference points. However, for examples # 3, 4, the top-ranking terms according to TF-IDF are broader and more applicable to a wider set of subjects making them more susceptible to false positive matches. In row #3 the top keywords relate to ice melt, sea water and the Antarctic which are general topics that occur multiple times within the corpus. Indeed the strongest TF-IDF match for the news article in pair #3 is "The Extreme Melt Across The Greenland Ice Sheet in 2012"[11] with a JSD of 0.53 and shared stemmed keywords including 'melt', 'ice', 'water', 'antarct' and 'sea'

Whilst TF-IDF representations are often good enough to match topically aligned documents together, the representations only provide a superficial insight into the semantic similarities between documents and do not help us to understand where information is being transferred. In document pair #5 from Table 6.4, the news article[12] summarises the main discovery from the scientific paper[13] "a gene, known as PIF4, activated the flowering pathway when a certain temperature was achieved" and paraphrases the author's suggestion that "...the findings could be used in future research to improve crops' resilience to projected changes in the climate". However, the TF-IDF score is based on the specific words shared between the two documents (PIF4, Arabidopsis thaliana) rather than the semantic similarity between the scientist's descriptions of the discovery, conclusions and the journalist's respective summaries and paraphrasings in the news article.

6.4.3 LDA

LDA yields modest IR performance when the number of topics k is set to 100 but increasing the number of topics significantly improves performance, yielding best performance for the BBC and Express after the BoW approaches (Word Count and TF-IDF).

Given the broad, diverse nature of the documents within the corpus, we hypothesise that fitting LDA with a larger number of topics allows the model to better place distinct themes and concepts within individual topics. We investigate this hypothesis by measuring c_v topic coherence, a measure which strongly correlates with human ratings of topic quality (Röder et al., 2015). This metric involves calculating the Normalized Pointwise Mutual Information (see Manning and Schiitze (1998, pg 68)) for the 'top words' associated with

[11] https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2012GL053611

[12] https://www.bbc.co.uk/news/science-environment-17447012

[13] https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC4972390&blobtype=pdf

Topics	Average Coherence (c_v)
100	0.455
300	0.469
500	0.490

Table 6.5: c_v Topic Coherence for LDA models trained on the full-text subset of the Linked HarriGT Corpus. With more topics, LDA is able to generate a more coherent model of the topics in the corpus.

a given topic the words that co-occur with them within the corpus, producing a series of co-occurrence vectors. The overall c_v for a topic is the mean pairwise cosine similarity for all co-occurence vectors. Values approaching 1 are good (top words frequently co-occur in the corpus) and values approaching 0 are bad (top words rarely co-occur in the corpus). The c_v coherence results for our three LDA models are shown in Table 6.5. For our corpus, c_v also increases with the number of topics in our model.

We also manually inspect words that are most strongly associated with randomly chosen topics from the best performing model ($k = 500$), shown in Table 6.6. This shows that many of the topics in the model do converge around meaningful, coherent subject matters although some of the topics (e.g. Topic #477) do appear random.

LDA also provides a summary of the information transfer between documents through the comparison of topic mixture representations of news and science documents as illustrated in Figure 6.5. The figure shows that the news article and scientific paper from example # 1 in Table 6.4 both correspond to topics 16, 38 and 87 but that proportionally, more coverage of topics 16 and 38 is assigned to the scientific paper than the news article. Likewise, the news article has a stronger link to topic 87 than the scientific paper. Both documents are also strongly associated with topics that are not shared by their counterpart document. The largest proportion of the scientific paper's topic mixture is assigned to 94 other topics with low likelihood whereas the news article is only associated with 8 topics.

Given the discrepancy in length between news articles and scientific papers and the fact that LDA is a probabilistic generative model, it follows that the longer scientific papers are more likely to be associated with a larger number of topics than the shorter news articles. However, the difference in document structures may also provide a partial explanation for

Topic	**Coherence (c_v)**	**term 1**	**term 2**	**term 3**	**term 4**	**term 5**
126	0.72	milk	mirna	transg	express	bovine
335	0.78	mass	boson	sect	model	ev
93	0.73	men	placebo	week	mg	testosteron
477	0.13	enzym	blog	carbapenem	piec	artic

Table 6.6: Coherence and most likely word-stem terms for selected topics from our LDA 500 topic model. Topics 126, 335 and 93 are clear and coherent topics whilst topic 477 appears less clear.

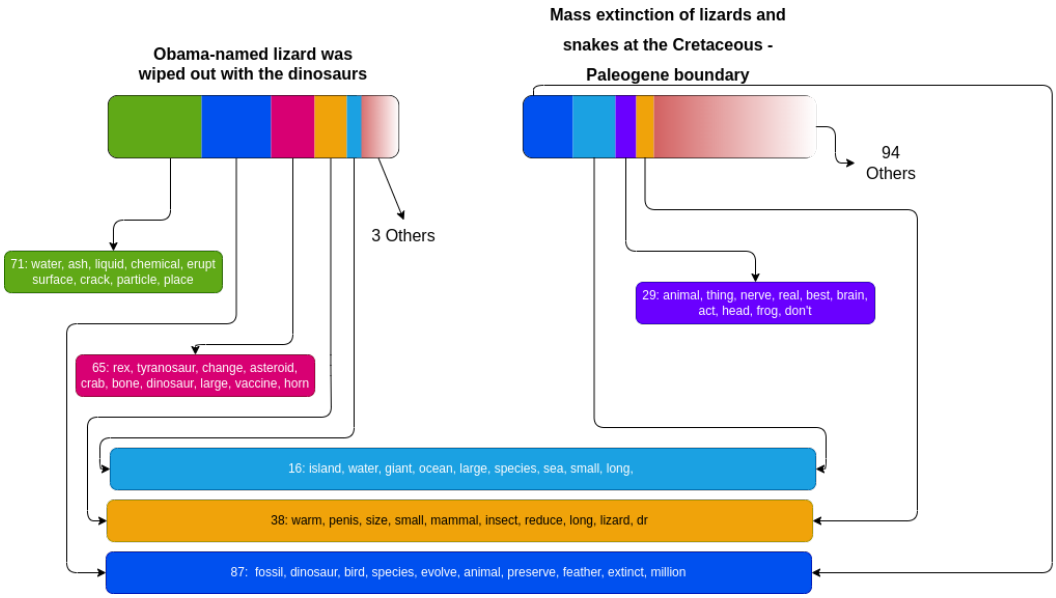


Figure 6.5: A visualisation of document topic distributions for example # 1 in Table 6.4. Topics 16, 38 and 87 are shared by the two documents although the news article is more strongly associated with topic 87 and less strongly associated with topic 16.

the differences in topic distributions.

We expect that the different document lengths and vocabularies lead to the LDA model converging upon news-specific and science-specific topics that are more associated with simplistic and more technical terms respectively. Louis and Nenkova (2013a) showed that news articles typically blend specific details about the scientific work they discuss (e.g. “They found that a gene, known as PIF4, activated the flowering pathway when a certain temperature was achieved.") with general statements that provide few details but provide a structure and narrative that make the articles easier to follow (e.g. “This understanding could allow researchers to modify the plants’ responses to temperature changes, either genetically or through breeding programmes."). Conversely, scientific papers typically contain very few general statements, focusing on the specific details of the work they describe in great technical detail. This phenomenon is also likely to contribute to the generation of document-type-specific topics.

We investigate by inspecting topics that are strongly associated with either news or science documents. We generate topic distributions for all documents and then calculate mean vectors for the two respective collections. We compare the two vectors in order to identify topics that are very strongly associated with either science or news. In Table 6.7 we show that the topic model does indeed learn collection-specific topics and that the associated terms are more technical for topics that strongly relate to science papers and more general/accessible for topics that strongly relate to news articles.

Whilst the ability to compare topic mixture summaries of documents is insightful,

Science		News	
Topic #	**Top 5 Word Stems**	**Topic #**	**Top 5 Terms**
102	mass, gev, wid, lepton, plot	209	fossil, bon, anim, palaeontolog, dinosa
96	psycholog, org, problem, soc, effect	255	eat, risk, diet, diabet, uk
207	jack, gold, specy, afr, euras	115	mat, phys, dark, partic, light
7	fibr, protein, vir, grain, diseas	3	cel, effect, mic, hum, sign
33	dat, ai, mass, vaccin, shel	213	brain, act, ear, just, sign

Table 6.7: Top 5 word stems of the topics strongly associated with either news or science within the $k = 500$ LDA model. Words in news-related topics tend to be more general (e.g. anim[al], partic[le], diet). Words in science topics tend to be more technical/specific (e.g. euras[ian] jack[al], lepton, protein)

it does not provide granular detail about the specific linguistic mechanisms involved in information flow from scientific papers to news articles. LDA's reliance on an underlying bag-of-words language model means that word ordering and document structure are discarded and whilst the topic distribution for each token is known, it's not possible to infer which instance of a token in the source document specifically corresponds to a given topic or token in another document.

LDA models documents as an independent distribution over all topics which means that relationships between topics are not modelled. This prevents the model from making automated inferences about which news topics are likely to be related to, or informed by, science topics. As a human it is easy to see that there is a clear relationship between topics #102 and #115 in Table 6.7 - both topics discuss particle physics. We discuss more recent topic modelling approaches that could be used to model such relationships in future works in section 6.6.2 below.

6.5 Doc2Vec

Doc2Vec provides an information retrieval performance profile that is comparable or better than LDA, yielding the best IR performance for The Guardian and The Daily Mail after the BoW approaches. Similarly to LDA, when the model is trained with a higher dimensional latent vector space, it gets better at retrieval. However, unlike LDA, the documents learned by Doc2Vec are arbitrary projections in a latent space that have no direct relationship with the likelihood of words used (as opposed to topic mixtures which relate directly to topic word distributions in LDA models).

6.6 Conclusion

In this chapter we explored and characterised the differences between scientific newspaper articles and the academic publications that they report upon. We have shown that, despite their shared subject-matter, the different communication goals and target audiences of the two document types lead to very different structures, vocabularies and lengths. We have explored a number of state-of-the-art and baseline document representations with the aim of finding suitable methods for comparing content similarity of long documents.

6.6.1 Findings

Our information retrieval benchmark exercise shows that bag-of-words based representations are a simple but effective method for retrieving relevant documents. It is likely that journalists' tendency to quote highly-specific technical terms mentioned in corresponding scientific documents but less relevant in the wider corpus (e.g. *obamadon gracilis*) leads to high-scoring TF-IDF matches on these terms. Conversely, news articles that paraphrase technical terms and use more generic terminology are less likely to be matched with their corresponding scientific paper. Our results are in line with recent findings that simple TF-IDF information retrieval approaches provide strong baseline performance and even outperforms state-of-the-art retrieval models which have not been fine-tuned to a given problem domain (Thakur et al., 2021). On the other hand, TF-IDF representations do not adequately encode information flow beyond providing confirmation that terms are shared by two documents.

Both LDA and Doc2Vec yield good document retrieval performance which improves as more latent variables (topics and dimensions respectively) are added to the models. Increasing the dimensionality of the models' respective representations facilitates better representation of the broad and varied subjects that are covered by the corpus. Doc2Vec offers limited explainability as documents are represented as arbitrary mappings in a random latent space. Conversely, LDA models can be more directly inspected and the differences in topic utilisation across documents can be examined. However, the topic mixture representations of documents offered by LDA only provide a high level view of how topics are shared and change between scientific papers and news articles. In order to better understand the specific linguistic mechanisms at play when journalists transform and supplement scientific information to make it more palatable for general audiences, a more nuanced approach is needed.

At this stage, there are two opposing strategies that we could explore in order to achieve the required level of nuanced comparison between impactful news and science writing. We could continue to pursue a top-down document-level approach where we attempt to develop models that are both capable of representing long documents as well as aligning and comparing information shared between documents at a low level. The second strategy

is a bottom-up approach in which we attempt to break down documents into more granular components (i.e. paragraphs, sentences), using automated approaches to identify and align comparable fragments of information. The remainder of this thesis focuses on the latter, bottom-up approach. However, here we briefly outline some possible future directions for of top-down approaches.

6.6.2 Future Work

Modelling and comparing long documents remains a significant challenge within Natural Language Processing at time of writing. Whilst LDA (Blei et al., 2003) and Doc2Vec (Le and Mikolov, 2014) remain popular for many use cases (Karas et al., 2022; Prabhakaran et al., 2016), recent works have focused on producing context-sensitive representations of sentences and paragraphs using attention-based models (Section 2.6.8). Historically, the significant compute and memory costs associated with these models has meant that these experiments are limited to short documents and excerpts of documents. However, new SOTA models such as Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2021) apply attention to longer documents by using heuristics to limit the scope (and thus memory and compute requirements) of their attention mechanisms to important areas of documents. Subsequent studies (Jiang et al., 2020; Zhou et al., 2022) have also demonstrated the efficacy of using task-specific heuristics for attending to long documents e.g. using search terms to determine which words in target documents may be attended to within an information retrieval model. Future work could explore strategies for building long transformer models that are sensitive to concepts that span news-science document pairs by using shared terms (Jiang et al., 2020) or topic information to guide transformer attention. Such models may be able to produce rich, dense representations of models that embed useful context about respective structures of the documents they represent as opposed to the simpler but less informational bag-of-words representations used in this work.

The underlying bag-of-words representations relied upon by LDA also remains one of its major drawbacks, providing very limited information about document structure. This prevents meaningful interpretation of document topic mixture representations at a granularity below document level. Whilst this may not be an issue when working with shorter texts (e.g single paragraphs), it precludes us from properly examining the alignment between sections of news articles and scientific papers. Moody (2016)'s lda2vec approach integrates global document-level Dirichlet topic distributions into Mikolov et al. (2013a)'s word-level skip-gram optimisation process to provide vector representations that encode both local and global semantic information for a given document. lda2vec can be used to generate and compare topic distributions for sections of documents. Future work could examine whether lda2vec provides representations that are rich enough to be used to identify and compare semantically similar areas of paired documents, or even areas chosen by discourse analysis or other automated means. Recent work has also shown that topic information from LDA

models can be used to supplement contextual embeddings from transformer models in order to improve task performance (Peinelt et al., 2020), the application of which may represent a promising direction for future work in document comparison.

As noted above, LDA models are also unable to model relationships between topics, representing documents as mixtures of independent topics. This prevents us from understanding how topics that predominantly feature in news articles and relate to simpler words and topics that predominantly feature in scientific articles and relate to more complex words may relate. Lafferty and Blei (2006) present an extension of LDA called Correlated Topic Models (CTM) which learns the dependencies between topics. Future work could explore whether CTM is able to improve information retrieval performance and whether it can yield better insight information transfer between topics.

Future work could also explore methods for better explaining information transfer such as neural rationale models. Neural Rationale models, such as UNIREX (Chan et al., 2022), jointly learn a primary task (e.g. sentiment analysis, review scoring) and also generate a text-based rationale of their final decision, made up of words or sentences from the input document. By using inputs based on robust document representations (e.g. Longformer embeddings discussed above), it may be possible to train models to predict the semantic similarity between pairs of news articles and scientific papers whilst automatically identifying and highlighting the most similar regions of the two documents.

6.6.3 What's Next?

In the next chapter we focus on bottom-up strategies for intra-document ranking of information and prominent excerpts from news articles and scientific papers as well as strategies for examining the semantic similarity of these excerpts.

7

Prominence & Semantic Similarity in News Articles & Scientific Papers

"Read carefully anything that requires your signature. remember the big print giveth and the small print taketh away."

H. Jackson Brown Jr.

7.1 Introduction

Following our document-level experiments in chapter 6 and our evaluation of their corresponding limitations, we seek to to understand lower-level mechanisms and processes involved in information transfer from scientific papers to news articles.

Beyond secondary education, the vast majority of people understand science via their interactions with the news. This places a great responsibility upon the shoulders of journalists and their editorial teams who "...act as gatekeepers to the scientific world, controlling what information reaches the public eye and how it is presented" (MacLaughlin et al., 2018). In order to write a scientific news article, journalists must sift through the detailed scientific papers pertaining to a given issue and decide what information to include in their article. Some details are likely to enthuse and excite their audience whilst others may be too technical or not interesting. Journalists' choices of narrative frame and style (as explored in section 6.2.3) also affects the extent to which a particular work is highlighted and the level of detail devoted to it.

In this chapter we address **RQ4** (section 1.2) and **CH2** (section 1.3.2) as we focus on the how the representation of scientific work in news articles can affect the generation of comprehensive impact for that work. We consider the nature of scientific news coverage: is it mentioned in passing? Is it part of a wider article about the state of a field? Or, is the scientific work in question the main component of the article? To achieve this we define a measure of *prominence* in news articles, and we develop a method to determine whether a scientific paper is prominently featured in the news article. Subsequently, we make use of a publicly available[1] system for scientific discourse segmentation (Liakata et al., 2012b) to help us characterise the scientific contributions of each sentence within scientific papers from the Linked HarriGT Corpus (discussed in Chapter 5). We inspect the semantic similarity between each discourse category and prominent sentences in news articles with the goal of understanding what kinds of scientific detail journalists are most likely to focus on in their articles. Finally, we experiment with the subset of the Linked

[1]Website: `http://www.sapientaproject.com/`, Public API: `https://sapienta.dcs.aber.ac.uk/`

HarriGT Corpus with connections to REF impact case studies to understand the extent to which prominent links to different kinds of scientific discourse can affect comprehensive impact.

This chapter is based on our pre-print entitled *'Measuring Prominence of Scientific Work in Online News as a Proxy for Impact'*. the work was carried out in early 2019 and was uploaded to arXiv on 28 July 2020. It can be found at `https://doi.org/10.48550/arXiv.2007.14454`. Since this work was carried out, new developments in summarisation, pre-trained language models and semantic textual similarity that could further enhance our findings in future work. We further discuss some possible directions for future work in section 7.5.1.

The specific contributions of this chapter are as follows:

1. We define the concepts of prominence and salience and their relationship to one another in the context of scientific news articles.

2. We introduce SimSemRank, a PageRank-inspired (Page et al., 1999) sentence scoring algorithm which we use to find prominent sentences in news articles in order to align them with relevant scientific discussion.

3. We explore the use of scientific discourse analysis to discover which parts of a scientific paper are most relevant.

4. We run a series of experiments using different state-of-the-art text representation methods to examine semantic similarity between prominent excerpts of news articles and sentences from different parts of scientific papers (Goals, Background, Methods, Outcomes)

5. We demonstrate a strong statistical relationship between prominent discussion of scientific work in news articles detail and comprehensive impact in terms of REF impact score.

7.2 Prominence and Salience

Prominence refers to the importance assigned to text by its author. It is a somewhat under investigated area of computational linguistics and even traditional linguistics and discourse (Becker and Egetenmeyer, 2018). Prominence has recently received some attention in the domain of argumentation mining with (Wachsmuth et al., 2017) taking the position that prominence may be considered “a product of popularity" rather than a measure of intrinsic quality or importance. In the same spirit Boltužić and Šnajder (2015) focus on repetition as a key indicator of prominence in order to automatically identify common arguments in online debates. They use clustering to group semantically similar arguments

together before manually analysing and labelling clusters. This approach gives some insight into prominence of arguments found in online debates but does not offer any narrative on the intrinsic importance of the arguments presented.

In contrast to prominence, salience corresponds to the intrinsic importance of a unit of text within a document regardless of its presentation. Boguraev and Kennedy (1997) uses salience as a way to measure the “aboutness" of a document.

Figure 7.1: An illustration of prominence and salience; the author wants the reader to focus on the prominent information about living your best life but the salient and important information about the APR is displayed in small print.

Whilst prominence and salience are separate concepts, the distinction is subtle and is often tricky to spot. This is because, as a rule of thumb, authors are trying to communicate the information that they consider to be most important (*salient*) in the most clear and *prominent* way. However there are some cases where salience and prominence are not aligned. In an advertisement for a credit card, the small-print about interest rates and terms of repayment provides salient information about the product but is not normally presented prominently as illustrated in Figure 7.1. Conversely, “clickbait" articles often present incomplete or misleading information (e.g. “scientists cure cancer") very prominently in the headline in order to capture the attention of the reader whilst the salient information (e.g. “in mice”) is often buried away near the end of the article. By examining prominence within news articles and scientific papers we can understand which aspects of scientific work journalists think are most important to present to their readers and how this aligns with the way that the original scientists presented their work.

We hypothesise that scientific papers which are discussed prominently in a news article are likely to generate more comprehensive impact than scientific works that are mentioned in passing. When an article’s sole focus is on a new high efficacy treatment for cancer, it signifies that journalist, who likely has a scientific background themselves, thought the work was noteworthy. The act of publishing such articles may also generate comprehensive impact in itself, in the form of increased awareness of the work from the authors’ peers and subsequent opportunities for new funding and career progression.

7.2.1 Prominence in News Articles

The layout and format of news articles often leads to prominent information in news articles being repeated a number of times (Grenander et al., 2019) as illustrated in Figure 7.2. The headline, subheading (often referred to as the lede or lead), and following paragraph often contain similar information, each providing slightly more detail in order to progressively engage readers. Articles will then typically follow the structure observed by Louis and Nenkova (2013b), interweaving specific scientific detail, which may relate to the headline and lede, with general sentences which strongly relate specific sentences back to the general topic.

Daily Blah.com

Mysterious Fossil Found in Supermarket Car Park

The fossil, found by a customer on sunday has paleontologists stunned.

Photography by Jose Luis https://www.pexels.com/photo/brown-dinosaur-mascot-2564846/

The global paleontology community has been shaken this week after a new unknown fossil was discovered in a supermarket car park in Milton Keynes. The specimen, thought to date back to the cretaceous period was discovered by Dan Walker, 32 as he left ASDA at 3pm with his weekly food shop.

Figure 7.2: News articles often repeat the most prominent information. The headline, lede (in bold above photo) and first few sentences are likely to contain very similar information, each more detailed than the previous.

In order to automatically identify prominent excerpts of news articles, we explore extractive summarisation approaches. Extractive summarisation algorithms generate sum-

maries of documents by extracting and combining the most salient excerpts of text from an original full document. We investigate the possibility of re-purposing such techniques in order to identify the most prominent rather than salient excerpts of texts from news articles, exploiting the formulaic structures of news articles.

LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are two popular and related methods for extractive summarisation that use graph centrality as a way to understand the salience of a text unit such as a sentence within a document; the intuition being that the more central an excerpt is to the parent document's general subject, the more salient it is. In TextRank and the continuous variant of LexRank, documents are represented as a fully connected graph where each sentence is a vertex and edges represent the semantic similarity between two sentences. Both approaches use a ranking approach based on PageRank (Page et al., 1999) to identify those sentences which are most central to the document graph and prioritise them for inclusion in the summary.

Given the formulaic structure and layout of news documents and the relative importance of a central narrative (Louis and Nenkova, 2013b), we hypothesize that TextRank and LexRank can be used to take into account the prominence of text in news articles. LexRank and TextRank scoring is based on the similarity of each sentence with respect to all others, where similarity is defined by Erkan and Radev (2004) and Mihalcea and Tarau (2004) respectively. Information repeated in multiple sentences (as described above) will boost the relative similarity of these sentences and thus the score of all sentences that discuss the repeated information.

We use a modified version of continuous LexRank to create a method for measuring and ranking all sentences in news articles from the Linked HarriGT Corpus in terms of the prominence of the information they contain. We call this method "SemSimRank".

For a given news article document D containing S sentences, pairwise semantic similarity $\theta(s_i, s_j)$ between all sentences $s_i; i \in \{0..S\}$ and $s_j; j \in \{0..S; j \neq i\}$ is stored in adjacency matrix E and used to create a fully connected weighted graph $G(D, E)$ as per Figure 7.3. Edges are normalised row-wise to help with ranking convergence.

We then use PageRank (Page et al., 1999) with damping factor α, max iterations N and convergence threshold δ to produce a set of rankings P for all sentences $s_i \in D$ (Mihalcea and Tarau, 2004; Erkan and Radev, 2004). This process is described in Algorithm 1.

The key difference between our ranking approach, which we designate SemSimRank, and the TextRank and LexRank approaches is how pairwise semantic similarity $\theta(s_i, s_j)$ is calculated. LexRank uses the cosine similarity of TF-IDF vectors (Erkan and Radev, 2004) whilst TextRank simply uses the word overlap between the two sentences (Mihalcea and Tarau, 2004). We evaluate a number of sentence-wise semantic similarity functions $\theta(s_i, s_j)$ based on text representations explored in section 7.3.3 in order to understand the information transfer between the most prominent sentences in the news articles

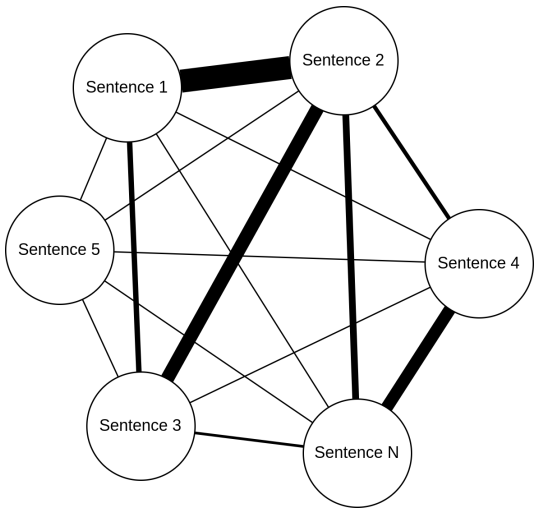


Figure 7.3: A fully connected, undirected, weighted graph representing the pairwise semantic similarity of sentences in a document. Edge weight is equal to the semantic similarity between vertices.

and excerpts of the linked scientific papers. For each article we select N top sentences with the highest PageRank for comparison with the scientific content.

7.2.2 Prominence in Scientific Papers & Scientific Discourse Modelling

Scientific papers also follow a formulaic structure like news articles. However, as discussed in section 6.2, scientific writing is generally longer and more complex than news writing with a focus on precision and reproducibility over readability and reader engagement. Scientific papers do not use headlines and leads to present prominent information but scientific authors may make information that they consider to be important prominent by strategically placing it in certain parts of their document (i.e. the abstract, discussion and conclusion sections) so that readers are more likely to notice key ideas during their initial reading or scanning of the work. The structure of scientific papers is likely to be too complex for SimSemRank without modifications taking into account document structure and length. However, the domain of scientific discourse which focuses on identifying the functional role of units of text within scientific articles is well explored and provides a useful starting point for understanding prominence within scientific work.

Argumentative Zoning (Teufel, 1999, 2010) (AZ) is a seminal piece of work in this area which defines an annotation scheme for rhetorical structures within scientific papers. AZ aims to "...[capture] the attribution of intellectual ownership in scientific articles, expressions of authors' stance towards other work, and typical statements about problem-solving processes." Variants of AZ have been applied to both scientific articles and article abstracts. Subsequently, Liakata et al. (2010) introduced an annotation scheme complementary to AZ, the Core Scientific Concept (CoreSC) annotation scheme. CoreSC aims

Algorithm 1 SemSimRank Sentence Selection Algorithm

Input

s	An array of sentence representations to be compared
θ	The similarity function used to evaluate pairwise similarity between sentences
N	The maximum number of iterations to carry out
δ	The minimum change in weights required at each iteration before stopping early
d	Damping factor hyperparameter, usually set to .85

Output

P	An array of real numbers indicating sentence ranking with the same length as s

let $S = |s|$
for i in [0..S] **do**
  for j in [0..S] **do**
    $E_i^j = \theta(s_i, s_j)$ ▷ calculate pairwise similarity for all sentences in s
  end for
  $E_i = E_i / \sum_{j=0}^{S} E_i^j$ ▷ Normalise similarities for each sentence wrt itself
end for
let $P_i = 1/S$ for i in S ▷ init p for all sentences based on the length of s
while num iterations $< N$ **do**
  for i in [0..S] **do**
    ▷ Calculate ranking for sentence i based on the summed products of respective
    ▷ rankings and similarities to s_i for other sentences
    $P_i^{new} = \frac{1-d}{S} + d \sum_{j=0}^{S} \frac{P_j}{\sum_{k=0}^{S} E_j^k}$
  end for
  if $\sum_{i=0}^{S}(||P_i^{new} - P_i||) < \delta$ **then**
    break/stop iterating
  else
    $P = P^{new}$
  end if
end while
return P ▷ Return the final ranking of sentences

to capture the content and structure of a scientific investigation rather than its rhetorical narrative and related arguments. Examples of CoreSC discourse labels are 'Background', 'Hypothesis', 'Methodology', 'Conclusion'.

CoreSC and AZ provide complementary metadata that is useful for a wide range of tasks including summarisation of scientific papers (Teufel and Moens, 2002; Liakata et al., 2013), information retrieval (Teufel, 2006; Duma et al., 2016) and prediction of a publication's communication style (Ravenscroft et al., 2013). We hypothesise that the functional information provided by scientific discourse annotations is likely to be useful for determining the prominence of sentences therein and that certain discourse categories are likely to be more prominent than others. For example, authors are likely to emphasise their novel contributions and findings by featuring them clearly in their conclusion and abstract. Details such as experiment parameters are still important but less central to the document, likely appearing in the method section of a paper.

The authors of AZ and CoreSC advocate for their combined use to leverage their individual strengths (Liakata et al., 2010). In our work, we utilise CoreSC annotations owing to the presumed relation between CoreSC content-based categories and our definition of prominence and to its larger training corpus and publicly available automated SAPIENTA classifier (Liakata et al., 2012b). The latter can be used via a web service[2]. It is also advantageous that SAPIENTA is trained primarily on biomedical papers which make up the majority of our linked scientific content due to journalists favouring these kinds of papers (MacLaughlin et al., 2018). We use SAPIENTA to assign CoreSC scientific discourse labels (e.g. Results, Hypotheses, etc.) to each sentence in each scientific paper within the Linked HarriGT Corpus.

Some types of CoreSC scientific discourse categories are quite rare (e.g. Hypotheses, Objectives) and therefore we aggregate the 11 discourse categories together into 4 CoreSC Groups: Background (containing Background and Motivation), Goals (containing Goals, Objectives and Hypotheses), Method (containing Method, Experiment and Model) and Outcomes (containing Observations, Results and Conclusions).

Using different methods described in Section 7.3.3, we measure the semantic similarity between the extracts from the news articles that are considered prominent by the SemSimRank algorithm and the linked scientific articles, considering a pairwise similarity with each sentence extracted from the scientific paper.

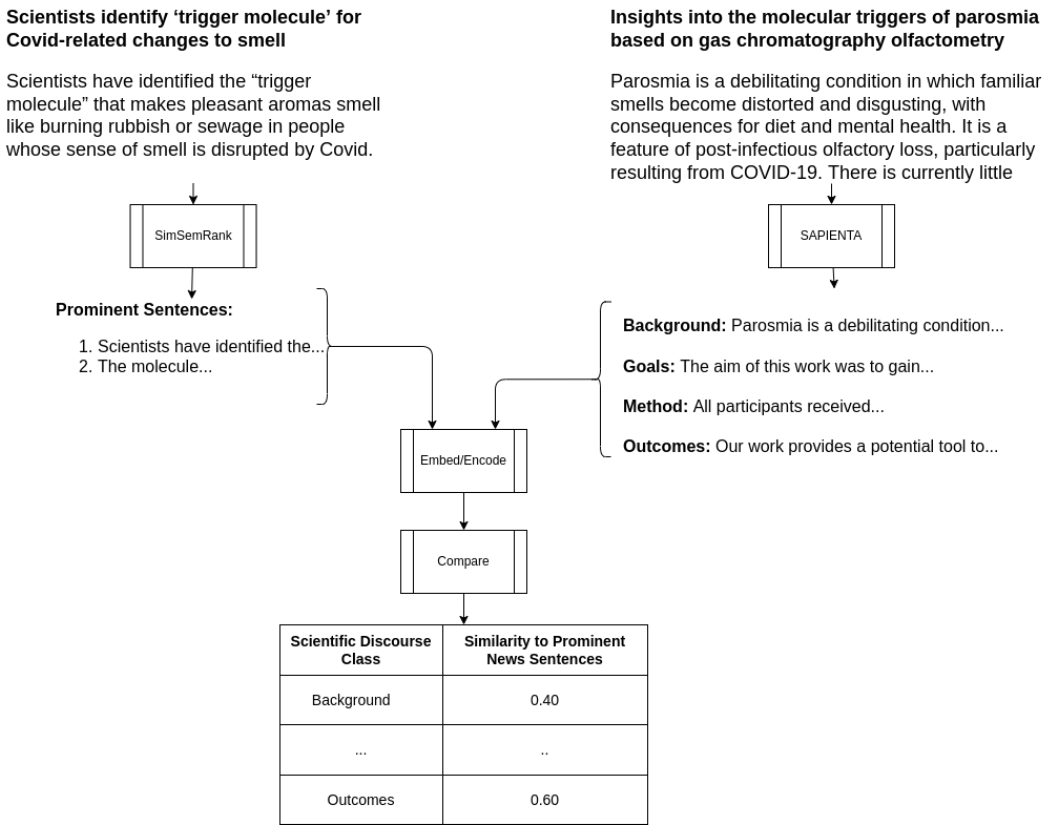


Figure 7.4: An overview of the prominence experiment process: news articles are put through SemSimRank and scientific papers are put through SAPIENTA. Resulting sentences are encoded and pair-wise comparisons made between prominent news sentences and sentences from scientific papers using methods from section 7.3.3. Mean similarities between prominent sentences and each scientific discourse class are reported.

7.3 Methods

7.3.1 Overview

We next aim to identify which aspects of the scientific work are discussed most prominently in each news article. We combine the techniques discussed above to evaluate the semantic similarity between prominent news sentences and different types of scientific sentences based on their discourse label. Figure 7.4 shows the preprocessing applied to each type of document before encoding and comparing the resulting sentences. For a given news and science document pair from the Linked HarriGT Corpus, the news article is given sentence-based prominence rankings by SimSemRank in order to identify the most prominent news sentences. The full text from the scientific paper is annotated with CoreSC labels using SAPIENTA and the sentences are then grouped into the 4 discourse categories described in section 7.2.2 (Background, Goals, Method and Outcomes).

[2] `http://sapienta.papro.org.uk/`

All sentences are encoded using one of the representations described in 7.3.3 and finally sentence-wise similarity is calculated between each of the top N sentences and each of the sentences in the scientific paper via an appropriate similarity metric (cosine distance or Jenson Shannon Distance). Results are aggregated by CoreSC discourse category in order to see which of the CoreSC categories is most similar to the prominent news sentences and thus, which section of the scientific paper is most prominently presented within the news article. In most cases, the pairwise sentence comparisons yield a low number of highly similar sentence pairs and a large majority of sentences that have a low semantic similarity. Given that the median of these similarity distributions is often close to zero, we report the mean similarity in our results. We also found that the smoothing effect of averaging our results was preferable over taking the maximum value which is even more prone to noise from outliers.

7.3.2 Sentence Selection

We use our SemSimRank algorithm to select the most prominent sentence from each newspaper article for pairwise comparison against each sentence from the linked scientific papers. We also measure our SemSimRank algorithm against two baseline approaches:

- Our “First Sentence" baseline simply takes the first full sentence from each newspaper article as the most prominent sentence. Since newspaper articles often start with an overview of their content, this is a simple but often effective strategy.
- Our “Random Sentence" baseline uniformly at random selects one of the sentences from the newspaper article as the most prominent sentence. We preserve the random choice across all experiments so that the results can be compared.

7.3.3 Text Representation & Semantic Similarity

We aim to identify a baseline approach for determining the semantic similarity between the news excerpts that are prominent and CoreSC passages extracted from linked scientific works such that it is possible to measure knowledge transfer between the documents. We compare a number of common feature representations and similarity metrics in order to find the most suitable for this task. This is especially important since, as discussed in Chapter 6, most semantic similarity tasks such as STS (Agirre et al., 2012) and SICK-R (Marelli et al., 2014b) compare sentences from documents with similar lengths and vocabularies, and so these representations and metrics may prove to be unsuitable for our task.

We use an initial bag-of-words (BoW) count vector representation as a baseline for sentences from both scientific papers and newspaper articles. We combine a static English stopwords list, a lowercase filter and a minimum word length check of 3 characters or more in order to identify sets of relevant unigram features. The combined corpus vocabulary is

very sparse. However, since we only consider pairwise semantic similarity between linked documents we need not attempt to build a global vocabulary model for the corpus, avoiding problems associated with a high dimensional feature space. Instead we construct a local feature space on-the-fly for each document pair allowing us to avoid feature pruning and retain all valid unigram features within specific document pairs. Using count vectors rather than simpler binary one-hot encoding allows us to account for word repetition, which may be indicative of prominence, within our similarity calculations.

As discussed in Section 2.5, BoW feature representations are unable to account for semantic relationships between distinct words. Therefore, they may fail to successfully encapsulate relationships between two related documents that use different vocabularies with minimal word overlap; even when discussing the same subject matter, scientific papers and newspaper articles are typically written in different grammatical styles and vocabularies for scientists and laypeople respectively (see section 6.2.1). This motivates us to consider alternative sentence representations that are more sensitive to semantic relationships between texts. We use pre-trained static GloVe embeddings to calculate mean vector representations which have been shown to provide a strong baseline for sentence similarity and matching tasks (Shen et al., 2018; Arora et al., 2017). We also use BERT (Devlin et al., 2019) to encode each sentence in order to provide context sensitive embedding and word disambiguation as well as support for rare and out-of-vocabulary terms via its built in wordpiece subword encoder (Schuster and Nakajima, 2012)

We use cosine distance for measuring the similarity between neural sentence vectors from pairs of news articles and scientific papers. For our BoW count vector representation we use Jensen-Shannon Distance (defined as square root of Jensen-Shannon divergence) which is more suited to sparsely populated non-normalised integer word count vectors.

7.3.4 Prominence & Comprehensive Impact

We examine whether there is an association between how prominently different types of scientific discourse are discussed in news articles and levels of comprehensive impact generated by using the REF impact scores from the Linked HarriGT Corpus (subset F in Figure 5.6) as a proxy. We carry out the process outlined above on documents in Harri GT Corpus Subset F which contains 108 scientific papers linked to 140 news articles and 103 REF impact case studies and subset D which contains 5655 news articles linked to 3553 scientific papers. We assume that subset D, which is not linked to any REF case studies, provides a representative and broad sample of news articles that refer to scientific work in a variety of different contexts and for a number of different reasons. On the other hand, we assume that news articles and scientific papers in subset F are representative of UK universities' self-selected most impactful work and will have a distribution of comprehensive impact levels skewed towards higher impact than subset D.

We evaluate pairwise semantic similarity for excerpts from all known linked pairs

of newspaper articles and scientific papers using both F and D paper collections and each of the above-described feature representations and respective similarity metrics. A statistically significant uplift in semantic similarity between prominent science and news excerpts from the fully linked collection F versus the collection D would indicate that our notion of prominence is related to the comprehensive impact of scientific papers featured in the news.

7.4 Results and Discussion

Our results are shown in Figure 7.5. For each CoreSC group we show the observed mean semantic similarity for sentences from pairs of REF-Linked documents and sentences from pairs of Non-REF-Linked. Our findings suggest that for these collections scientific work is more prominently discussed in REF-Linked newspaper articles than Non-REF-Linked articles. For the 'Outcomes and 'Methods' CoreSC groups in particular, almost all of the experiments in Figure 7.5 show stronger similarity for the REF-Linked documents than the Non-REF-Linked documents. However, both the encoding approach set and newspaper sentence selection approach have a significant effect on what is captured.

All BoW-JSD approaches (top row Figure 7.5) consistently capture a significant positive difference in prominence for REF-linked documents across all 4 CoreSC discourse groups. The clearest demonstration of this relationship is from BoW-JSD + SemSimRank but BoW-JSD + FirstSentence also captures this difference particularly well. A KS-2 test comparing the outputs of BoW-JSD + SemSimRank and BoW-JS + FirstSentence for sets E and F confirms that the difference is significant ($p < 0.05$).

The relative success of the BOW-JSD + FirstSentence approach may be down to the large number of news articles within the linked corpus that begin with an informative lede which briefly summarises the key goals and outcomes from the linked scientific document for the reader as discussed in section 7.2.1. However, there are also a significant number of articles that do not start in this way, engaging readers in a more chatty, informal style (see Table 7.2). In these instances, BoW-JSD + SemSimRank typically outperforms BoW + FirstSentence by identifying a more relevant summary sentence in the newspaper article. Table 7.1 shows examples of FirstSentence and SemSimRank selected sentences from the same article.

Strong similarity between two BoW representations of sentences necessitates a high degree of exact word overlap between them. Since we make use of a comprehensive stopwords list [3], it is likely that these overlapping terms are salient features that correspond to direct quotation, technical terms and named entities mentioned by both documents. We posit that the higher average semantic similarity between REF-linked sentence pairs suggests that REF-linked articles are more likely to quote or make use of technical terms used

[3] https://github.com/scikit-learn/scikit-learn/blob/36958fb240fbe435673a9e3c52e769f01f36bec0/sklearn/feature_extraction/_stop_words.py

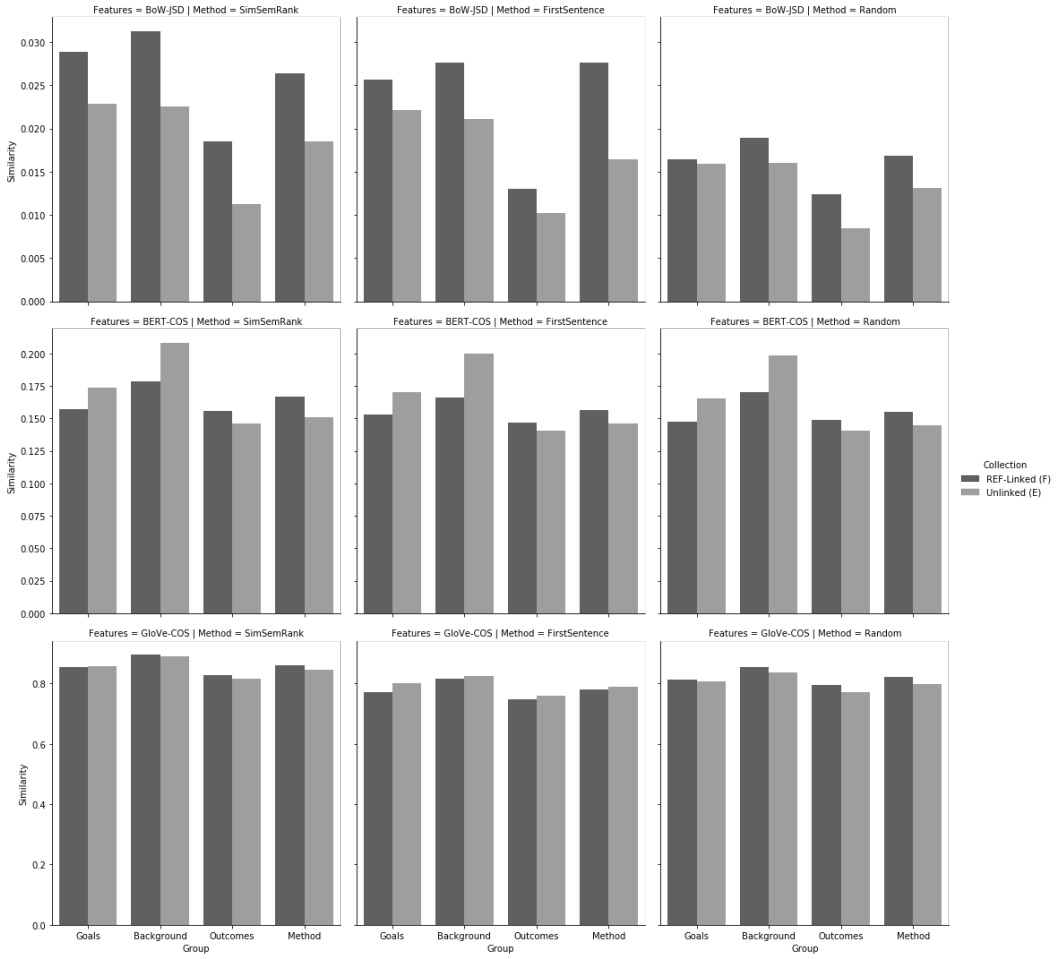


Figure 7.5: Mean Pairwise % Difference in Semantic Similarity between newspaper articles and scientific papers that are associated with REF case studies versus those that are not. y scale is not comparable across feature sets (rows) due to the use of different semantic similarity metrics.

News Article - First Sentence	News Article - SemSimRank Sentence
It isn't often that science and pop culture overlap, but the two fields are in agreement when it comes to the familiar trope of the forgetful stoner.	But with the recent changes in drug policy, the chances are that more people will be smoking cannabis than ever before, and the more potent and more popular high-THC/low-CBD marijuana that is available today will increase their risk of dependence.
Writing in the British Medical Journal they say a 15% cut in consumption could save 8.5 million lives around the world over the next decade.	The report - by researchers at the Universities of Warwick and Liverpool - says that after cutting tobacco consumption, getting people to eat less salt would be the most cost effective way to improve global health.
Several prehistoric creatures developed elaborate body traits in order to attract members of the opposite sex, according to new research.	Co-author Dr Dave Martill from the University of Portsmouth said: "Pterosaurs put even more effort into attracting a mate than peacocks whose large feathers are considered the most elaborate development of sexual selection in the modern day".

Table 7.1: Example pairs of First Sentences and corresponding "most prominent" sentences discovered by SemSimRank from the same articles.

within the scientific papers they discuss.

News Article First Sentences - High Semantic Similarity	News Article First Sentences - Low Semantic Similarity
One in three adults aged over 65 in England have difficulty understanding basic health-related information, suggests a study in the BMJ.	Like many patient groups, the Alzheimers' Society isn't happy with the state of scientific research..
Acne drug not found to increase suicide risk	It has all the makings of a pub quiz teaser: what do Barack Obama, Emma Watson, Jake Gyllenhaal and the British TV presenter Fiona Bruce have in common?
University College London researchers found a 3.6% decline in mental reasoning in women and men aged 45-49.	Lately, it seems as if everyone is anti-antidepressants.

Table 7.2: Example 'First Sentence' extracts that are helpful (left) and not helpful (right) for prominence task.

BERT-COS consistently asserts that Goal and Background sentences in the Non-REF-Linked collection are more semantically-similar than those from the REF-Linked collection for all three news sentence selection strategies. It is likely that articles which prominently feature a scientific paper's background statements are less likely to generate compre-

hensive impactful for that paper (e.g. they may refer to a scientific work in passing to lend credibility to a general observation). However, BERT's subword embedding mechanism may also struggle with rare, specialised words, entity names and numbers (Wu et al., 2016), which are important features in our task.

The GloVe-COS approach is the least consistent of the methods, generating semantic similarities that are barely discernable for the two document collections. As discussed in Chapter 6, the REF-Linked HarriGT Corpus contains a very sparse and varied vocabulary and many highly specialised terms. Our experiment uses pre-trained GloVe embeddings trained on the CommonCrawl corpus and we hypothesise that a number of the important terms in these documents, many of which are named entities and noun-phrases, are likely to be missing from the model vocabulary.

All three document encoding strategies consistently show Goal and Background sentences as having the strongest semantic similarity to prominent news sentences and Outcomes as having the weakest. This may be due to vocabulary gap between scientific and journalistic explanations of results; findings in scientific papers are often phrased very carefully and specifically which must be simplified for a general audience to understand. For example in one study[4] the results are summarised prominently in the news article as "The authors report that people dependent on the drug – both healthy individuals and patients with schizophrenia – show impairments in memory compared with healthy volunteers and non-smoking schizophrenia patients." whereas the scientific paper concludes that "...our findings suggest that a remote CUD [cannabis use disorder] may be associated with differences in WM-related subcortical morphology in both control and schizophrenia subjects".

7.5 Conclusion & Future Work

In this chapter we have explored prominence in news articles and scientific papers and the effect that prominent journalistic discussion of scientific work has on the comprehensive impact that the work receives. By taking advantage of the document structures of news articles and scientific papers we are able to identify prominent excerpts of news articles and group excerpts from scientific papers by discourse functions. Our sentence similarity experiments suggest that prominent discussion of scientific outcomes and methodology is more likely in scientific news articles linked to REF impact case studies. This is intuitive since the outcome sections of a scientific work usually feature prominent statements by scientists about the real world impacts and ramifications of their work. Although our results are promising, similarity of sentence representations is imprecise and does not necessarily imply semantic similarity, nor does it tell us much about what makes a sentence likely to convey comprehensive impact.

[4] https://www.theguardian.com/science/blog/2014/jan/17/cannabis-memory-loss-cbd

As discussed in section 6.3.2.5, news articles and scientific papers that they discuss often have large and disjoint vocabularies and journalists typically make complex scientific work accessible to a general readership through creative language use, summary and paraphrasing. However, the success of BoW based representation methods in both this chapter and Chapter 6 suggests that journalists also make extensive use of quotation and verbatim repetition of important terms from scientific papers. The difference in mean similarity between REF-Linked pairs of sentences and those not linked to REF could suggest that prominent quotation or use of specialist terms in news articles is linked to more comprehensive impact. On the other hand, REF-Linked news articles are also subject to selection bias by the authors of impact case studies who must be aware of articles in order to include them. It may be the case that news articles which directly and prominently quote scientific papers were more available to scientists than those that do not; for example, the authors of the scientific work and impact case study may have been interviewed by the journalist who wrote the article article or it may simply be easier to find these articles using online search engines due to the overlap of specific vocabulary. The limited size of our labelled data and lack of comprehensive impact metrics that could be easily applied to new documents also present significant barriers to further exploration of this phenomenon.

When scientific writing is too detailed or technical for general readers, journalists often summarise long excerpts into a single sentence which highlights the details that their readership may find interesting and omits anything that's too complex. It stands to reason that aligning sentences with a single counterpart sentence from the other document based on their similarity is unlikely to adequately capture the semantic relationships between the two documents. Therefore, aligning document content without relying on sentence-wise similarity is a high priority.

7.5.1 Future Work

There are a number of clear directions that future work could explore. Firstly, future work could explore whether language models that have been pre-trained using scientific writing, such as SciBERT (Beltagy et al., 2019), are able to provide better semantic comparison than BERT which is normally pre-trained on large corpora of books and literature and english language wikipedia (Devlin et al., 2019). Newer transformer models such as RoBERTa (Liu et al., 2019b) which outperform BERT at a range of NLP tasks may also yield better results. Alternatively, separate pre-trained embedding spaces could be used to embed news and science documents and these embeddings could be aligned to facilitate direct comparison (Conneau et al., 2018; Kutuzov et al., 2018). Future investigation into sentence-wise content alignment could consider models trained specifically at Semantic Textual Similarity (as defined in section 2.8.4), such as XLNet (Yang et al., 2020), which may yield results that better represent the semantic similarity between paraphrasing sentences that are worded very differently. Ways to adapt models trained on commonly used STS datasets such as the

Quora Question Pairs [5] and SICK (Marelli et al., 2014a) to a scientific domain could also be explored in order to improve STS model performance when working with rare and specialist wording. Approaches for automatically aligning journalistic summaries with scientific prose could draw inspiration from multi-hop reading comprehension in which models combine disjoint pieces of textual evidence in order to complete a task (Welbl et al., 2018). Shortly after this work was carried out, Grenander et al. (2019) developed a model that aims to generate extractive summaries of news articles (i.e. extractive in that the text in the summary is sourced entirely from the document as opposed to generated) which are not influenced by lede bias (discussed in Section 7.2). Future work could investigate whether an inversion of this model could be used to identify prominent sentences and more effectively than SimSemRank. Finally, more recent Scientific Discourse Labelling approaches such as those found in Brack et al. (2021) who use multi task learning across a number of scientific discourse tasks including CoreSC (Liakata, 2010) to train a state-of-the-art sentence classification model could be explored.

7.5.2 What's Next?

In the remainder of this thesis we focus on the two key challenges outlined above. Firstly, in Chapter 8 we explore an alternative strategy for aligning semantically similar excerpts of news articles and the scientific papers that they discuss by focussing on the entities that make up the subject matter of each sentence. Subsequently, in Chapter 9 we devise a comprehensive impact metric based on linguistic features of the pairs of articles which can be used to quantify and compare perceived comprehensive impact between two sentences.

[5] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

8

Cross-Document Cross-Domain Coreference Resolution for Conceptual Alignment between News and Science

"Similar souls wander in the similar places! They may not know each other, but often they touch the same winds, they step on the same leaves, their looks are lost in the same horizons!"

Mehmet Murat ildan

8.1 Introduction

In Chapter 7 we aligned sentences across news articles and scientific papers by directly comparing their feature representations. We discussed the fact that these approaches may not fully capture semantic relationships between pairs of documents both due to the fidelity of the representations themselves and because of the asymmetry between the length and detail of documents and the information therein.

We know that exact 1:1 semantic alignment between sentences may not be possible since information conveyed across multiple sentences in a scientific paper may be summarised in a single sentence in a news article (as discussed in section 7.2.1). However if two sentences discuss the same entity, be it a novel chemical compound; manufacturing technique; species of animal or an as-yet unnamed star observed with a new telescope, we can be reasonably sure that the sentences are related. Furthermore, capturing co-referent mentions of the same entity which have very different or even opposing semantics may help us to better understand how journalists transform scientific writing in order to make it more appealing to readers or generate more comprehensive impact.

In this chapter we tackle **RQ5** (section 1.2) and challenges **CH1** (section 1.3.1) and **CH2** (section 1.3.2). We devise a new task setting called Cross Document, Cross Domain Co-reference Resolution (CD^{2}CR) which allows us to identify when the same entity is mentioned in a news article or a scientific paper in order to align sentences that discuss the same topics, disjoint vocabulary notwithstanding. We develop a CD^{2}CR annotation scheme, guidelines document and tool, the latter of which uses a neural language model to suggest annotations, reducing manual effort involved in annotation and accelerating the data collection process. We use our tool to assemble a large, challenging cross-document co-reference resolution corpus based on the Linked HarriGT Corpus (as described in Chapter 5) and we train a series of machine learning models to complete the new task automatically.

This chapter is based on our publication *'CD^{2}CR: Co-Reference Resolution across Documents and Domains'*. which was published in the Proceedings of the 16th Conference

of the European Chapter of the Association for Computational Linguistics, 2021 and can be accessed at https://www.virtual2021.eacl.org/paper_main.134.html Our specific contributions in this chapter are:

- A novel task setting for CDCR that is more challenging than those that already exist due to linguistic variation between different domains and document types (we call this CD^{2}CR).
- An open source English language CD^{2}CR dataset with 7602 co-reference pair annotations over 528 documents and detailed 11 page annotation guidelines (section 8.4.1).
- A novel annotation tool to support ongoing data collection and annotation for CD^{2}CR including a novel sampling mechanism for calculating inter-annotator agreement (Section 8.4.4).
- A series of experiments on our dataset using different baseline models and an in-depth capability-based evaluation of the best-performing baseline (Section 9.6)

8.2 Cross-document Cross-Domain Co-Reference Resolution

In section 2.8.2 we discuss two flavours of co-reference resolution. Intra-document co-reference resolution, which is often referred to as simply co-reference resolution, which is concerned with finding and resolving within-document references to the same entity. Conversely, Cross-document co-reference resolution (CDCR) is the task of recognising when multiple documents mention and refer to the same real-world entity or concept. In this chapter, we are primarily concerned with CDCR which allows us to identify co-referent mentions of the same concept across pairs of news articles and scientific papers.

Recent CDCR work (Dutta and Weikum, 2015; Barhom et al., 2019; Cattan et al., 2020) has primarily focused on resolution of entity mentions across news articles. Despite differences in tone and political alignment, most news articles are relatively similar in terms of grammatical and lexical structure. Work based on modern transformer-based (Vaswani et al., 2017) networks such as BERT (Devlin et al., 2019) and ElMo (Peters et al., 2018b) have been pre-trained on large news corpora and are therefore well suited to news-based CDCR (Barhom et al., 2019). However, CD^{2}CR, or CDCR across different domains such as news articles and scientific paper is hitherto unexplored.

In this work we focus on CD^{2}CR as a way to conceptually align excerpts of news articles and the scientific works that they discuss for downstream analysis of the surrounding text. For example, if a news article mentions “cancer cells” and the related scientific paper mentions “carcinoma cells” we may want to compare the surrounding context sentences to understand differences in semantics and structure. There are many other examples where

CDCR across documents from different domains is useful. Beyond comprehensive impact, conceptual alignment of news sentences and scientific papers could also facilitate sentence-wise natural language inference and enable fact checking of journalists' work (Wadden et al., 2020). For example if we know that 'cancer cells' and 'carcinoma cells' refer to the same, does "the new drug eradicated all cancer cells" mean the same as "the compound eradicated carcinoma cells in the mouse model"? Outside of these domains, a chatbot or recommender that is able to resolve references to current affairs in both formally presented news articles and informal user input could be more effective at suggesting topics that interest the user. Additionally, it may be helpful for e-commerce companies to know when product reviews gathered from third party websites refer to one of their own listings.

The objective of CD^{2}CR is to identify co-referring entities from documents belonging to different domains. In this case co-reference resolution is made more challenging by the differences in language use (lexical but also syntactic) across the different domains that we discussed in section 6.2. Success at the CD^{2}CR task in this setting is dependent on context sensitive understanding of how the accessible but imprecise writing of journalists maps on to precise terminology used in scientific writing. For example, a recent study has found that "convalescent plasma derived from donors who have recovered from COVID-19 can be used to treat patients sick with the disease" [1]. A news article[2] discussing this work says that "...blood from recovered Covid-19 patients in the hope that transfusions...[can help to treat severely ill patients]" . In this example the task is to link 'blood' to 'convalescent plasma' and 'recovered Covid-19 patients' to 'donors'. These cross-document, cross-domain co-reference chains can be used as contextual anchors for downstream analysis of the two document settings via tasks such as natural language inference, stance detection and frame analysis.

Many recent CDCR models Cattan et al. (2020); Barhom et al. (2019) and corresponding training datasets Cybulska and Vossen (2014) facilitate the detection of co-referring entity mentions (e.g. specific names of people, places, organisations) and events (e.g. "the protest", "the experiment", "the attack", "the match [between Manchester United and...]"). We focus only on entity mentions as a starting point for the CD^{2}CR corpus. However, the addition of event co-reference annotations and subsequent training of an event detection model could serve as the basis for future work.

8.3 Related Work: Entity Linking

CD^{2}CR is of particular interest in scenarios where entity linking approaches are not feasible. Entity linking has been developed as a subarea, particularly in BioNLP, to ground in-document mentions of entities to concepts defined in knowledge resources (Ji et al.,

[1] DOI: 10.1101/2020.03.16.20036145

[2] https://tinyurl.com/ycnq9xg7

2010) such as SNOMED CT[3] or DBPedia[4].

Entity Linking is challenging due to the large number of pairwise comparisons between document mentions and knowledge resource entities that may need to be carried out. Raiman and Raiman (2018) provide state of the art performance by building on Ling et al. (2015)'s work in which an entity type system is used to limit the number of required pairwise comparisons to related types. Yin et al. (2019) achieved comparable results using a graph-traversal method to similarly constrain the problem space to candidates within a similar graph neighbourhood. Entity Linking can be considered a narrow sub-task of CDCR since it cannot resolve novel and rare entities or pronouns (Shen et al., 2015). Moreover, Entity linking relies heavily on the availability of manually curated external knowledge resources which are are expensive to create and maintain, reliant on scarce subject matter expertise and usually highly specialised such that it is not possible to build a 'general science' entity linking model. Therefore a robust approach that generalises well for unknown entities and concepts is required. Furthermore, scientific journalists often write about novel scientific outputs that have been recently peer reviewed or in some cases are still in pre-print. These documents are likely to contain information not yet reflected by existing knowledge resources which are usually manually maintained and updated. Therefore, an approach robust against unseen concepts and entities is required.

8.4 Dataset creation

Our dataset is composed of pairs of news articles and scientific papers from the Linked HarriGT Corpus (Chapter 5). We also use the scraping tools described in section 5.3.5 to extend and augment the existing collection of document pairs. Our annotation process begins by obtaining summaries of the news and science document pairs (extractive news summaries and scientific abstracts, respectively) (Section 8.4.2). Candidate co-reference pairs from each summary-abstract pair are identified and scored automatically (Section 8.4.3). Candidate co-reference pairs are then presented to human annotators via a bespoke annotation interface for scoring (Section 8.4.4). Annotation quality is measured on an ongoing basis as new candidates are added to the system (Section 8.4.5).

8.4.1 Data Collection

We have developed a novel data set that allows us to train and evaluate a CD^{2}CR model. The corpus is approximately 50% the size of the ECB+ corpus (918 documents) (Cybulska and Vossen, 2014) and is split into training, development and test sets (statistics for each subset are provided in Table 8.1). Each pair of documents consists of a scientific paper and a newspaper article that discusses the scientific work. In order to detect pairs of documents,

[3] https://tinyurl.com/yy7g4ttz
[4] https://wiki.dbpedia.org/

Subset	Documents	Mentions	Clusters
Train	300	4,604	426
Dev	142	1,821	199
Test	86	1,177	101

Table 8.1: Total individual documents, mentions, co-reference clusters of each subset excluding singletons.

we follow the approach automated matching approach from Section 5.3.5, whereby we assert relationships via mentions of scientific paper DOIs in news articles or by following all hyperlinks within a news story and checking each page for relevant academic metadata.

We extend the web scraper built in Section 5.3.5 to scan for new articles from the 'Science' and 'Technology' sections of 3 well-known online news outlets (BBC[5], The Guardian[6], New York Times[7]) and press releases from Eurekalert[8], a widely popular scientific press release aggregator. Once a newspaper article and related scientific paper are detected, the full text from the news article and the scientific paper abstract and metadata are stored. Where available the full scientific paper content is also collected. We ran the scraper between April and June 2020 collecting news articles and scientific papers including preprints discussing a range of topics such as astronomy, computer science and biology (incl. coverage of COVID-19). New relevant content is downloaded and ingested into our annotation tool (see Section 8.4.4) on an ongoing basis as it becomes available.

8.4.2 Article Summarisation

Newspaper articles and scientific papers are long and often complex documents, usually spanning multiple pages, particularly the latter. Moreover, as discussed in section 6.2.2, the two document types differ significantly in length. Comparing documents of such uneven length is a difficult task for human annotators. We also assume that asking human annotators to read the documents in their entirety to identify co-references would be particularly hard with a very low chance for good inter-annotator agreement (IAA). We therefore decided to simplify the task by asking annotators to compare summaries of the newspaper article (5-10 sentences long) and the scientific paper (abstract).

For each document pair, we ask the annotators to identify co-referent mentions between the scientific paper abstract and a summary of the news article that is of similar length (e.g. 5-10 sentences). Scientific paper abstracts act as a natural summary of a scientific work and have been used as a strong baseline or even a gold-standard in scientific summarisation tasks (Liakata et al., 2013). Furthermore, abstracts are almost always available rather than behind paywalls like full text articles. For news summarisation, we used a state-of-the-art

[5] https://www.bbc.co.uk
[6] https://www.theguardian.com
[7] https://www.nytimes.com
[8] https://www.eurekalert.org/

extractive model (Grenander et al., 2019) to extract sentences forming a summary of the original text. This model provides a summary de-biasing mechanism preventing it from focusing on the repetitive parts near the start of the article (as discussed in 7.2.1), preserving the summary's informational authenticity as much as possible.

The difference in style between the two documents is preserved by both types of summary since abstracts are written in the same scientific style as full papers and the extractive summaries use verbatim excerpts of the original news articles.

8.4.3 Generation of Pairs for Annotation

Here we outline the process used to populate our annotation tool with possible examples of co-referring mentions which is illustrated in Figure 8.1.

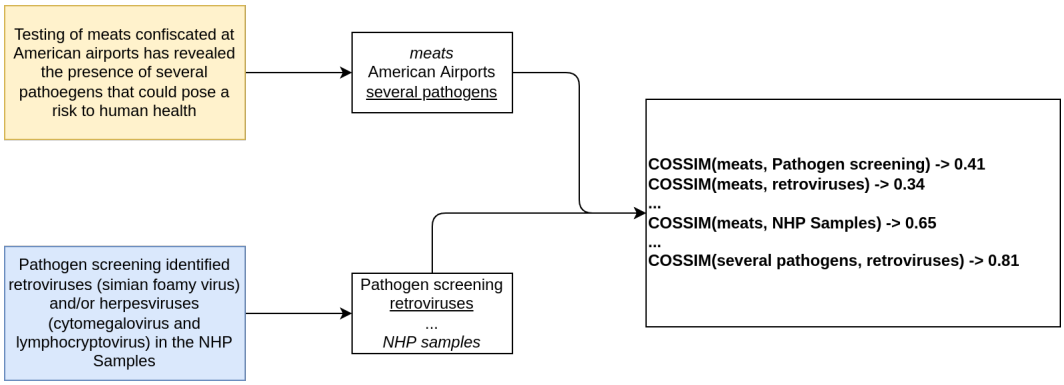


Figure 8.1: Illustration of the generation process for pairs of potentially co-referring expressions, left boxes represent related news summary (top) and abstract (bottom), co-referent entity pairs in middle boxes shown with same formatting (underline,italic).

A mention is a span of text found within the news summaries and scientific abstracts that likely corresponds to a named entity (e.g. person, location, company), noun phrase (e.g. 'a pill', 'dolphins', 'scientists') or an anaphoric reference or pronoun (e.g. 'they', 'them', 'the team', 'the dolphin'). Each mention may consist of one word or multiple words. To populate our annotation tool, we generate candidate pairs of mentions that may co-refer to be evaluated by the user. Candidate mentions are automatically identified by using spaCy (Honnibal and Montani, 2017) for the recognition of noun phrases and named entities from each input document pair (abstract & news summary). For each pair of documents, pairs of all possible mention combinations are generated and stored for annotation.

In any given pair of documents, the majority of mention pairs (M_0, M_1) generated automatically in this way will not co-refer thus resulting in a vastly imbalanced dataset and also running the risk of demotivating annotators who are much more likely to encounter non co-referent pairs. To ensure that annotators are exposed to both positive and negative examples, we use a similarity score to rank examples based on how likely they are to co-refer. The first step in generating a similarity score s is to concatenate each abstract-news-

summary pair together: “summary [SEP] abstract" into a pre-trained BERT$_{large}$ model, yielding contextualised word embeddings for all tokens in the document pair. We use these contextualised embeddings to calculate mean word vectors (see section 2.5.6.3) for each candidate mention. Then, we calculate the pairwise cosine similarity for all combination of candidate mentions within each pair of documents. Finally, for each document pair, we sort candidate pairs of mentions in descending order of cosine similarity, presenting the most similar pairs of mentions to the users of our annotation tool (see section 8.4.4 below) first since they are more likely to co-refer. We find that this BERT-based similarity score performs well in practice. We also use it in combination with a thresholding policy as one of our baseline models in Section 8.5.

8.4.4 Annotation Tool & Interface

We developed an open source annotation tool[9] that allows humans to identify cross document co-reference between each pair of related documents. Whilst designing this tool, we made a number of decisions to simplify the task and provide clear instructions for the human annotators in order to encourage consistent annotation behaviour.

To maximise the quality and consistency of annotations in our corpus, we simplified the task as much as possible for the end user. Annotation tasks were framed as a single yes or no question: “Are **x** and **y** mentions of the same entity?". Mentions in context were shown in bold font whereas mentions already flagged as co-referent were shown in green. This enabled annotators to understand the implications for existing co-reference chains before responding (see Figure 8.2). For example, if a pair of articles discusses two microbes, knowing about the existence of a link between the first microbe and an endophoric mention (‘the germ’) should prevent the user from linking the same instance of ‘the germ’ to the second instance, thus, signalling that the two separate microbes refer to the same entity. Questions were generated and ranked via our task generation pipeline (see Section 8.4.3 above).

We added two additional features to our annotation interface to improve annotators’ experience and to speed up the annotation process. Firstly, if the candidate pair is marked as co-referent, the user is allowed to add further mentions to the coreference cluster at once. Secondly, inspired by Li et al. (2020), if the automatically shown mention pair is not co-referent, the user can select a different mention that is co-referent.

The upstream automated mention detection mechanism can sometimes introduce incomplete or erroneous mentions, leading to comparisons that don’t make sense or that are particularly difficult. Therefore, annotators can also move or resize the mention spans they are annotating.

We use string offsets of mention span pairs to tokens to check that they do not overlap with each other in order to prevent the creation of duplicates.

[9] `https://github.com/ravenscroftj/cdcrtool`

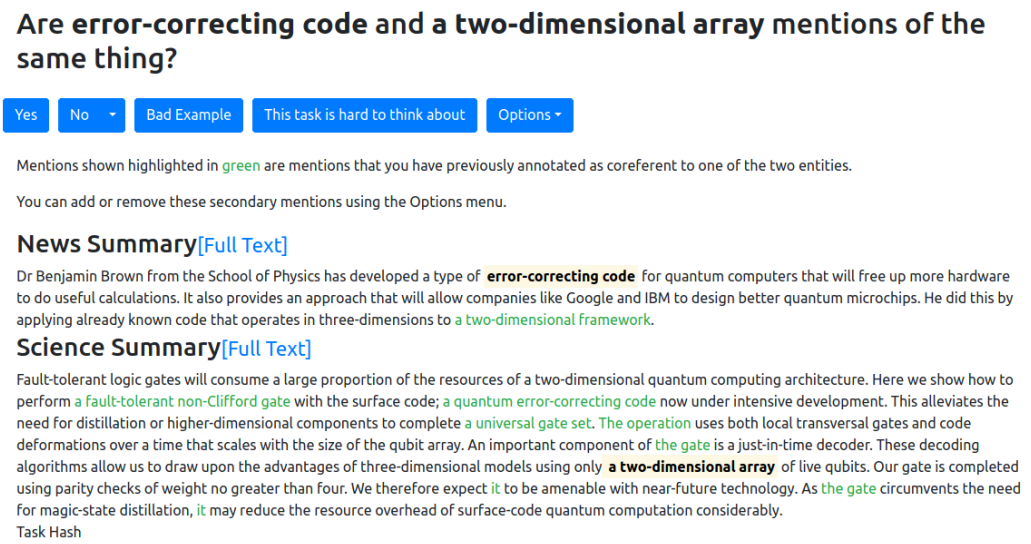


Figure 8.2: An example of a cross-document co-reference task presented within our annotation tool.

8.4.5 Annotation Protocol

We recruited three university-educated human annotators and provided them with detailed annotation guidelines for the resolution of yes/no questions on potentially co-referring entities in pairs from the ordered queue described above. By default each entity pair resolution is carried out once, allowing us to quickly expand our data set. However, we pseudo-randomly sample 5% of mention pairs in order to calculate inter-annotator-agreement (IAA) and make sure that data collected from the tool is consistent and suitable for modelling. New entity pairs for IAA are continually sampled as new document pairs and mention tuples are added to the corpus by the web scraper (Section 8.4.1). The annotation system puts mention pairs flagged for IAA first in the annotation queue. Thus, all annotators are required to complete IAA comparisons before moving on to novel mention pairs. This allows us to ensure that all annotators are well represented in the IAA exercise. To avoid annotators being faced with a huge backlog of IAA comparisons before being able to proceed with novel annotations, we also limited the number of comparisons for IAA required by each user to a maximum of 150 per week.

8.4.6 Task Difficulty and Annotator Agreement

We anticipated that annotation of the CD^{2}CR corpus would be difficult in nature due to its dependencies on context and lexical style. We invited users to provide feedback regularly to help us refine and clarify our guidelines and annotation tool in an iterative fashion. Users could alert us to examples they found challenging by flagging them as difficult in the tool. Qualitative analysis of the subset of 'difficult' cases showed that the resolution of mention

	# Annotations	A1	A2	A3
A1	10,685	-	0.492	0.600
A2	3,051	0.492	-	0.500
A3	9,847	0.600	0.500	-

Table 8.2: Number of Annotations and Pairwise Cohen's Kappa scores κ_{cohen} demonstrating 'moderate agreement' between annotators.

pairs is often perceived by annotators as difficult when:

- Deep subject-matter-expertise is required to understand the mentions, e.g. is "jasmonic acid" the same as "regulator cis -(+)-12-oxophytodienoic acid".
- Mentions involve non-commutable set membership ambiguity e.g. "Diplodocidae" and "the dinosaurs"
- Mentions are context dependent e.g. "the struggling insect" and "the monarch butterfly".

This feedback prompted the introduction of highlighting for existing co-reference chains in the user interface (as described in section 8.4.4 above) to make it easier to tell when non-commutable set membership would likely introduce inconsistencies into the dataset. For mention pairs requiring subject-matter-expertise, annotators were encouraged to research the terms online. For context sensitive mention pairs, annotators were encouraged to read the full news article and full scientific paper in order to make a decision.

In our 11 page annotation guidelines document (Appendix A) we describe the use of our annotation tool and illustrate some challenging CD^{2}CR tasks and resolution strategies. For example precise entities mentioned in the scientific document may be referenced using ambiguous exophoric mentions in the news article (e.g. 'a mountain breed of sheep' vs 'eight ovis aries'). Our guidelines require resolving these cases based on the journalist's intent (e.g. 'a mountain breed' refers to the 'ovis aries' sheep involved in the experiment).

We evaluated the final pairwise agreement between annotators using Cohen's Kappa (Cohen, 1960) (κ_{cohen}) and an aggregate 'n-way' agreement score using Fleiss' Kappa (Fleiss, 1971) (κ_{fleiss}). Pairwise κ_{cohen} is shown in Table 8.2 along with the total number of tasks each annotator completed. Annotator 3 (A3) shows the most consistent agreement with the other two annotators. Our Fleiss' Kappa analysis of tasks common across the three annotators gave $\kappa_{\text{fleiss}} = 0.554$. We note that Fleiss' Kappa is a relatively harsh metric and values, like ours, between 0.41 and 0.60 are considered to demonstrate 'moderate agreement' (Landis and Koch, 1977). We also carried out Fleiss' Kappa analysis on the subset of mention pairs that were completed by all annotators and were also marked as difficult by at least one user (180 mention pairs in total). We found that for this subset of pairs, $\kappa_{\text{fleiss}} = 0.399$ which is considered to be fair agreement (Landis and Koch, 1977).

8.5 Model

Next, we describe several baseline models, including state of the art CDCR models, that we used to evaluate how well current approaches can be used in our CD^{2}CR task setting.

8.5.1 BERT Cosine Similarity (BCOS) Baseline

In this model we calculate the cosine-similarity between embeddings of two mentions in context (M_0, M_1) encoded using a pre-trained BERT model as discussed above in section 8.4.3. We define a thresholding function f to decide if M_0 and M_1 are co-referent ($f(x) = 1$) or not ($f(x) = 0$):

$$f(x) = \begin{cases} 1, & \text{if COSSIM}(M_0, M_1) \geq t \\ 0, & \text{otherwise} \end{cases}$$

During inference, we apply this function to all pairs M_0, M_1 and infer missing transitive results such that if $f(A, B) = 1$ and $f(B, C) = 1$ then $f(A, C) = 1$.

Based on Figure 8.3, we test values for threshold cut off t in increments of 0.01 between 0.3 and 0.8 inclusive. We evaluated the baseline by measuring its accuracy at predicting co-reference in each mention pair in the CD^2CR development set. The best performance was attained when $t = 0.65$. A visualisation of the BERT Cosine Similarity (BCOS) distributions of co-referent and non co-referent annotated mention pairs can be seen in Figure 8.3.

Co-referent mention pairs tend to have a slightly higher BERT cosine similarity than non co-referent mention pairs but there is significant overlap of the two distributions suggesting that in many cases BERT similarity is too simplistic a measure.

8.5.2 Entities Only Baseline (CA)

We use a state-of-the-art model (Cattan et al., 2020) (CA) for cross-document co-reference resolution. In this model, each document is separately encoded using a RoBERTa encoder (without fine-tuning) to get contextualized representations for each token. Then, similarly to the within-document co-reference model by Lee et al. (2017), the mention spans are represented by the concatenation of four vectors: the vectors of the first and last token in the span, an attention-weighted sum of the span token vectors, and a feature vector to encode the span width. Two mention representations are then concatenated and fed to a feed-forward network to learn a likelihood score for whether two mentions co-refer. At inference time, agglomerative clustering is used on the pairwise scores to form coreference clusters.

The CA model is trained to perform both event and entity recognition on the ECB+ corpus (Cybulska and Vossen, 2014) In our setting there is no event detection subtask so, for

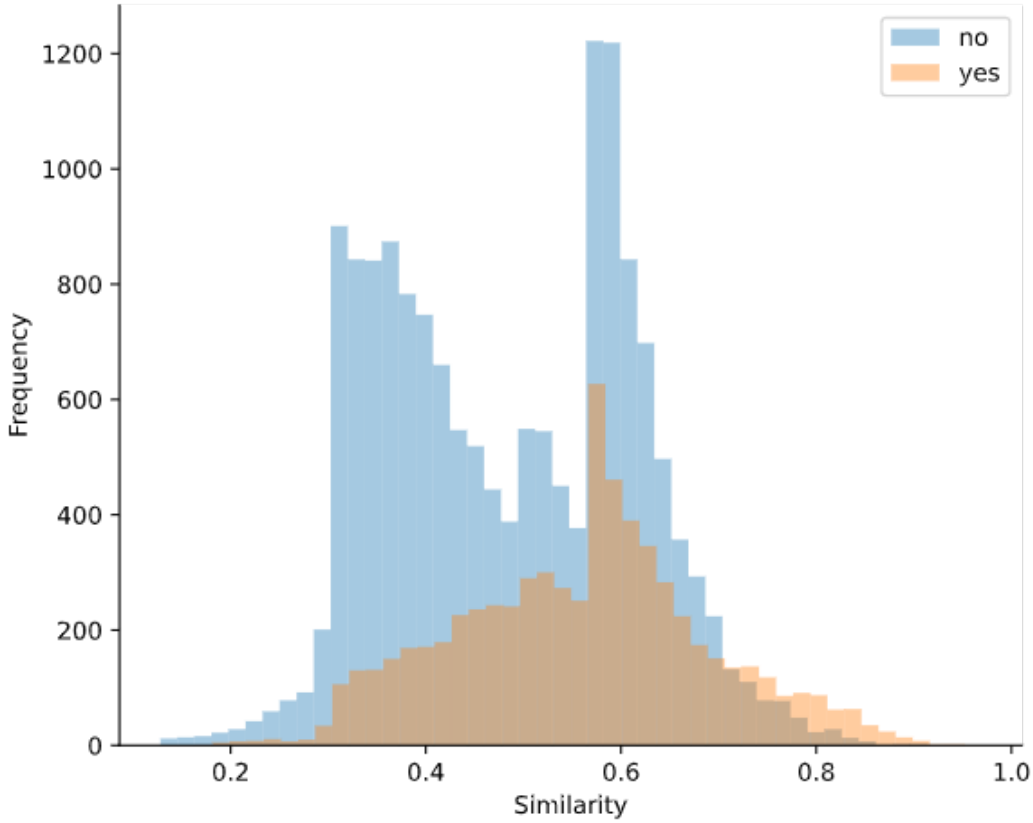


Figure 8.3: BERT Cosine Similarity frequency distribution for co-referent (Yes) and non-co-referent (No) mention pairs in the CD^{2}CR corpus. Significant overlaps in the distributions suggests that BERT similarity is too simplistic a measure.

fair comparison, we pre-train the CA model on ECB+ entity annotations only and evaluate it on our new CD^{2}CR task to see how well it generalises to our task setting.

8.5.3 CA + Fine-Tuned (CA-FT) Baseline

Here we aim to evaluate whether fine tuning the CA model from section 8.5.2 using the CD^{2}CR corpus can improve its performance in the new task setting. The CA model is first trained on the ECB+ corpus in the manner described above. We then further fine-tune the feed-forward model (without affecting the RoBERTa encoder) on the CD^{2}CR corpus for 10 epochs with early stopping. Pseudo-random sub-sampling is carried out on the training set to ensure a balance of co-referent and non-co-referent mention pairs.

8.5.4 CA - Vanilla (CA-V) Baseline

Here we aim to evaluate whether training the CA model on the CD^{2}CR dataset from the RoBERTa baseline without first training on the ECB+ corpus allows it to fit well to the new task setting. We re-initialise the CA encoder (Section 8.5.2) using weights from RoBERTa

(Liu et al., 2019b) and randomly initialise the remaining model parameters. We then train the model on the CD^{2}CR corpus for up to 20 epochs with early stopping with pseudo-random sub-sampling as above.

8.5.5 CA - SciBERT (CA-S) Baseline

This model is the same as CA-V but we replace the RoBERTa encoder with SciBERT (Beltagy et al., 2019), a version of BERT pre-trained on scientific literature in order to test whether the scientific terms and context captured by SciBERT improve performance at the CD^{2}CR task compared to RoBERTa. Similarly to CA-V in section 8.5.4, we initialise the BERT model with weights from SciBERT$_{\text{scivocab-uncased}}$ (Beltagy et al., 2019) and randomly initialise the remaining model parameters, training on the CD^{2}CR corpus for up to 20 epochs with early stopping.

8.6 Results and Discussion

We evaluate each of the model baselines described in section 8.5 above on the test subset of our CD^{2}CR corpus. Results are shown in Table 8.3.

For the purposes of evaluation, we use named entity spans from the manually annotated CD^{2}CR as the “gold standard" in all experiments rather than using the end-to-end Named Entity Recognition capabilities provided by some of the models. We evaluate the models using the metrics described by Vilain et al. (1995) (henceforth MUC) and Bagga and Baldwin (1998) (henceforth B^3). MUC F1, precision and recall are defined in terms of pairwise co-reference relationships between each mention. B^3 F1, precision and recall are defined in terms of presence or absence of specific entities in the cluster. When measuring B^3, we remove entities with no co-references (singletons) from the evaluation to avoid inflation of results (Cattan et al., 2020).

The threshold baseline (BCOS) gives the highest MUC recall but also poor MUC precision and poorest B^3 precision. The B^3 metric is highly specific with respect to false-positive entity mentions and strongly penalises BCOS for linking all non-coreferent pairs

Model	**MUC**			**B^3**		
	P	**R**	**F1**	**P**	**R**	**F1**
BCOS	0.42	**0.94**	0.58	0.01	0.45	0.00
CA	0.41	0.51	0.46	**0.39**	0.33	0.35
CA-V	0.50	0.69	**0.58**	0.35	0.57	**0.44**
CA-FT	0.47	0.71	0.52	0.30	**0.62**	0.41
CA-S	**0.58**	0.46	0.51	0.32	0.53	0.39

Table 8.3: MUC and B^3 results from running baseline models on CD^{2}CR test subset, BCOS threshold=0.65

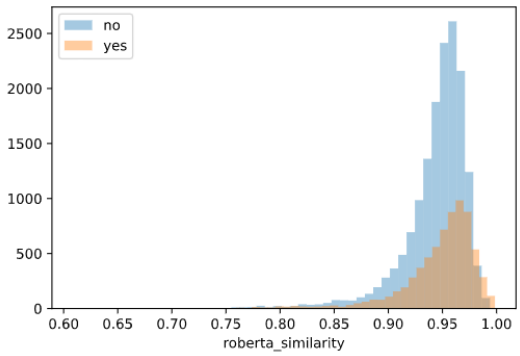


Figure 8.4: RoBERTa Cosine Similarity frequency distributions for co-referent (Yes) and non-co-referent (No) mention pairs in the CD^{2}CR corpus. Distribution is compressed between 0.8 and 1.0.

with $COSSIM(M_0, M_1) \geq 0.65$. Furthermore, Fig. 8.3 shows that a thresholding strategy is clearly sub-optimal given that there is a significant overlap of co-referent and non-co-referent pairs with only a small minority of pairs at the top and bottom of the distribution that do not overlap. Therefore, despite its promising MUC F1 score, it is clear that BCOS is not useful in practical terms.

Whilst our thresholding baseline above uses BERT, RoBERTa is used by Cattan et al. (2020) as the basis for their state-of-the-art model and thus for our models based on their work. Although the two models have the same architecture, RoBERTa has been shown to outperform BERT at a range of tasks (Liu et al., 2019b). However, as shown in Figure 8.4, the cosine similarity distribution of mention pair embeddings produced by RoBERTa is compressed to use a smaller area of the potential distribution space compared to that of BERT (Figure 8.3). This compression of similarities may imply a reduction in RoBERTa's ability to discriminate in our task setting. Liu et al. (2019b) explain that their byte-pair-encoding (BPE) mechanism, which expands RoBERTa's sub-word vocabulary and simplifies pre-processing, can reduce model performance for some tasks, although this is not further explored in their work. We leave further exploration of RoBERTa's BPE scheme and its effects on the CD^{2}CR task setting to future work.

All of the models specifically trained on the CD^{2}CR corpus (CA-V, CA-FT, CA-S) outperform the CA model by a large margin. Furthermore, the CA-V model (without pre-training on ECB+ corpus) outperforms the CA-FT model (with ECB+ pre-training) by 6% MUC and 3% B^3. These results suggest that the CD^{2}CR task setting is distinct from the CDCR and ECB+ task setting and that this distinction is not solvable with fine-tuning.

In terms of both MUC and B^3, CA-S performs much worse than CA-V suggesting that SciBERT embeddings are less effective than RoBERTa embeddings in this task setting. We hypothesise that SciBERT's specialisation towards scientific embeddings may come at the cost of significantly worse news summary embeddings when compared to those produced by RoBERTa.

We next evaluate our best performing CD^{2}CR baseline model (CA-V) at the entity resolution CDCR task using the ECB+ test corpus, to see how well it generalises to the original CDCR task. Results are presented in Table 8.4 along-side Cattan et al's original model results (CA). The CA-V model still shows good performance, despite a small drop, when compared to the original CA model. The drop in B^3 F1 is more pronounced than MUC but is still broadly in line with other contemporary CDCR systems (Cattan et al., 2020). The CA-V model demonstrates a promising ability to generalise beyond our corpus to other tasks and reveals an interesting correspondence between CDCR and CD^{2}CR settings.

Model	**MUC**			B^3		
	P	**R**	**F1**	**P**	**R**	**F1**
CA	0.86	0.82	0.84	0.63	0.68	0.65
CA-V	0.82	0.81	0.81	0.56	0.53	0.55

Table 8.4: MUC and B^3 results from running the CD^{2}CR baseline model (CA-V) on ECB+ dataset compared with original (CA).

Finally, the best model (CA-V) is analysed using a series of challenging test cases inspired by Ribeiro et al. (2020). These test cases were created using 210 manually annotated mention-pairs found in the test subset of the CD^2CR corpus according to the type of relationship illustrated (Anaphora & Exophora, Subset relationships, paraphrases). We collected a balanced set of 30-40 examples of both co-referent and non-coreferent-but-challenging pairs for each type of relationship (exact numbers in Table 8.5). We then recorded whether the model correctly predicted co-reference for these pairs. The results along with illustrative examples of each relationship type are shown in Table 8.5. The results suggest that the model is better at identifying non-co-referent pairs than co-referent pairs and that it struggles with positive co-referent mentions for all three types of relationship. The model struggles to relate general reader-friendly descriptions of entities from news articles to precise and clinical descriptions found in scientific papers. The model often successfully identifies related concepts such as 'the carnivore's skull' and 'Agriotherium africanum'. However it is unable to deal with the complexity of these relationships and appears to conflate 'related' with 'co-referent', which is likely due to lack of lexical knowledge, which we discuss further in the conclusion of this chapter. Figure 8.5 shows significant overlap between co-referent and non-co-referent RoBERTa-based cosine similarities, which can also be observed for the wider corpus in Figure 8.4, but is especially bad for these test examples. This overlap suggests that disentangling these pairs is likely to be a challenging task for the downstream classification layer in the CA-V model. These challenges are less likely to occur in homogeneous corpora like ECB+ where descriptions and relationships remain consistent in detail and complexity.

We examine the pairwise similarities produced by RoBERTa (Liu et al., 2019b), SciBERT (Beltagy et al., 2019) and BERT (Devlin et al., 2019) for mentions from our

Test Type	Co-referent?	Pass Rate & Total Tests	Example test case and outcome for test case
Anaphora and Exophora resolution	Yes	47.1% (16/34)	M1: ...to boost **the struggling insect**'s numbers... *[PASS]* M2: the annual migration of the **monarch butterfly**...
	No	76.5% (26/34)	M1: ...**monarchs** raised in captivity... *[FAIL]* M2: ... wild-caught **monarchs** in an indoor environment...
Subset relationship resolution	Yes	24.3% (9/37)	M1: ...it was in fact a hive of **human activity**... *[FAIL]* M2: ...**Pre-Columbian cultural developments**...
	No	60.0% (18/30)	M1: ... **the carnivore's skull**... *[FAIL]* M2: ... **the gigantic extinct Agriotherium africanum**
Paraphrase resolution	Yes	33.3% (13/39)	M1: ...**a giant short-faced bear**... *[PASS]* M2: ...**the gigantic extinct Agriotherium africanum**...
	No	80.5% (29/36)	M1: ...the energy that **existing techniques** require *[FAIL]* M2: ...the lack of **efficient catalysts** for ammonia synthesis

Table 8.5: A breakdown of specific tests carried out on CA-V model against three challenging types of relationships found in the CD^2CR corpus. [PASS] or [FAIL] indicates CA-V model correctness. Pass Rate is mathematically equivalent to Recall for test sets.

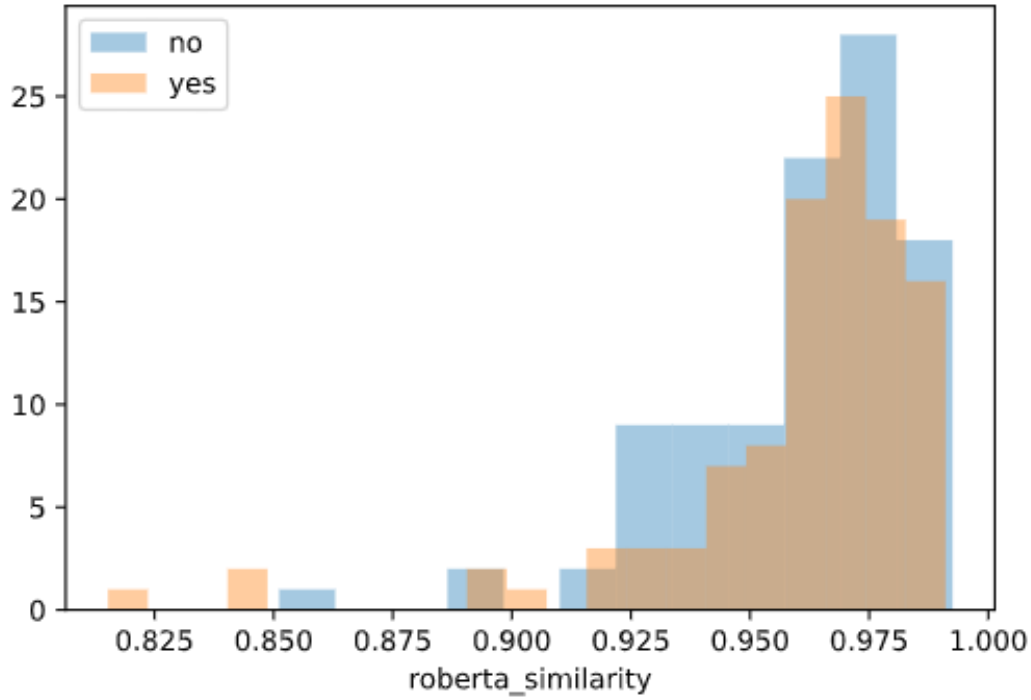


Figure 8.5: RoBERTa-based mention pair similarity frequency distributions for co-referent ('yes') and not co-referent ('no') mention pairs for test examples from the test types shown in Table 8.5.

Mention 1	Mention 2	Co-referent	SciBERT	BERT	RoBERTa
a home refrigerator	Vapour compression refrigerator system	Y	0.76	0.56	0.97
the findings	molecular analyses	Y	0.59	0.61	0.94
fatal brain injuries	abusive head trauma	N	0.63	0.67	0.97
dinosaur	Massospondylus carinatus embryos	Y	0.49	0.46	0.95
the immune system	macrophages	N	0.35	0.52	0.93

Table 8.6: The transformer cosine similarities for some example cases demonstrating where SciBERT's domain specialisation helps and hinders the model from identifying relationships correctly.

dataset to further explore the characteristics of each model. Table 8.6 shows some cases where SciBERT's domain specialisation helps and hinders the model from identifying relationships correctly. It also shows the complexity of some entity relationships (e.g. abusive head trauma can sometimes be fatal, macrophases are part of the immune system, Massospondylus carinatus is a member of the dinosaur family). In most cases, all three language models are able to identify some degree of semantic similarity between concepts. However, they all struggle with the specific nature of the semantic relationships between entities. Approaches for better representing relationships between entities to improve model performance that could be explored as part of future work are discussed below in section 8.7.

8.7 Conclusion

In this chapter we have defined cross-document, cross-domain co-reference resolution (CD^{2}CR), a special and challenging case of cross-document co-reference resolution for comparing mentions across documents of different types and/or themes. We have constructed a specialised CD^{2}CR annotated dataset, available, along with our annotation guidelines and tool, as a free and open resource for future research and which have already been used by others to the evaluate new CDCR and CD^{2}CR approaches (Held et al., 2021) and inspired new CDCR data annotation approaches (Cattan and Johnson, 2021).

We have shown that state-of-the-art CDCR models do not perform well on the CD^{2}CR dataset without specific training. Furthermore, even with task-specific training, models perform modestly and leave room for further research and improvement. Finally, we show that the understanding of semantic relatedness offered by current generation transformer-based language models may not be precise enough to reliably resolve complex linguistic relationships such as those found in CD^{2}CR as well as other types of co-reference resolution

and relationship extraction tasks.

As discussed in section 2.5, static neural vector language models like word2vec (Mikolov et al., 2013a) and GloVe Pennington et al. (2014b) and contextual language models such as BERT (Devlin et al., 2019) and ElMo (Peters et al., 2018b) all use distributional knowledge (Harris, 1954) inherent in large text corpora to learn word embeddings that can be used for downstream NLP tasks. However, these models do not learn about formal lexical constraints, often conflating different types of semantic relatedness (Ponti et al., 2018; Lauscher et al., 2020). This is a weakness of all distributional language models that is particularly problematic in the context of CD^{2}CR for entity mentions that are related but not co-referent (e.g. “Mars" and “Jupiter") as shown in section 9.6. A number of solutions have been proposed for adding lexical knowledge to static word embeddings (Yu and Dredze, 2014; Wieting et al., 2015; Ponti et al., 2018) and more recently to contextual language models (Lauscher et al., 2020; Majewska et al., 2021) as well as novel architectures like ERNIE (Zhang et al., 2019) which fuses both text and lexical information inputs in order to carry out downstream tasks. Whilst we do not experiment with semantic specialization in this work, it does provide an interesting direction for future work on CD^{2}CR. Incorporation of relational similarity prediction (Jurgens et al., 2012) in which the model must predict the nature of a relationship between two mentions (e.g. equivalent, subset, superset) in order to improve CD^{2}CR performance may also be a promising direction for future work.

In the broader context of this thesis, CD^{2}CR provides us with a mechanism for aligning sentences that discuss the same concepts and entities without relying heavily on semantic similarity which, as discussed above, may be problematic due to differences in style between news articles and scientific papers. Although our CD^{2}CR models yield modest performance, our CD^{2}CR corpus provides us with large a number of aligned sentence pairs which can be used to further experiment with differences in representation between the two document types. In the next chapter we focus on how textual representations of comprehensive impact may be detected and quantified using the CD^{2}CR corpus as our starting point.

9

Detecting and Measuring Scientific Impact with Multi-Task Few-Shot Learning

"Half of wisdom is learning what to unlearn"

Larry Niven

9.1 Introduction

Thus far in this thesis we have sought, through the application of NLP techniques, to understand the relationships between scientific works, news articles and REF Impact score in order to identify new, scalable ways to detect and quantify the comprehensive impact of scientific work and that help scientists, funding bodies and the general public to understand how scientific works impact society and the environment.

In the preceding chapters, we:

1. identified news as a ubiquitous and rich source of information evidencing comprehensive impact and constructed a dataset of news articles linked to scientific papers (Chapter 5).
2. explored the similarities and differences between different text representations of news articles and the scientific papers that they discuss (Chapter 6).
3. investigated which information journalists are most likely to display prominently in their articles (Chapter 7).
4. explored how CD^2CR can be used to align grammatically distinct discussions of co-referent concepts across the two document types (Chapter 8).

Our work in this chapter is motivated by **RQ5: How can we detect and quantify comprehensive impact implied by language use in scientific news articles and the academic works that they discuss?** (section 1.2) and within we tackle challenges **CH1** (section 1.3.1) and **CH3** (section 1.3.3). We leverage the aforementioned techniques and datasets to construct a new dataset and set of models for quantifying comprehensive impact implied by language used in news articles and scientific papers.

This chapter is based on our paper *'Detecting and Measuring Scientific Impact with Multi-Task Few-Shot Learning'.* which is currently pending review for publication in Transactions of the Association for Computational Linguistics at time of writing.

As discussed in Chapters 5 and 6, the relationship between scientific works and news articles that discuss them is complex and multifaceted; journalists must help readers

to confront a number of challenges in understanding and contextualising the comprehensive impact described within these distinct types of document:

- They may need specific background knowledge to understand the importance of certain outcomes. For example, knowing that *P. falciparum* is a vector for malaria furnishes the reader with an appreciation of the impact of discovery of markers for detecting it.
- They must be able to contextualise the scope of outcomes and their relative effect. E.g. the discovery of a new species of dinosaur which likely yields academic impact but limited societal impact (perhaps it inspires the next *'villain'* in the next Jurassic Park movie) vs development of new vaccines for virulent fatal illnesses which may yield significant societal impact and save thousands or millions of lives.
- They must be able to appropriately interpret ambiguity and uncertainty in scientific works e.g. this work has the *potential* for new drugs.

When communicating scientific outcomes, scientists and journalists engage in a cooperative constraint optimisation process. Scientists are constrained by the need to write precisely and objectively such that their work is correctly described and reproducible; clear communication is rewarded with the potential for further funding, career opportunities and new collaborations with other academics. Journalists are rewarded for communicating complex works in an accessible, engaging way to as large an audience as possible (Hodgetts et al., 2008) whilst constrained by their ability to simplify complex work without misleading readers (Louis and Nenkova, 2013b; Molek-Kozakowska, 2017). In this process, a precise scientific article with clear, captivating outcomes paired with an accurate, engaging news article will bring success to both parties. Failure of either party, through exaggeration, misrepresentation or poor communication, makes the other party's role harder or at worst brings disrepute to both parties.

We define the Scientific Impact Scoring (SIS) *regression task* (see Section 2.1.1) which takes sentences from scientific abstracts and news articles and provides a real-valued score in terms of the comprehensive impact they convey. This is a challenging task for humans, where recent developments in natural language processing can help improve outcomes.

Our specific contributions are:

- A pilot investigation examining the relationship between comprehensive impact detection and linguistic connotation (Section 9.3).
- A new task, Scientific Impact Scoring (SIS), that re-frames detection and quantification of statements about comprehensive impact in text as a regression problem (Section 9.3).

- A new few-shot learning corpus of 430 sentences for our SIS task as well as detailed annotation guidelines and an open-source annotation tool (Section 9.4).
- A high performing regression model which outperforms the SOTA impact classification model by 18% (Section 9.6).
- A discussion and comparison of few-shot and multi-task learning approaches for our task and the benefits for such approaches in low resourced tasks (Section 9.6).

9.2 Related Work

Here we explore works relating to scientific discourse and communication of scientific impact as well as learning in low resource settings.

9.2.1 Communication of comprehensive impact

Communication of comprehensive impact is a non-trivial activity. In order to accurately detect and measure such communication, a number of linguistic mechanisms and phenomena must be taken into account. August et al. (2020) assembled a corpus of scientific news articles and press releases annotated with communication goals or 'writing strategies' at sentence-level including communication of scientific impact. They train RoBERTa-based (Liu et al., 2019a) classifiers to detect usage of each strategy respectively. Their scientific impact model, performs well on news articles and press releases but has not been tested on the more formal language of scientific papers. Furthermore, the binary framing of their task makes the quantitative comparison of impact across documents impossible.

Molek-Kozakowska (2017) carried out a linguistic analysis of science news headlines in order to identify how they grab readers' attention. They focus on 'newsworthiness', a qualitative product of different aspects of the framing used to describe scientific work including the work's scope, scale and novelty, timeliness, cultural relevance and how relatable it is. Likewise, Louis and Nenkova (2013b) built a corpus of scientific news articles of varying quality which they analyse, finding that good scientific journalists often use creative language and metaphor to improve engagement and more clearly communicate complex scientific concepts to their readers. These linguistic mechanisms could be used by journalists to engineer sensationalism or, as noted by Fuoco (2021), to improve communication of findings via the cooperative process discussed in Section 9.1.

The strength and exaggeration of claims are also key aspects of the impact within scientific journalism. Pei and Jurgens (2021) trained a regression model to assess claim strength using a corpus of pairs of paraphrased claims from news articles and scientific abstracts that they assembled. Their findings suggest that journalists may be less prone to exaggeration than first suspected, although the simple paraphrase detection approach used in their work relies on token overlap and may under-represent more complex instances of

paraphrasing. Similarly, Wright and Augenstein (2021) modelled exaggeration of scientific claims in pairs of paraphrasing sentences identified in abstracts and corresponding press releases. They used a multi-task formulation of Pattern-Exploitative Training (PET) (Schick and Schütze, 2021) to jointly detect claim strength and exaggeration on pairs of paraphrased claims but did not draw conclusions about the extent to which claims are exaggerated in press releases.

9.2.2 Scientific Discourse

Context is a key determinant of whether a statement in a scientific work may be considered impactful. For example, background statements such as "malaria kills millions of people every year" provide a sense of cultural relevance and grab readers' attention but do not confer the impact of the work itself. As discussed in section 7.2.2, schemes like CoreSC (Liakata, 2010) can provide useful context for downstream impact detection models by providing a way for downstream models to discern contextual relevance.

Recent work by Brack et al. (2021) combined the original modelling task described by Liakata et al. (2012a) with more recent deep learning models, carrying out multi-task learning on related scientific discourse classification datasets in order to achieve SOTA CoreSC classification performance.

9.2.3 Learning in Low Resource Settings

In this chapter we work with a small labelled dataset (see Section 9.4 below) and thus we explore techniques for training NLP models in a low resource setting. In Section 2.7.1 we outline how modern neural PLMs have facilitated performant few-shot training schemes. Here we describe two specific techniques for carrying out few-shot learning on PLMs.

9.2.3.1 Prompt-based Learning

Prompt-based Learning is a new fine-tuning paradigm for Pre-trained Language Models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019a) (section 2.5.7) that provides SOTA performance in few-shot task settings (Gao et al., 2021; Schick and Schütze, 2021). Prompt-based learning was introduced by (Radford et al., 2019) who showed that large PLMs with decoder-like architectures (see sections 2.6.7 & 2.6.8.2) can be conditioned to carry out a number of NLP tasks by framing them as language generation tasks and exploiting the next-word-prediction task that is often used as part of PLM pre-training. In this paradigm the model is fed an input sequence x, the prompt, and must generate a corresponding output. For example, the prompt provided to the model may be "the cat sat on the mat translated to french is ..." and the model completes the sequence "le chat s'est assis sur le tapis". Prompt-based learning has been recently popularised by

GPT-3 (Brown et al., 2020), a very large pre-trained language model which is capable of reliably generating a large and diverse set of sequences dependent on an input prompt.

In Radford et al. (2019) and Brown et al. (2020)'s work, no traditional supervised learning (i.e. back propagation or model weight adjustment) takes place. Instead, the models are conditioned to produce a helpful answer by being fed a prompt or in the latter case, a sequence of prompts and corresponding answers, increasing the probability that the autoregressive decoder component generates a useful response. Conversely, Schick and Schütze (2021) and Gao et al. (2021) do train the models used in their experiments using back-propagation. We describe this process in more detail in section 9.5.1 and use it to train some of our models.

9.2.3.2 Multi-Task Learning

Multi-Task Learning (MTL) is a machine learning paradigm in which multiple tasks are learned by the same model at the same time in order to exploit commonalities between them. MTL is similar but tangential to transfer learning (see section 2.7) in the sense that, in transfer learning the goal is to improve performance at a new task based on learnings from a previous task which we no longer evaluate whereas in MTL the goal is to use common learnings to improve performance at all tasks (Zhang and Yang, 2021). MTL for NLP tasks is often achieved in a deep learning setting by constructing a neural network consisting of shared representation layers via transformers or recurrent layers and then attaching task-specific transformation layers and output heads for each of the tasks to be learned by the model (Crawshaw, 2020). These models are typically trained by iteratively alternating through each task, performing one back-propagation step for each iteration, although there are also a number of more complex approaches to task scheduling (*ibid*).

Radford et al. (2019) assert that large transformer-decoder language models that they train are unsupervised multitask learners and demonstrate reasonable performance at a number of NLP tasks including machine translation, summarisation, and reading comprehension simply by formulating the task as a prompt. More recent work in prompt-based MTL has focused on improving few or zero-shot performance at a primary task by having the model learn related *auxiliary* tasks (Sanh et al., 2021; Mishra et al., 2021).

These works use bidirectional encoder-decoder language models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) which can output long-form responses to input prompts and are thus more flexible than masked PLMs like BERT which provide encoder functionality only and are trained to predict one masked word at a time. This makes the encoder-decoder models easier to use for MTL with a wide range of language tasks. However, the increased flexibility of these comes at the cost of increased complexity and expense. On the other hand, Wright and Augenstein (2021) show that MTL also works well for prompt-based learning in smaller PLMs by modifying Schick and Schütze (2021)'s PET to support MTL.

In this work we benchmark prompt-based learning and shared representation transformer MTL approaches against each other in order to develop performant scoring models for scientific impact.

9.3 Defining Scientific Impact Scoring (SIS)

Our experiment focused on a 430 sentence subset of the CD^2CR corpus developed in Chapter 8. We call this subset the Few-shot Science Impact Scoring (`FSIS`) corpus.

We used the strong link between the linguistic connotations of scientific writing to its perceived impact (Molek-Kozakowska, 2017) and Webson et al. (2020)'s disentangled representation model to separate and contextualise predictions from August et al. (2020)'s impact model.

9.3.1 Connotation & Denotation

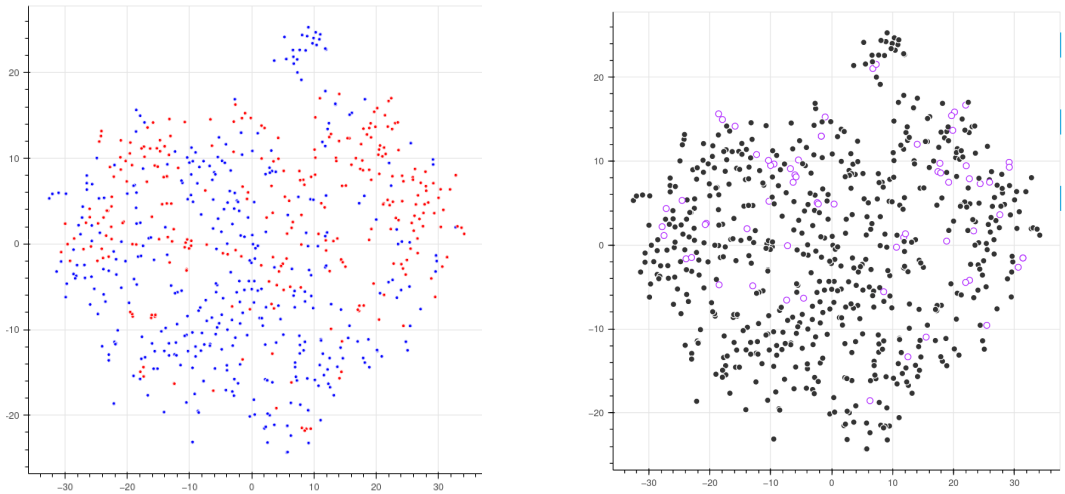
Words in natural language carry both a literal, specific meaning, known as word's denotation, and cultural or emotional associations, known as connotations. Journalists can drastically alter our perception of news articles by picking words with similar denotations that have very different connotations (Hamborg et al., 2019) allowing them to increase the newsworthiness of the work that they are describing (Molek-Kozakowska, 2017). Webson et al. (2020) use their model to disentangle connotation and denotation of political terms e.g. "illegal alien" and "undocumented worker" have similar denotation but different connotations. Their model learns to separate a pre-trained dense word embedding space $V_{pretrained}$ into distinct embedding spaces V_{conno} for connotations and V_{deno} for denotations which can both be used independently for downstream tasks.

We trained Webson et al. (2020)'s model using pairs of sentences from the full CD^2CR corpus, each sentence receiving a $c \in \{science, news\}$ label, approximating its connotation, depending on whether it came from a science or news article, and a d label, approximating denotation, corresponding to which of the 264 topically aligned[1] document pairs it came from. As in the original work, we initialise the $V_{pretrained}$ space with GloVE embeddings (Pennington et al., 2014a). We then generated mean vector embeddings (Arora et al., 2017) from V_{conno} for each of the FSIS sentences and visualised via a 2-dimensional t-SNE (Maaten and Hinton, 2008) projection as shown in Figure 9.1a. We used the plot to facilitate qualitative inspection and exploration of the corpus.

The plot shows distinct grouping of news (red) and science (blue) sentences suggesting that differences in connotation and language use between the two types of documents are discernible by Webson et al. (2020)'s model.

The August et al. (2020) impact predictions (impactful vs non-impactful) for the FSIS sentences are then overlaid onto our visualisation as shown in Figure 9.1b to see how

[1] topic here refers to the subject matter of the documents as opposed to topic model



(a) Sentences from news (red) and science (blue) articles, suggesting discernible differences in connotation and language use.

(b) Sentences predicted as impactful, highlighted in purple, are more numerous in the upper half of the y-axis.

Figure 9.1: t-SNE plot providing a 2-dimensional visualisation of sentence vectors based on V_{conno} space.

well the model characterises scientific impact in FSIS sentences. There is visible separation of sentences predicted as impactful along the y-axis. Impactful sentences are more likely to be nearer the top of the projection (mean=4.86, std=8.11) and non-impactful nearer the bottom (mean=-0.65, std=10.7).

Manual inspection of outliers of the projection shows that whilst the model is able to detect clear cases of impactful communication (e.g. "...reductions of this magnitude would have a substantial effect in preventing heart attacks...") it struggles with more nuanced cases. It also conflates motivating statements with wide scope as impactful ("The unprecedented pandemic... has created worldwide shortages of [PPE]") and fails to identify scientific outcomes where the wider impact is not immediately clear (e.g. "a study... showed how a drug could be created which sticks to tumours, but is only activated when hit by specific waves of light").

Importantly, quantification and comparison of model results is limited by the framing of the task as a binary classification problem. For example, one might consider the discovery of a new species of monkey to be impactful but less so than the development of life-saving medical treatments.

9.3.2 Scientific Impact Scoring (SIS) as a Regression Task

Given the limitations of the methods explored above, we re-frame the detection and measurement of scientific impact as a regression task, Scientific Impact Scoring (SIS). SIS assigns sentences a real-valued score between 0 and 1 depending on the level of comprehensive impact they encapsulate, thus allowing quantification and comparison of impact

across documents.

9.4 Annotating the FSIS corpus

We created the FSIS corpus for training and evaluating supervised models on SIS scoring by annotating sentences in terms of impact through best/worst scaling (BWS) and following a set of guidelines.

9.4.1 Best-Worst Scaling

Best-Worst Scaling (BWS), sometimes referred to as MaxDiff (Jurgens et al., 2012), is an annotation scheme for regression tasks in which respondents are presented with a number of options and asked to partially rank them by picking the most extreme examples (i.e. 'best' and 'worst' or 'highest' and 'lowest'). BWS is an alternative to Likert scales in which respondents are asked to mark one of a series of checkboxes that correspond to positions along a scale (e.g. 1 for worst, 5 for best). BWS has been shown to provide more reliable results versus explicit ranking on a Likert scale for a number of regression-based language tasks including sentiment analysis (Kiritchenko and Mohammad, 2017), semantic similarity (Jurgens et al., 2012) and summarisation (Jurgens et al., 2012; Tang et al., 2021; Suhara et al., 2020).

BWS is also a more efficient way of collecting data than using Likert scales where each task corresponding to a single data point; in BWS each task provides ranking information about multiple data points. If a user is asked to rank four items A, B, C, D, and they select A as worst and D as best, we can infer that $A < B$, $A < C$, $A < D$, $D > C$ and $D > B$. Users would have to complete 4 sets of Likert questions to reveal the same information.

9.4.2 Annotation Guideline Development

SIS is a nuanced and multifaceted task and it is not possible to objectively specify which impact factors may be more or less important as this may vary on a case-by-case basis. An individual's assessment of impact may also depend on their educational background and personal biases. For example, a biologist familiar with *P. falciparum* (as described in section 9.1), may be more likely to rank a sentence about a new medicine for "clearing the dormant *P. falciparum* parasites in infected patients" highly. Conversely, a layperson who is not familiar with the parasite may not make this connection without further research or the inclusion of an explicit reference to malaria.

In order to build consensus, the annotation guidelines for the task were developed collaboratively by three annotators with varied professional and educational backgrounds using an iterative process. The annotators began by listing the impact factors that they felt

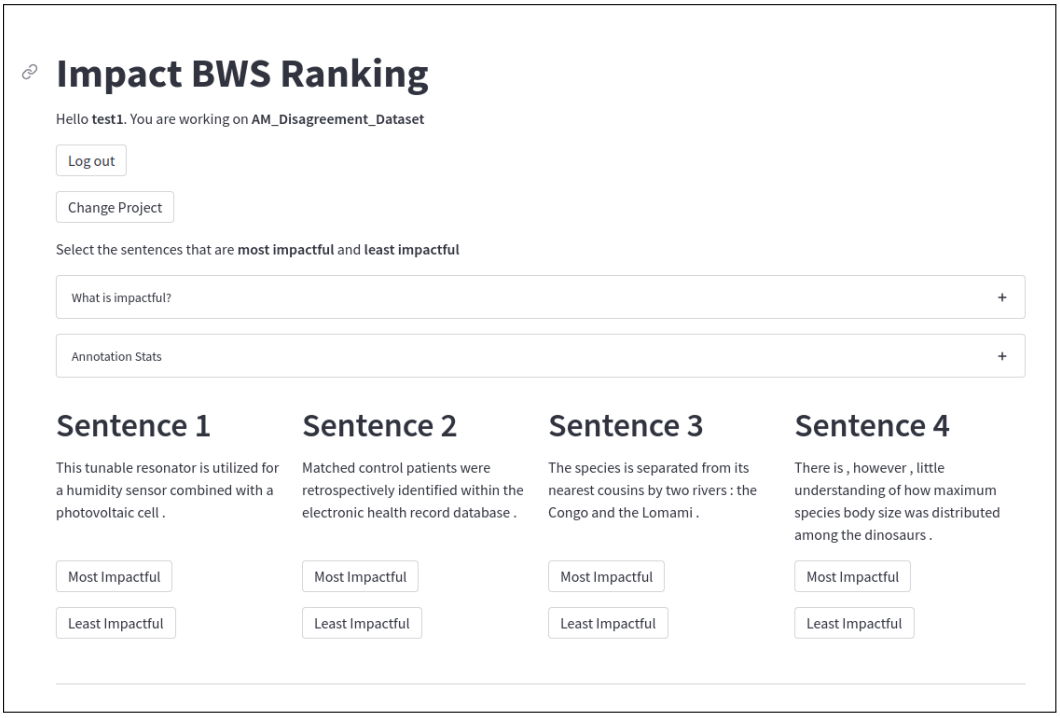


Figure 9.2: A screenshot from our Best/Worst Scaling Tool. The user is asked to select the most and least impactful sentences based on our guidelines.

may indicate that a sentence is impactful. This included newsworthiness (as per Molek-Kozakowska (2017)), the scientific context in which a sentence appears (e.g. penalising sentences that describe background and no new outcomes), scope and strength of the claim (e.g. penalise hedging such as 'may' or 'could'). The annotators then worked together to manually rank a small number of sentences from the unlabelled corpus, discussing their choices and incrementally tweaking and updating the initial set of criteria.

The final guidelines document, along with the code for the annotation tool, are available from our GitHub repository [2].

9.4.3 Annotation Tool

We developed a web-annotation tool which presents annotators with four sentences at a time from either a scientific article or a news article and asks them to pick the most impactful (best) and least impactful (worst) of the four, as shown in Figure 9.2.

Comparison tasks are pre-generated by pseudo-randomly selecting groups of four sentences from the corpus with each sentence appearing no more than once in a single task. Sentences may appear in up to 8 distinct BWS tasks to maximise the chance that they are compared with a variety of other sentences with high and low impact. Following Kiritchenko and Mohammad (2017), we generate 2N tasks (where $N = 430$, the number

[2] https://github.com/ravenscroftj/scientific-impact-scoring

of sentences in the FSIS corpus), resulting in 860 distinct tasks. A tally of the number of times each sentence s is voted as best C^s_{best} or worst C^s_{worst} is kept. For each sentence s, a raw impact score $-1 \leq I^s_{BWS} \leq 1$ is then calculated and scaled to $0 \leq I^s \leq 1$ to better facilitate neural modelling:

$$I^s_{BWS} = \frac{(C^s_{best} - C^s_{worst})}{(C^s_{best} + C^s_{worst})}$$

$$I^s = \frac{I^s_{BWS} + 1}{2}$$

By requiring annotators to consider each sentence multiple times in different contexts and then combining their observations, we are able to generate robust and stable scores for each sentence that are more likely to be in consensus with the other annotators.

9.4.4 Inter-Annotator Agreement and Contentious Subset

The three annotators were asked to annotate an overlapping set of 20% of the BWS tasks so that three independent sets of labels were available for calculation of annotator agreement as an approximation of task difficulty and data reliability.

Rather than randomly sampling 20% of the dataset, we use the t-SNE visualisations (Figure 9.1a) to guide the selection of challenging or contentious sentences. We picked sentences assigned a low I score by annotator 1 that are near the top of the projection ($y \approx 30$) and items assigned a high I score near the bottom of the projection ($y \approx -20$). We call this the “contentious subset".

We tested the statistical significance and validity of the contentious subset to confirm that the sentences therein are outliers with respect to the remaining FSIS dataset. Using the connotation model's output representation vectors, we calculate a centroid vector and measure the cosine distance of each sentence from this point. We found that both the mean and median cosine distances of contentious sentences were greater than those of the non-contentious set, indicating that the contentious set are more likely to be outliers. We validated the significance of this finding by confirming that the two subsets are normally distributed using D'Agostino (1971)'s omnibus test ($p = 0.003$ and $p = 1.21 \times 10^{-121}$ respectively) and using Student (1908)'s t-test which rejects the null hypothesis ($= 0.004$).

We confirmed that the annotators' ratings are normally distributed ($p = 1 \times 10^{-22}$) (D'Agostino, 1971) then measured agreement between annotators (A1, A2, A3) by taking the pairwise Pearson Correlation Coefficient of their I scores: $p(A1, A2) = 0.69$, $p(A2, A3) = 0.72$, $p(A3, A1) = 0.76$

We found strong positive correlation between all three annotators, annotator 2 having a marginally weaker correlation with annotators 1 and 3. Following Suhara et al. (2020), for sentences annotated by all three annotators, the final score is calculated by summing together C^s_{best} and C^s_{worst} from all three people before calculating I^s.

9.4.5 Auxiliary Tasks for Multi-Task Learning

As per our guidelines (section 9.4) the linguistic manifestation of scientific impact is multi-faceted. Thus we examine several related tasks to see whether jointly learning them improves SIS performance. We describe each task and the respective configurations used in our experiments. Since the FSIS dataset is relatively small, we under-sample training data for each auxiliary task to balance learning across the aux task and SIS task during training.
Scientific Discourse Classification (SciClasses) predicts the discourse function of a sentence in a scientific document. This, as discussed in § 9.2, is tightly coupled with whether or not a sentence may exhibit impact. Here we use 200 randomly selected sentences from the RCT-200k dataset (Dernoncourt and Lee, 2017) containing 2.3 million sentences tagged with their scientific discourse label (BACKGROUND, CONCLUSION, METHODS, RESULTS, OBJECTIVE). Although scientific discourse classification is usually framed as a sequential classification task (Liakata et al., 2012a; Dernoncourt and Lee, 2017), here we consider it as a sentence-level classification task.
Scientific Claim Strength Detection (Claim Strength) predicts the strength of a claim made about scientific outcomes in scientific articles and press releases. Claim strength is directly linked to impact, where stronger claims are more likely to confer impact. We use the few-shot claim strength dataset from Wright and Augenstein (2021).
Reading Level Classification (Readability) indicates the ease with which a passage of text may be read. A sentence that is easier to understand by lay-readers is more likely to be perceived as impactful than a complex sentence that uses a lot of jargon. We use a stratified sample of the OneStopEnglish corpus (Vajjala and Lučić, 2018) which contains examples of English sentences taken from newspaper articles that have been rewritten for adult elementary, intermediate and advanced learners.

9.5 Model

We explore both prompt-based and multi-task-fine-tuning learning approaches as well as combinations thereof to attempt to model the SIS task even with a limited quantity of training data available. Here we describe the experimental setup used to test both approaches on our FSIS corpus.

9.5.1 Prompt-Based Modelling and PET

As discussed in section 9.2, Prompt-based learning has been shown to perform well in low resource NLP settings like ours (We have 430 sentences for SIS) Schick et al. (2020). It facilitates few-shot learning for classification tasks by PLMs by exploiting their masked word prediction pre-training task. The model is fed a prompt e.g. “I hated the movie, it was <M>" and predicts the most likely value of the masked token <M>. It learns to maximise

Task	Prompt	Label => Verbalisers
Impact Scoring	<S_1> => impacts: <M> <S_1> => affects: <M>	0 => "reference","details","source", "documentation", "methods" 1 => "society", "everything","world","climate", "humans", "everyone", "profound"
SciClasses	<M>: <S_1>	BACKGROUND => "background" CONCLUSION => "conclusions", "findings", "discussion", "output" METHODS => "method", "controls" OBJECTIVE => "objective","aim", "motivation", "proposed", purpose" RESULTS => "result", "results", "outcomes", "increase", "decrease", "reduce", "improve"
Claim Strength	<M>: <S_1>	0 => "other" 1 => "related" 2 => "sometimes" 3 => "causes
Readability	<S_1> is <M> to understand	Elementary => "simple","easy","straightforward" Advanced => "hard","difficult","complex"

Table 9.1: Prompts and label verbalisers for the PET tasks in this paper. <M> is the token to be guessed and <S_1> is replaced with the full text from the sentence under evaluation.

the likelihood that the masked word belongs to a subset of 'verbalizer' tokens, manually chosen to correspond to the training example's class e.g. "bad" could map to a negative review class. While ours is a regression task, work by Gao et al. (2021) has shown how prompt based learning can also be ported to regression.

9.5.1.1 Pattern Exploitative Training (PET)

Given the versatility of the Pattern Exploitative Training (PET) Schick et al. (2020) framework for prompt based learning we leverage it to train a series of models that can predict the comprehensive impact score for a given input sentence from our corpus in a few-shot setting. The PET framework starts with a pre-trained language model (see section 2.5.7) M with vocabulary V, a pattern function $P(\mathbf{x}) \in V^*$ which transforms each input sentence x into a cloze question with a mask token to be predicted and a verbaliser function $v : \updownarrow \rightarrow V$ to map the label $l \in \updownarrow$ onto V. **p** denotes a Pattern-Verbaliser-Pair (PVP) (P, v).

For a given PVP **p** and input sequence **x**, a score $s_{\mathbf{p}}(l|\mathbf{x}) = M(v(l)|P(\mathbf{x}))$ is calculated for each label $l \in \mathcal{L}$, allowing a probability distribution over all possible labels to be calculated as

$$q_{\mathbf{p}}(l|x) = \frac{exp(s_{\mathbf{p}}(l|\mathbf{x}))}{\sum_{l' \in L} exp(s_{\mathbf{p}}(l'|\mathbf{x}))} \tag{9.1}$$

The categorical cross-entropy (see section 2.6.3.7) between the predicted distribu-

tion $q_{\mathbf{p}}(l|x)$ and the true distribution of the training example $(\mathbf{x}, l)$ is used to fine-tune M on new classification tasks via standard back-propagation as illustrated in Figure 9.3.

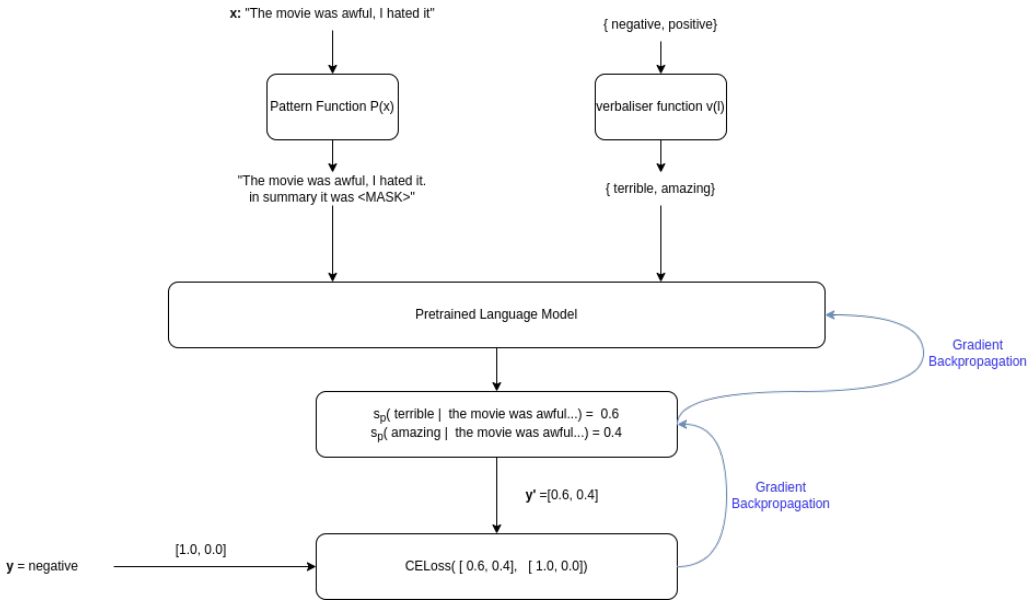


Figure 9.3: Simplified PET model architecture for classification

9.5.1.2 Adapting PET for Regression

As SIS is a regression task, we follow Gao et al. (2021)'s approach which treats regression as the interpolation between two opposing poles $\{y_u, y_l\}$. Specifically we make the following assumptions: Maximum and minimum impact scores (c_u and c_l) are awarded when all annotators rank a sentence 'best' or 'worst' in all contexts respectively and the predicted impact score I^{pred} can be represented as a mixture of probabilities $p(y_u)$ and $p(y_l)$ such that:

$$I^{pred} = c_u \cdot p(y_u|\mathbf{x}) + c_l \cdot p(y_l|\mathbf{x}) \tag{9.2}$$

Verbalisers $\{v_u, v_l\} \in V$ are assigned to represent extreme values $\{y_u, y_l\}$ respectively, and the probability of y_u (similar for y_l) can be calculated:

$$p(y_u|\mathbf{x}) = \frac{exp(s_p(v_u|\mathbf{x}))}{\sum_{v' \in \{v_u, v_l\}} exp(v_{v'}|\mathbf{x})} \tag{9.3}$$

The language model is then fine-tuned the using KL-divergence (KL-Div) (Kullback and Leibler, 1951) between the inferred $p(y_u|\mathbf{x})$ and the true weight mixture I^s.

9.5.1.3 Prompt Selection and Knowledge Distillation

Prompt-based model performance is strongly dependent upon the specific prompts and verbalisers used for each task (Schick and Schütze, 2021; Gao et al., 2021). These are normally defined manually but approaches such as PETAL (Schick et al., 2020) and that of Gao et al. (2021) have been used to automatically generate suitable prompts and verbalisers. However, these approaches require their counterpart output as input; either a manually constructed set of prompts or verbalisers is always required as a starting point. The prompts and verbalisers used for SIS (shown in Table 9.1) were developed through a combination of manual iteration and use of PETAL.

PET allows the combination of different prompts by training an ensemble of models using different PVPs which are used to generate soft labels for a large unlabelled set of sentences sourced from full text news and science articles. These labels are then combined using to train a final model knowledge distillation, a simple method for compressing the knowledge in an ensemble of models into a single model (Hinton et al., 2015).

9.5.2 Multi-Task PET (MT-PET)

We are the first to introduce both regression and classification into an MT-PET (Wright and Augenstein, 2021) setting. We test whether PET-based models perform better at SIS when jointly learning related tasks. Each model in the ensemble is trained on two tasks at a time: a primary task (P_m, v_m) and an auxiliary task (P_a, v_a). v_m and v_a map task labels onto the same vocabulary V and P_m and P_a both generate prompts in V^*. Thus, the model can learn both tasks without architectural changes by randomly alternating between them at each training step. The ensemble provides soft labels for an unlabelled set of examples for the main task only which are then used to train a final model. As per Wright and Augenstein (2021)'s implementation, tasks with smaller datasets are repeated during each epoch in order to balance the model's exposure to each task. We trained MT-PET configurations for all auxiliary tasks described in Section 9.4.5 using the prompts shown in Table 9.1.

9.5.3 Fine-Tuned RoBERTa Models

We compare our PET models to RoBERTa-based models trained using standard fine-tuning for impact score prediction:

FT RoBERTa: A randomly-initialized feed-forward layer with a single regression output standard is attached to a RoBERTa model initialized using weights from Liu et al. (2019a). Embeddings produced by RoBERTa are pooled and passed through the dense layer which is fine-tuned using the mean-square error loss (MSE) versus I_s.

FT August Impact: These experiments use the same configuration but initialise RoBERTa using weights from August et al. (2020)'s Impact model.

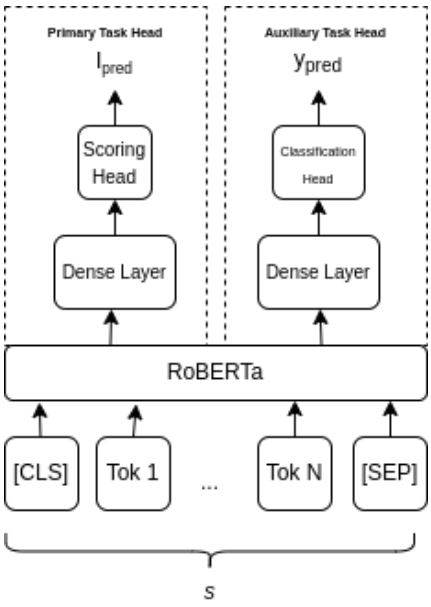


Figure 9.4: Simplified model architecture for Fine-Tuned Impact + multitask RoBERTA

When impact score prediction is jointly learned with a supporting task, the model architecture is initialised as above but an additional randomly-initialised feed-forward classification head is attached and fine-tuned on one of the auxiliary tasks via categorical cross-entropy loss. During training, the optimiser switches between each task after each mini-batch, effectively allowing the model to learn both in tandem. Both configurations of our RoBERTa architecture are represented in Figure 9.4.

9.5.4 Experimental Setup

SIS Task: We create a few-shot training set for SIS by randomly selecting 50 and 100 sentences from our annotated FSIS data set. The remaining 380 and 330 sentences respectively are then used to evaluate model performance. We re-run each experiment using three random seeds for sampling of FSIS and aux. data and model parameter initialisation and report averaged results.

Sciclasses: We developed verbalisers and prompts for this task and used them to train a PET-based sentence classifier with $F1_{micro} = 0.74$ which we jointly learn with SIS in our MT-PET experiment. We use categorical cross-entropy optimisation on sentence/label pairs when we use this task in our multi-task RoBERTA model.

Claim Strength: We use the same setup (data set, prompts and verbalisers) as Wright and Augenstein (2021) which we jointly learn with SIS in our MT-PET experiment and we use categorical cross-entropy optimisation on sentence/label pairs in our multi-task RoBERTA model experiment.

Readability: A standalone PET-based model trained on the three OneStopCorpus com-

	F1 Score ↑		R^2 Score ↑		MAE ↓	
Few Shot Examples	k=50	k=100	k=50	k=100	k=50	k=100
Majority Class Baseline	0.558	0.520	-	-	-	-
August IMPACT Classifier - no fine tuning	0.521	0.539	-	-	-	-
PET RoBERTa Large	0.568	0.655	-0.158	0.104	0.234	0.206
MT-PET RoBERTa Large + SciClasses	0.642	0.693	0.076	0.119	0.206	0.191
MT-PET RoBERTa Large + Claim Strength	0.622	0.652	0.043	0.140	0.213	0.200
MT-PET RoBERTa Large + Readability	0.583	0.585	-0.103	-0.019	0.226	0.221
FT RoBERTa Base	0.533	0.631	-0.031	0.180	0.226	0.203
FT RoBERTa Base + SciClasses	0.618	0.712	0.059	**0.319**	0.205	**0.175**
FT RoBERTa Base + Claim Strength	0.667	0.678	0.162	0.272	0.197	0.180
FT RoBERTa Base + Readability	0.613	0.657	-0.046	0.014	0.224	0.211
FT August Impact	0.674	0.726	0.230	0.243	0.192	0.184
FT August Impact + SciClasses	**0.739**	**0.731**	**0.263**	0.240	**0.178**	0.179
FT August Impact + Claimstrength	0.664	0.710	0.090	0.259	0.202	0.181
FT August Impact + Readability	0.640	0.689	0.082	0.125	0.208	0.197

Table 9.2: Results for all experiments. ↑ = higher is better, ↓ = lower is better. Best in model group for k is shown underlined. Best overall result for k is shown in **bold**.

plexity levels yielded $F1_{micro} = 0.58$, most often confusing adjacent classes (Elementary with Intermediate, Intermediate with Advanced). We simplified the task to classify just Elementary versus Advanced, using the 100 samples from those two classes only, yielding $F1_{micro} = 0.87$. We use this simplified task as our auxiliary in both PET and multi-task RoBERTa settings.

9.6 Results & Discussion

Results for all experiments are reported in Table 9.2.

SIS is a regression task and model performance is is reported in terms of Mean Absolute Error (MAE) and coefficient of determination (R^2 score). However, classification performance is also measured for comparison with earlier work. We compare against two baselines: the unmodified August et al. (2020) impact classification model and a majority class baseline given the slight imbalance towards $I_s \geq 0$ in our dataset. Compatible class labels y^s from our normalised, continuous impact scores $0 \leq I^s \leq 1$ are produced by thresholding for each sentence s:

$$y^s = \begin{cases} 1 & \text{if } I^s \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

The same thresholding technique is applied to the output of each of our regression models I^{pred} in order to produce y^{pred} against which we calculate micro-F1 for our models.

The August et al. (2020) impact classifier is outperformed by the majority class

baseline suggesting that even when re-framed as a classification task, our setting is different enough that this model cannot generalise to it without adaptation. We believe this is due to the different nature of the scientific papers and press releases. However, fine-tuned RoBERTA-based models initialized with August et al. (2020)'s model weights (FT August Impact) give the best overall performance for $k = 50$, suggesting that there is commonality between the two tasks that allows the model to adapt effectively to the new task setting with limited training. FT August Impact gives the best impact classification performance for $k = 100$ but is outperformed by FT RoBERTa in terms of MAE and R^2, suggesting that it may be trapped in a local performance maxima by its original training.

We experimented with both RoBERTa-large and RoBERTa-base configurations and pre-trained weights (Liu et al., 2019a) for all PET and FT tasks. We found that all PET experiments initialised with RoBERTa-base performed worse than those using RoBERTa-large and conversely, all FT RoBERTa experiments initialised with RoBERTa-Large performed worse than those using RoBERTa-Base. We hypothesise that this may be a side effect of the different training regimes (prompting & KL-Div vs new parameters & MSE) which could be explored in future work. The worse configuration is omitted for each architecture respectively.

When no aux. task is learned, PET outperforms FT RoBERTa at classification but not regression for both $k = [50, 100]$. MT-PET + SciClasses also outperforms FT RoBERTa Base + SciClasses for $k = 50$ in terms of both F1 and R^2 and yields a similar MAE. However, the remaining MT-PET configurations yield similar but slightly worse performance. PET's relatively poor regression predictions compared to fine-tuned RoBERTa settings may be explained by the fact that it is a classification architecture adapted for regression-like tasks using the method discussed in Section 9.5.1 whereas FT RoBERTa has a task-specific regression head. A more robust exploration of PET's regression capabilities could be the focus of future work.

Furthermore, comprehensive scientific impact is a complex, abstract concept that is difficult to describe succinctly and despite following best practices for selection of prompts and verbalisers, these may also be limiting PET performance. These results highlight the complexity involved in configuring PET, particularly for regression-like tasks and subsequently, the trade-offs between PET and standard fine-tuning approaches, especially when fine-tuning involves models trains on similar tasks.

For all three model architectures, jointly learning impact scoring with either SciClasses or Claim Strength as the supporting task always improves performance compared to learning the impact scoring task alone, supporting our hypothesis that these subtasks are intrinsically linked with the core impact scoring task as discussed in Section 9.3. SciClasses is the most helpful aux. task, yielding best regression performance for $k = 100$ via FT RoBERTa and for $k = 50$ via FT August. Readability has an inconsistent effect, in some cases improving classification performance a modest amount and in almost all cases,

worsening both R^2 and MAE. This could imply that language complexity is not important when it comes to describing scientific impact. However, another possibility is that the task is too different to the impact scoring task to confer any advantage.

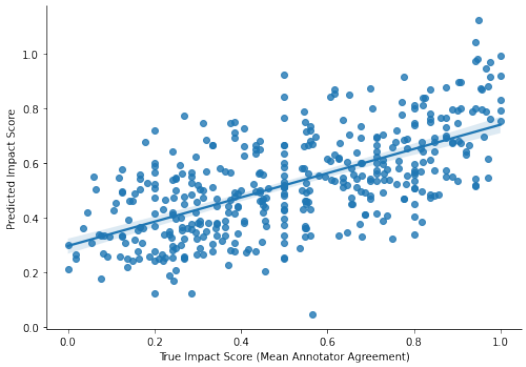


Figure 9.5: Scatter plot comparison of Mean Predicted vs Ground Truth Impact Score for FT RoBERTa + SciClasses

9.6.1 Error Analysis

Figure 9.5 shows sentence-wise comprehensive impact score versus averaged predicted scores from the best performing model (FT RoBERTa + Sciclasses, k=100). The model yields tighter plots at either end of the scale where impact, or lack thereof, is relatively straightforward to detect. The averaging strategy employed to combine annotations (see S9.4) assigns middling scores to sentences where annotators strongly disagree and this uncertainty is reflected by the model which yields more sparse predictions around the middle of the scale and a vertical stripe for $I = 0.5$.

We carried out a qualitative analysis of a small number of sentences which are selected due to high variance of human annotations indicating annotator disagreement or due to large absolute difference between mean score and predicted score indicating poor model performance. These sentences along with normalised impact scores assigned to them by annotators (A1, A2, A3) and the mean predictions of FT RoBERTa + Sciclasses + k=100 model are shown in Table 9.3.

Sentence #1 exemplifies the smoothing effect of the BWS score aggregation on varied annotator scores. Whilst the sentence does mention specific outcomes, as noted by A1 and A3, these outcomes are fairly abstract and likely of limited practical use to wider society, reflected by A2. On balance, the mean score assigned seems to reasonably capture these considerations and the model is able to accurately reproduce it. Sentence #2 yields some annotator disagreement and the model overshoots the mean impact score significantly. The model likely picks up on the explicit statement that impact is produced but the context (asteroid impacts not scientific impact) is implicit and requires reading comprehension and background knowledge to discern. The FSIS corpus contains a number of sentences allo-

#	Sentence	A1	A2	A3	Mean Score	Predicted Score
1	Our results suggest that the African wolf is a relatively ancient gray wolf lineage with a fairly large , past effective population size , as also suggested by the Pleistocene fossil record	0.8	0.2	1.0	0.67	0.69
2	A steeply - inclined impact produces a nearly symmetric distribution of ejected rock and releases more climate - changing gases per impactor mass than either a very shallow or near - vertical impact	0.2	0.66	0.5	0.45	0.88
3	At least one adverse reaction within the first 7 days after the vaccination was reported in 30 (83 %) participants in the low dose group , 30 (83 %) participants in the middle dose group , and 27 (75 %) participants in the high dose group	0.5	1.0	1.0	0.83	0.69

Table 9.3: Some challenging sentences, their normalised annotator scores (A1,A2,A3), mean annotator score and mean predicted score from all three random seeds for FT RoBERTa + SciClasses + k=100

cating high impact scores to mentions of "climate change", which is also mentioned as a by-product of the asteroid impact in sentence #2 and could also trigger the production of a high impact score. A2 and A3 are in agreement that sentence #3 has a high impact. However, A1 is more cautious, possibly due to ambiguity about what the vaccine was for and what the implications of adverse reactions on the trial groups are. Despite these ambiguities, the model is able to generate a prediction 0.14 from the actual mean score.

9.7 Conclusion & Future Work

In this chapter, we have presented the Scientific Impact Scoring (SIS) task and accompanying Few-shot Scientific Impact Scoring (FSIS) corpus, reframing detection of comprehensive scientific impact as a regression task. We evaluate a number of few-shot, multitask model architectures and for the first time, we combine regression and classification tasks in an MT-PET multitask prompting setting. We outline a set of auxiliary tasks that boost performance when jointly learned with SIS. Our best model improves on the legacy classification task by 18% compared to the baseline.

Our comparison of few-shot learning approaches on the SIS task unveils the complexity involved in using prompt-based learning and its trade-offs with standard fine-tuning training regimes. Whilst prompt-based models have been shown to be highly effective in domains similar to ours, our classifier trained with standard fine-tuning outperforms our prompt-based methods by a significant margin.

Future work could investigate whether sequential rather than independent sentence

models could improve regression performance. Further investigation into the performance of PET versus 'traditional' fine-tuning for this task and other regression tasks would also be of interest. Furthermore, new automated approaches to help with selection of prompts and verbalisers may make it easier to train more performant models with PET.

Our corpus, annotation guidelines, code and model weights are available via our Github repository[3].

[3] https://github.com/ravenscroftj/science-impact-scoring

10

Conclusions & Future Works

In this chapter we outline the primary findings of our work with respect to the research questions posed in section 1.2 before briefly outlining a number of interesting and promising directions for possible future work in this field.

10.1 Primary Findings

RQ1: To what extent does the academic impact of scientific work relate to its impact beyond academia (comprehensive impact)?

RQ1.1 What are commonly used metrics for the academic impact of scientific work?

In Chapter 3 we explore and discuss existing academic metrics for scientific impact. We find that most existing academic impact metrics focus on citation and publication counts varying by how these counts are aggregated. JIF (Garfield, 2006) is widely used for comparing journals whereas h-index (Hirsch, 2005) is commonly used for comparing and measuring the impact of academics and departments. Since citation behaviour varies across disciplines and between time periods, MNCS (Waltman, 2015) can be used to normalise citation counts to facilitate fairer comparisons of citation-based impact across disciplines and between different periods in time. We also briefly explore altmetrics which aggregate data about online interactions with scientific works e.g. numbers of re-tweets or likes. Compared to citation-based metrics, altmetrics are relatively new and whilst some metrics are beginning to see wide usage, their coverage varies significantly by discipline (Banshal et al., 2019).

RQ1.2 What metrics exist for measuring impact beyond academia and, how widely are they used?

We also discuss a selection of comprehensive impact metrics in Chapter 3. Since comprehensive impact is broad and expressed in many ways, it is difficult and often expensive to describe metrics that generalise across multiple disciplines. This challenge has given rise to partial impact metrics like the work of Steingard et al. (2022) that focuses on sustainability and Sørensen et al. (2022) which is only applicable within Occupational Health. Broader metrics like STAR METRICS (Lane and Bertuzzi, 2010) and REF Impact Score (HEFCE, 2012, 2017) provide more robust metrics that facilitate interdisciplinary comparison but are costly to apply and rely on centralised government funding and coordination which makes them difficult or impossible to apply independently.

RQ1.3 What statistical relationships exist between widely used academic impact metrics and scientific impact outside of academia?

In Chapter 4 we conduct an empirical investigation into whether academic impact metrics and scientific impact metrics correlate by comparing widely used citation metrics and altmetrics with REF Impact scores. We demonstrated that none of the academic impact metrics correlate well with REF Impact score ($r \in [-0.080, 0.141]$) and our attempt to train a regression model that predicts REF impact from academic impact yielded poor results ($R^2 = -0.113$).

RQ1 Findings Summary

We have introduced and defined the two types of scientific impact: academic impact that concerns a work's affect and influence within academia and comprehensive impact that concerns a work's wider effects and influence on society, economy, policy and the environment. We have discussed current, widely used academic impact metrics including both citation-based and alternative metrics. We have discussed the strengths and weaknesses of these metrics and the objections raised about how whether they are being used appropriately by the wider academic community. We discussed current state-of-the-art comprehensive impact metrics and the challenges associated with trying to measure comprehensive impact which has a very broad and multi-faceted set of corresponding outcomes. We introduce the Research Excellence Framework (REF) Impact Score, a general comprehensive impact metric that is used by all academic institutions within the UK but which is only calculated once every 4 years at great expense. Finally, we construct experimental datasets consisting of over 7 million open-access scientific papers and 643 REF impact case studies from the 2014 REF assessment and carry out statistical analyses to understand the extent to which commonly used academic impact metrics correlate with REF impact. We find a lack of correlation between academic metrics and REF impact score, motivating us to find new ways to measure comprehensive impact.

RQ2: How does evidence of impact beyond academia affect a scientific work's performance in terms of existing comprehensive impact metrics?

RQ2.1 What data sources provide evidence of non-academic impacts of scientific work?

In section 5.2 we discuss several sources of evidence for non-academic impact including commercial records, parliamentary proceedings and news articles.

Commercial information such as patents and statutory financial information can demonstrate that a scientific work has led to commercial success via a spin-out company. However, this sort of information is not available in all countries and may be incomplete in cases where it is available due to complex ownership structures (e.g. holding companies)

and companies making the decision not to use patents to protect their invention. Furthermore, commercial success may be irrelevant or inappropriate for measuring some scientific impact. For example in medicine we may be more interested to understand the reduction in mortality rates provided by a new treatment rather than how profitable it is to produce said medicine.

Parliamentary proceedings can capture evidence that scientific work has had an impact on public policy and likewise, governing bodies (e.g. UK's National Health Service) usually cite any research that supports the policies that they implement in corresponding documentation. However, like commercial evidence, policy documents and parliamentary proceedings may be more difficult to obtain in some geographies. Furthermore, these data sources have a selection bias whereby only a very small number of works end up affecting policy versus other works which are still likely to produce comprehensive impact in other ways.

Finally, we explore News as a source of evidence for comprehensive impact. Scientific journalism is ubiquitous internationally and news is often easy to access. Whilst news still has some selection bias, since not all scientific work that produces impact may be considered newsworthy, news coverage is more comprehensive than that of financial or political records. Furthermore, coverage is often more diverse, there are a large number of reasons that a scientific work may be considered newsworthy.

RQ2.2 How can we automate the extraction and linking of such evidence to the scientific work it relates to?

In Chapter 5 we construct a pipeline for semi-automatically matching news articles to scientific papers. We process news articles found within a large, general purpose web archive of all websites with a *.uk* domain name. We find that many scientific news articles provide hyperlinks or unique document identifiers that link to the scientific papers that they discuss and that these identifiers can usually be used to automatically retrieve the corresponding scientific paper for processing.

We also produce a data collection and annotation platform which automatically analyses news articles, identifying those that discuss scientific work and retrieving candidate scientific papers from the Microsoft Academic Knowledge Graph dependent on the publication date of the news article and the named entities discussed within (for example using 'Dr Foster' and 'Oxford University' from 'Dr Foster from Oxford University led the work...' as search criteria). We also address **CH3: Data collection and annotation with multiple annotators** by providing users of our system with clear instructions detailing how to use it and by measuring the extent to which they agree with each other's article matches.

In Chapter 6 we introduce an information retrieval task in which we separate the pairs of news articles and corresponding scientific papers and, using the news article as the query, we attempt to retrieve its paired scientific paper from a large collection of scientific

papers that includes distractors. We show that it is possible to use TF-IDF representations of news articles to retrieve their corresponding scientific paper with nDCG $\in [0.76, 0.93]$ for our dataset. In section 10.2.4 below we propose further exploration of this technique as a way to fully automate news/science pair linking.

RQ2.3 In terms of current comprehensive impact metrics, how does the presence or absence of evidence of comprehensive impact affect a work's associated score?

In section 5.4 we carry out a statistical analysis of the relationship between news coverage and REF Impact score for scientific papers that are associated with both one or more news articles and a REF impact case study from the 2014 REF assessment. We show through a Kolmogorov-Smirnov 2-sample test (Massey, 1951) that scientific works *with* associated news coverage and scientific works *without* associated news coverage can be treated as two separate populations ($p = 0.007$). Furthermore, we apply a two-sample bootstrap test of mean difference (Hesterberg, 2015) which yields a 95% bootstrap percentile confidence interval of [0.07,0.27] which implies that the set of scientific papers *with* associated news coverage tend to have a higher REF Impact score than those *without*. This finding is useful and has implications for future REF assessments. However, it does not yield any insight into how or why this may be the case. We suggest possible extensions to this analysis in section 10.2.3 below.

RQ2 Findings Summary

We have shown that news articles provide a ubiquitous, easy to access source of evidence for comprehensive impact, limitations notwithstanding. We have constructed a pipeline for automatic collection of scientific news articles and for linking those news articles with corresponding scientific papers by using hyperlink metadata from news stories to identify the scientific works that they discuss. In cases where this metadata is missing, we have built a corresponding annotation tool which uses machine learning to identify candidate matches and, we have proposed some strategies for automating this part of the process. Finally, we have shown via statistical analysis that presence of news articles corresponds to higher REF impact scores within our dataset.

RQ3: How can we identify semantically similar statements in topically aligned news articles and scientific papers despite the disjoint language use and grammatical styles of the two document types?

RQ3.1 What are the statistical characteristics of news articles and scientific papers and, how do they differ?

In section 6.2 we characterise the news articles and scientific papers in our corpus. We also characterise news articles by source: quality news articles which tend to be more complex

and target an intellectual audience and tabloid articles which tend to be simpler and shorter targetting a younger, less educated audience and focussing on topics that appeal to a broader audience such as celebrity and sport (Bastos, 2016).

We find that the mean length of a scientific paper is 7,718 words whilst that of a news article is 1,012. Furthermore, we also show that broadsheet news articles are typically longer (mean 804 words) than tabloid articles (mean 524 words). We note that existing research (Pitler and Nenkova, 2008) shows an inverse relationship between length and readability and our findings align with this conclusion; scientific papers are verbose and precise because they explain in detail how scientific work was carried out and how it can be reproduced whereas broadsheet articles provide a summary of scientific work and tabloid articles, an even simpler, more succinct summary.

We find that scientific papers typically have vocabulary sizes that are between 14-28% larger than their corresponding news articles. Again, this is intuitive since scientific papers are written for a niche, highly technical readership and use very precise, discipline-specific words whereas news articles are written for a broad, diverse readerships and use simplified, common wording. We record a mean vocabulary overlap between news articles and the scientific papers that they discuss of between 7-12% again highlighting the difference between the two document types despite their shared subject-matter.

In section 6.4.3 apply LDA (Blei et al., 2003) topic modelling to pairs of news articles and scientific papers and find that pairs of news articles and the scientific papers that they discuss tend to have significantly different topic distributions. This is likely a reflection of the differing structures of the two documents; scientific papers typically follow a conventional set of core regions, describing the background of the work, the methodology, observations and conclusions (Liakata, 2010) whereas news articles tend to summarise the scientific work, interspersing specifics of the research conducted with general statements about how this may affect the reader (Louis and Nenkova, 2013b).

RQ3.2 What techniques can we use to best represent news articles and scientific papers for comparison?

In sections 6.3.1 and 6.3.3 we describe and apply a selection of text encoding strategies to pairs of news and scientific papers. We evaluate these methods quantitatively via an information retrieval task in which we use news articles to retrieve corresponding scientific papers from a pool of several thousand. We found that, of the encoding techniques that we evaluated, the most technique effective for this task was TF-IDF (Jones, 1972) bag-of-words encoding which achieves an nDCG between 0.76-0.93 across the different news sources. We hypothesize that, whilst typically limited, the overlapping vocabulary between scientific papers and news articles must contain specific and relevant terms that can be used to identify the correct corresponding document. We show that this is the case for randomly chosen pairs of documents in Table 6.4. All other encoding strategies achieve comparably

poor results.

RQ3.3 What contextual information do the document representations provide?

In addition to the quantitative evaluation of encoding methods described above, we also carry out a qualitative exploration of the encoding mechanisms, aiming to understand the extent to which context about how information is transferred between news articles and scientific papers is captured by the different encoding mechanisms. Whilst TF-IDF and word count vectors yield the best retrieval performance, the simplicity of these approaches provides little in the way of context beyond the ability to see which terms overlapped. On the other hand, LDA provides insight into the structure of each document with respect to the learned topics defined within the model. The topic distributions themselves can also give insight into which words commonly appear together in the same context. However, LDA is built on top of a bag-of-words representation does not retain document structure information. Doc2Vec representations are arbitrary, multidimensional vectors which, unlike LDA, do not have any intrinsic meaning. We conclude that lower-level text representations are required in order to be able to understand how the same information may be discussed across two related documents.

RQ3 Findings Summary

We show that news articles and the scientific papers that they discuss have very different natures despite their shared subject-matter and that this aligns intuitively with the distinct goals and readerships of these two document types. We evaluate a series of document encoding techniques with the objective of understanding how to identify corresponding excerpts of the two documents that discuss the same subject-matter despite differences in structure and vocabulary. We find that, counterintuitively, TF-IDF provides the best performance in our evaluation but that none of the encoding approaches evaluated provide enough context to help us align similar excerpts from documents. We note that this work was carried out before the widespread adoption of Attention (Bahdanau et al., 2016), Transformers (Vaswani et al., 2017) or BERT (Devlin et al., 2019) and that there are likely a number of more recent works that could better facilitate alignment of similar excerpts. We propose some new investigations into these technologies below in section 10.2.5

RQ4: How does information flow from scientific papers to the news articles that discuss them and what is the connection between REF impact scores and how works are presented in the news?

RQ4.1 How do the structures and styles of news articles and scientific papers affect the perceived importance of information presented within them?

In Chapter 7 we discuss the concept of Prominence, how important information *seems* based on its presentation. On the other hand, salience describes the actual importance of information regardless of its presentation. We illustrate these concepts with a billboard: the bold font telling the reader to buy the product is prominent but not salient and the small print detailing the terms and conditions of the offer is salient but not prominent. We discuss the structure of news articles and how journalists will take advantage of news publishing conventions including use of large fonts and repetition to prominently feature information that they consider important. Likewise, we discuss how scientific discourse annotation schemes like CoreSC (Liakata, 2010) can help us to understand which parts of scientific papers are most frequently presented prominently in news articles.

RQ4.2 Which information in scientific papers is most often presented as most important in news articles?

In section 7.3 we carry out an experiment to identify which information in scientific papers is most frequently presented prominently in news articles. We split pairs of news articles and corresponding scientific papers into collections of sentences. A text-rank based graphical ranking algorithm is then developed to identify the most prominent sentences in news articles, taking advantage of the repetitive structure of these documents to identify the most prominent sentence in the document. The sentences from the scientific paper are labelled with CoreSC discourse labels. Finally, we use text representation techniques, some from Chapter 6 and some new, to carry out pairwise comparisons between prominent sentences from news articles and sentences from corresponding scientific papers. Our findings indicate that sentences that discuss the background or goals of the scientific work are often most similar to prominent sentences from news articles.

RQ4.3 What characteristics are different for document pairs with known links to REF Impact Case Studies?

In section 7.4 we show that for news articles and scientific pairs with a known link to a REF Impact case study result, the mean semantic similarity between sentences is usually higher than for pairs of documents for which we do not have a link. This effect is most visible in the Bag-of-Words encoded sentence pairs and most pronounced for Outcomes and Method sentences.

RQ4 Findings Summary

We explore the document structures of news articles and scientific papers and illustrate prominence and salience and how these concepts relate to news articles. We exploit journalistic conventions for article structure in order to automatically identify a 'most prominent' sentence for news articles within our corpus and use similarity between sentence representations to find corresponding similar sentences from scientific papers. We apply a sentence-level scientific discourse labelling scheme to scientific papers in our collection in order to identify which sections of scientific papers sentences that are most similar to prominent news excerpts appear in. Our findings suggest that that prominent sentences in news articles are, on average, most similar to background and goals sentences and least similar to outcome sentences. Furthermore, when we compare the mean similarity of sentence pairs from documents associated with REF Impact case studies and those *not* associated with REF studies, we see higher semantic similarity for REF associated sentence pairs.

RQ5: Can we identify co-referring entities, such as people, institutions and companies across news articles and scientific papers and use them to align semantically similar sentences despite the stylistic and linguistic differences between the documents?

RQ5.1 How does cross-document co-reference resolution differ for documents in separate domains?

In Chapter 8 we introduce a novel setting for Cross-Document Co-reference Resolution (CDCR) which we designate *Cross-Domain* Cross-Document Co-reference Resolution (CD^2CR). In standard CDCR, both documents are normally from the same domain e.g. they are both news articles. However, as we discuss in section 6.2, scientific papers and news articles have different structures, vocabularies and lengths due to their distinct audiences and communication goals. Therefore, on top of the same set of challenges faced by CDCR, CD^2CR additionally addresses the need to translate between mentions that are expressed using disjoint vocabulary or contexts (E.g. "catalyst chemicals" vs "zeolytes").

RQ5.2 How can pairs of co-referent phrases in news articles and scientific papers be identified and annotated efficiently and reliably?

As CD^2CR is a new setting, no prior training data for this task was available, necessitating the collection and annotation of new data. In section 8.4 we describe a workflow that we built for annotating CD^2CR data based on entity mentions within new/science document pairs in our HarriGT Corpus (described in chapter 5). We devise a set of data annotation guidelines and build a custom annotation tool which seeks to make the task of collecting and annotating CD^2CR data simple. We also attempt to accelerate the task of annotation by

suggesting possible co-references, generated via a BERT-based (Devlin et al., 2019) scoring approach. We asked all annotators to provide overlapping labels for a randomly sampled subset of the data in order to evaluate inter-annotator agreement and we achieved moderate agreement ($\kappa_{fleiss} = 0.554$).

RQ5.3 How well do current state-of-the-art cross-document co-reference resolution models perform at the cross-domain task and where do they struggle?

In section 8.5 we use our CD^2CR corpus to train and evaluate a set of state-of-the-art CDCR models, also comparing with a pre-trained baseline from previous work (Cattan et al., 2020). Across all model implementations, we achieve only modest results ($F1_{muc} = 0.58$, $F1_{B^3} = 0.44$). Subsequent error analysis shows that the models tend to struggle with three particular challenges: cross-document anaphora and exophora resolution and word meaning disambiguation (e.g. knowing the difference between monarch butterflies and monarchs in the royal sense), subset and membership relationships (e.g. knowing that 'Pre-columbian cultural developments' count as 'human activity') and general paraphrase resolution (e.g. 'existing techniques' versus '[existing] efficient catalysts'). We also note that RoBERTa's BPE mechanism may be compressing text representations and reducing performance. A recent state-of-the-art work (Held et al., 2021) which builds on our corpus has achieved $F1_{muc} = 0.87$ and $F1_{B^3} = 0.79$

RQ5 Findings Summary

We introduce a new cross-domain setting for CDCR which translates between domains as well as documents as opposed to standard CDCR which usually involves processing documents from within the same domain. We build annotation tooling and guidelines for our new task setting, collecting a 528 document dataset with moderate inter-annotator agreement. We show that CD^2CR is a challenging task for transformer-based models which struggle to discern some types of entity relationships. Although our models yield modest performance, our dataset is a useful asset for downstream development, containing 7,000 mentions that can be used to identify topically aligned sentence pairs. Furthermore, recent work (Held et al., 2021) which yields significantly improved model performance on our dataset could be used to generate new news/science sentence pairs.

RQ6: How can we detect and quantify comprehensive impact implied by language use in scientific news articles and the academic works that they discuss?

RQ6.1 What is the current state-of-the-art approach for detecting impactful sentences in scientific news articles and press releases, and what are its limitations?

In chapter 9 we explore existing work by August et al. (2020) who provide a binary classification model for detecting scientific impact in news articles and university press releases.

We note that whilst the model performs well at detecting major impacts (e.g. detection and prevention of heart attacks), it struggles with more nuanced impact (e.g. discovery of a new species of monkey).

RQ6.2 Can we define an appropriate task for quantifying comprehensive impact score based on text and can we reliably collect corresponding labelled data from multiple annotators?

In section 9.3.2 we define Scientific Impact Scoring (SIS), a regression task in which sentences are allocated a real-valued score depending on how impactful they are perceived to be. Subsequently, in section 9.4 we outline a protocol for collecting SIS annotations. We iteratively develop a set of annotation guidelines, inspired by Molek-Kozakowska (2017) and Louis and Nenkova (2013b) and taking into account linguistic phenomenon that increase the perceived newsworthiness of the science being reported. We design an annotation tool which captures multiple data points per user interaction via best-worst scoring. We focus our annotation on a subset of the CD^2CR corpus developed in Chapter 8. We combine data from three annotators to create a 430 sentence dataset.

RQ6.3 Is it possible to train a statistically performant, low-resource model for detecting and scoring text that implies comprehensive scientific impact?

We experiment with low-resource learning techniques to allow us to train statistically performant models despite the relatively small number of data points we were able to collect. We use a multitask variant of a popular, transformer-based prompting model (Wright and Augenstein, 2021; Schick and Schütze, 2021) to jointly learn our SIS task alongside auxiliary tasks which we hypothesised would improve the model's ability to perform SIS. We adapt the model to allow it to jointly learn auxiliary classification tasks alongside a primary regression task. We also evaluate a RoBERTa-based (Liu et al., 2019a) multitask model to which we add task-specific output heads for primary and auxiliary tasks. We find that our auxiliary tasks improve performance for almost all model configurations. The best performance is obtained by initializing our RoBERTa-based architecture with model weights from August et al. (2020)'s legacy model.

RQ6 Findings Summary

After evaluating a current state-of-the-art model (August et al., 2020) for detection of comprehensive impact in news articles and press releases, we define a regression task in which sentences are assigned a real-valued score depending on how impactful they are perceived to be. We construct a set of annotation guidelines and tools, constructing a dataset of 430 sentences with associated impact score. We use few-shot, multitask learning to develop a series of models to carry out the SIS task, yielding a model that outperforms the previous

state-of-the-art model by 18%. We demonstrate that whilst prompt-based modelling is a promising low-resource learning strategy for NLP, it is complicated to configure and tune and does not always yield the best results.

10.2 Directions for Future Work

10.2.1 Exploration and Characterisation of SIS Task

In Chapter 9 we introduce SIS, a sentence-level comprehensive impact scoring task that assigns a real-valued score $I_s \in [0, 1]$ depending on level of comprehensive impact implied in a given sentence, and we show that the task can be carried out well by machine learning models. Future work could seek to characterise SIS and understand how it relates to other metrics and works in this field. For example, SIS could be applied to documents in the corpora constructed in this thesis and the results could be used to carry out correlation studies of SIS compared to other impact metrics such as REF Impact Score, Altmetric score and citation metrics.

10.2.2 Exploring Heterogeneous Data Sources for Evidencing Comprehensive Impact

Comprehensive impact is broad, multi-faceted and sparse in nature; one scientific work could impact global health policy and increase global life expectancy whilst a similarly impactful work could result in a new low-power, eco-friendly microchip that finds its way into millions of electronic devices. As we discuss in Chapter 5, there are numerous data sources that can be used to evidence comprehensive impact including financial records, parliamentary proceedings and policy documents. Furthermore, given the findings of (Wooldridge and King, 2019) (discussed in section 4.7), it may also be possible to use Altmetric data as evidence of comprehensive impact. However, none of these sources provide complete coverage. Even scientific news, which we focus on heavily in this work due to its global ubiquity and ease of access, has a selection bias towards newsworthy scientific works. Models and techniques that combine multiple sources of evidence could provide more complete comparisons of the comprehensive impact of different types of scientific works.

10.2.3 Re-visiting the Dynamics Between REF Impact Score and News

In section 5.4 we show that REF Impact case studies that are associated with one or more news articles are more likely to have a higher impact score than those that are not associated with news articles. Our experiment was conducted on a relatively small data set due to the limited availability of news articles with known links to REF impact case studies found during our data mining process (section 5.3.5). Future work could incorporate additional data from the more recent REF 2021 assessment in order to provide more robust

conclusions. Linguistic analyses could also be carried out on scientific works referenced in REF impact case studies with and without associated news articles to see if there are any qualitative differences between the two sets. For example, are the scientific works in the set *with* associated news articles simply more newsworthy? Furthermore, future investigations could examine other possible confounding variables. For example, could the Matthew Effect (Merton, 1968), in which highly cited authors find it easier to accrue new citations by virtue of already being well known, also apply to academics who have existing relationships with journalists and news outlets?

In section 7.3 we observe that pairs of sentences originating from news and science articles that are associated with a REF impact case study are, on average, more semantically similar to each other than pairs of sentences without associated REF impact case studies. As above, future work could also focus on verifying this observation using a bigger dataset that incorporates news articles, scientific papers and impact case studies from REF 2021. Future work could also explore how similarity may relate to newsworthiness of a scientific statement. We intuit that a strong similarity between sentences from news articles and scientific papers may that the scientific work is already close to what a journalist is happy to report and requires less paraphrasing to become newsworthy. Such work should also investigate the differences in semantic similarity between pairs of sentences based on the CoreSC designation of the sentence from the scientific paper. We note in section 7.3 that pairs with an Outcome sentence generally have the weakest semantic similarity, likely due to assertive and simplistic reporting of results in news articles and more careful discussion of results in scientific papers. Future work may also benefit by using a CD^2CR-based sentence alignment technique to find topically aligned pairs of sentences rather than working with mean similarities as in our work.

10.2.4 Fully Automated Matching of News/Science Article Pairs using Content Features

In Chapter 5 we propose a semi-automated pipeline for linking news articles to scientific papers. In the event that a news article does not contain a link or DOI for the scientific work that it discusses, we use named-entity-recognition to possible candidate names for academics and institutions, and we search for candidate scientific papers in an external citation network. Human annotators are then required to review candidate papers and confirm links between documents where appropriate. Assuming it is possible to retrieve the full text content for candidate papers, future work could use techniques developed later in our thesis, specifically retrieval of scientific papers as proposed in chapter 6 and resolution of co-referring named entities via CD^2CR as proposed in 8 to automate the process of finding links between related documents. A decision threshold could be set based on similarity score or a classifier model could be trained to decide when to create links. A quantitative evaluation of the new automated pipeline could be carried out using the existing tooling

created in this thesis.

10.2.5 Modelling and Comparison of Long Sequences

In Chapter 6, we explore a number of encoding techniques for full news articles and scientific papers. Although we found that TF-IDF bag-of-words document representations were able to sufficiently encode similarity between news articles and the scientific papers that they discuss, none of the strategies that we investigated sufficiently explain how information flows between pairs of related documents. Consequently, we chose to align and model impact in news articles and scientific papers at *sentence level*. However, as we observed in Chapter 9, document-level and paragraph-level context can be an important indicator of comprehensive impact (e.g. motivation vs findings in a scientific paper).

Future work could investigate whether recent language models, such as Longformer (Beltagy et al., 2020), which have been designed to generate contextualised encodings for long documents could be used to encode and compare full news articles and scientific papers. Document representations could be examined to understand the extent to which they characterise information transfer between related documents. Ginzburg et al. (2021) introduce a RoBERTa-based (Liu et al., 2019a) model that conditionally applies attention to tokens long documents with respect to an input query in order to carry out information retrieval tasks. Future work could build on this work to conditionally encode scientific papers dependent on prominent sentences for news articles or even full news articles in order to provide better linking of related documents and facilitate SIS scoring.

Finally, future work could investigate whether rationale models such as UNIREX (Chan et al., 2022), which seek to generate human-readable rationales for decisions made by language models, could be used to explicitly characterise the linguistic prompts in news articles and scientific papers that signify comprehensive impact.

10.2.6 Relating Academic and Comprehensive Impact and Credit Attribution

This thesis has primarily focused on comprehensive impact and the development of natural language processing tools for its detection and measurement. However, we refrain from discussing or exploring how comprehensive impact and academic impact relate to each other.

In practice, new scientific work is made possible by "standing on the shoulders of giants"; scientists read and are inspired by the work of other scientists whose work they then extend and cite. Many works that generate comprehensive impact are made possible by theoretical frameworks and intermediate discoveries that paved the way for new practical applications and discoveries to occur. These relationships can be direct (e.g. we built a new model for detecting cancer based on a novel neural network architecture) or indirect

(we were able to discover a new deep sea fish because of a new industrial manufacturing process that allowed us to build a glass dome for our submarine that can withstand the pressure in the Mariana Trench) in nature.

We propose that a fair assessment of comprehensive impact should allow credit generated by new papers to also be shared with influencing works. A credit attribution system could improve engagement and collaboration between theoretical and applied branches of science, allowing the former to understand the real world impact that their work has and motivate authors whose works may otherwise become "sleeping beauties" (van Raan, 2004).

Ghosal et al. (2022)'s recent study aims to automatically discover research lineage for scientific work via citation classification models (Pride and Knoth, 2017; Valenzuela et al., 2015b). Future work could extend these methods, using SIS to attribute comprehensive impact to a novel study and propagating residual impact backwards through the citation network depending on how influential each preceding work is determined to be on its successor.

A

CD^{2}CR Annotation Guidelines

CD^{2}CR:News Articles and Scientific Paper Annotation Guidelines

By James Ravenscroft

Version 1.0 PUBLISHED

18/4/2020

Introduction

The purpose of this task is to identify examples of words and phrases that refer to the same object or concept across two related documents, even though they may use different language and style.

For example, a newspaper report about a new scientific work may say “Red Meat may contain cancer causing chemicals” and the related scientific work may say “an analysis of carcinogenicity of red meat”. The task here is to understand that “red meat” refers to the same thing in both sentences and possibly less obviously that “carcinogenicity” and “cancer causing” are the same thing in this context, forgiving any differences in tense and grammar.

As in the example above, this task is generally fairly intuitive with occasional, more difficult scenarios. This document aims to cover general ground guidelines for how you should approach the task and gives some examples of more difficult scenarios and how they should be resolved.

Core Principles

An overview of how the task works

You will be presented with a summary of a newspaper article and the abstract from a related scientific paper. You will also be presented with two sets of words or keywords that represent concepts in each document. By reading the two short documents you should use your human judgement to indicate whether the two sets of keywords refer to the same concept across the two documents or not by clicking the “Yes” or “No” buttons respectively. (see **Figure 1** below for an example of how this will be presented to you during the task). If a task is particularly challenging because the link between the sets of words is not simple then you can click “This task is difficult to think about”.

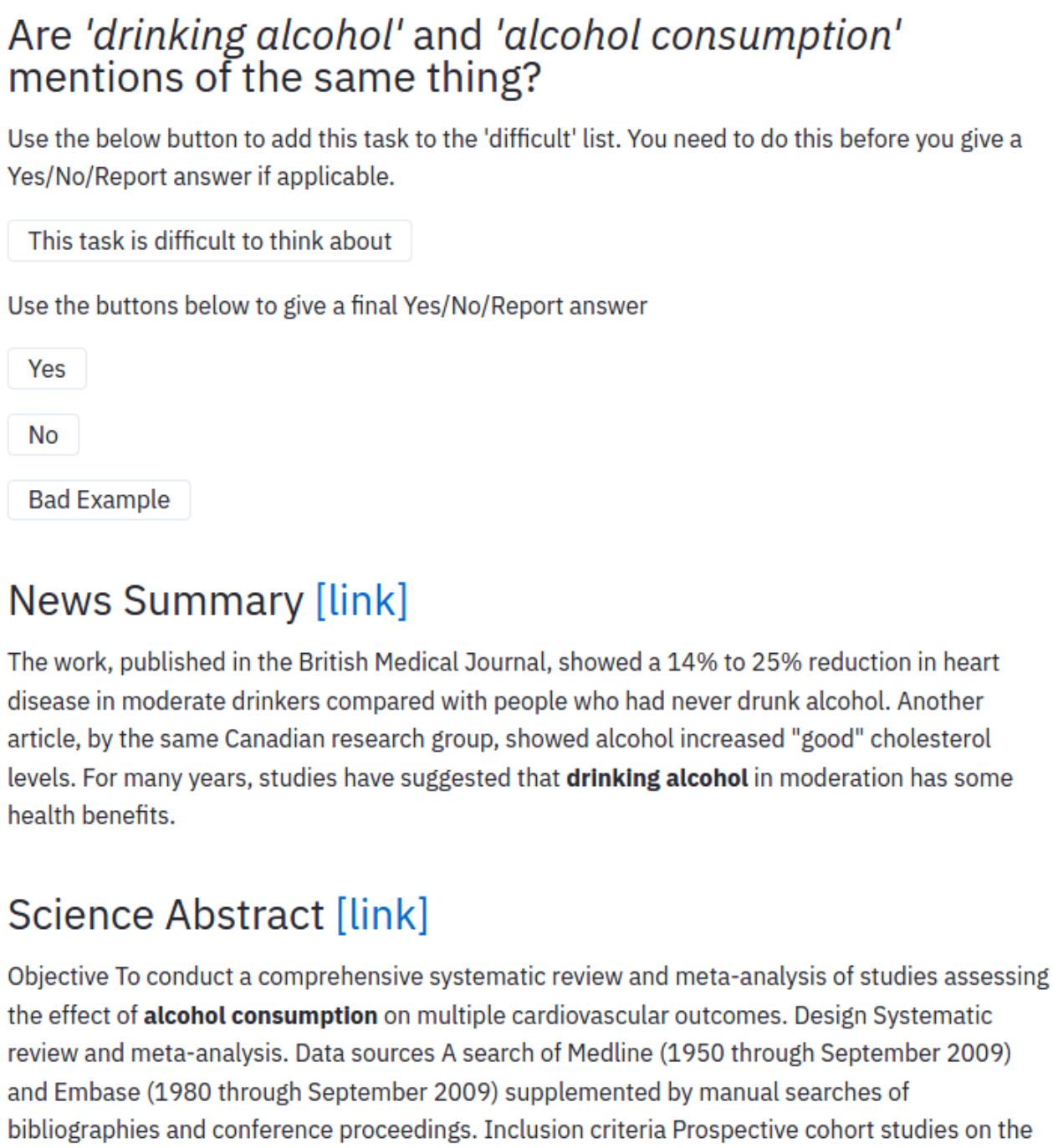


Figure 1: Screenshot of an example within the CDCR Tool.

There may be cases where the two mentions are the same but have different wording but there are also going to be cases where they have identical wording like in **Figure 2** below.

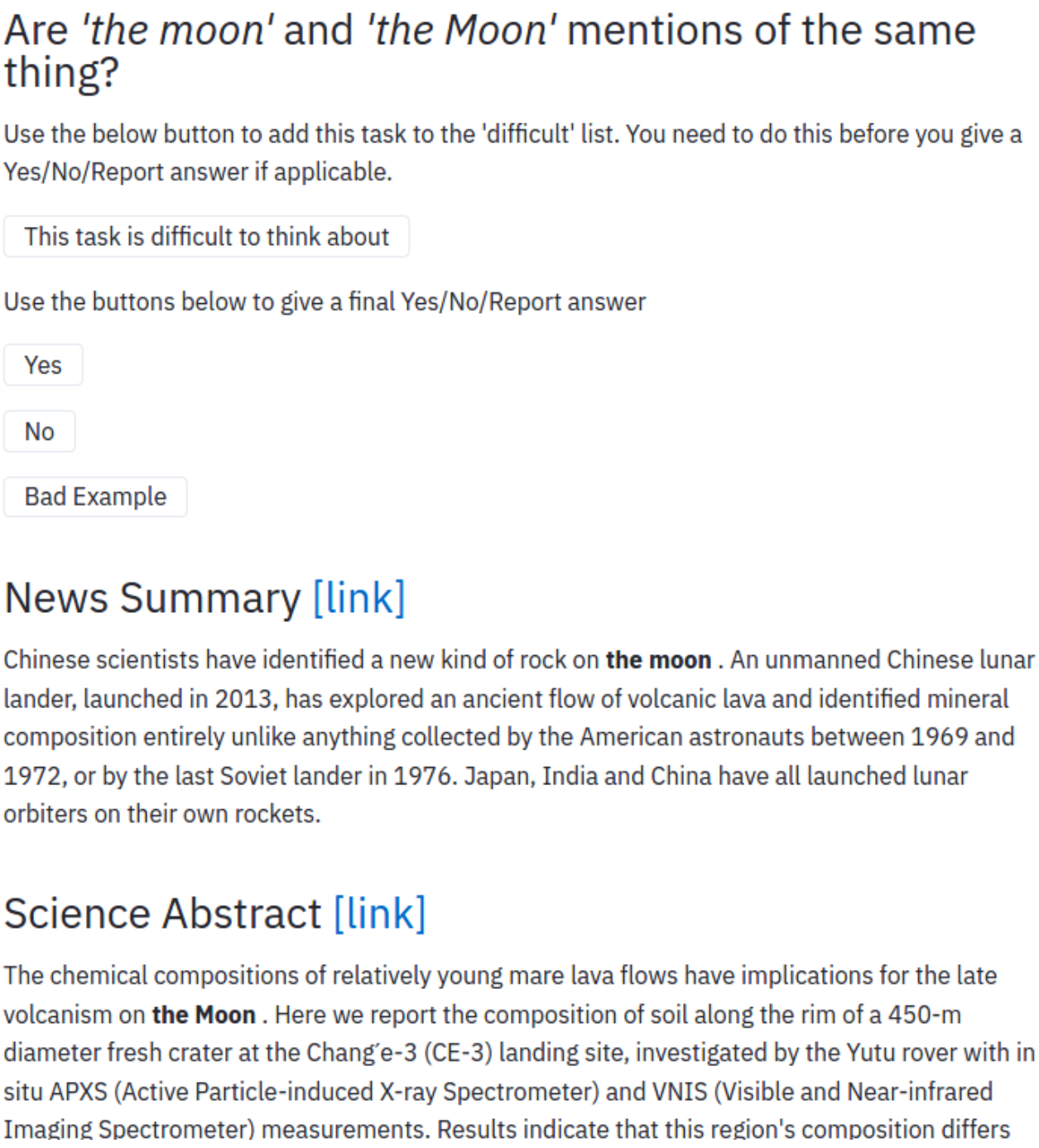


Figure 2: Screenshot of an example where the two mentions are identical

How long does the task go on for?

The task will continue until the system runs out of examples to present to the user. You can take a break at any point because the system will remember where you got to and take you back to the most recent example that you haven't yet annotated.

How the examples are generated and why it's important

The examples are generated automatically using an algorithm that discovers related pairs of news articles and scientific papers and then compares all concepts within the two documents against each other (every 'key phrase' is compared with every other 'key phrase' from the other document). We use an algorithm to provide a crude 'similarity' between pairs of key phrases allowing us to sort them from most similar to least similar in our annotation tool.

This means that you will see pairs that are most similar - possibly exactly the same word or phrase - to begin with and they will become more challenging or less likely to be the 'same' as you work through the examples.

Of course sometimes our automatic system gets things wrong. You might occasionally find that the newspaper article and scientific paper pairing are completely unrelated. If that is the case then simply click the "Report Bad Example" button to remove that example from the system. There are other scenarios when a task is a "bad example". Please refer to the section below for further guidance.

Resolving Complicated Examples

Please use the following instructions to resolve complicated examples.

Good example or bad example?

The following instructions should help you to decide whether to use the "Report bad example" button.

The key phrases are completely unrelated to each other

- Read the two summaries. If they are about the same topic but the key phrases are unrelated then this is simply a 'No' example.
- If the newspaper article and scientific paper summaries appear to be completely unrelated then see "Newspaper and Scientific Paper aren't about the same thing" below.

Newspaper and scientific paper aren’t about the same thing

- If the newspaper article and scientific paper appear to be unrelated, please open the links and read the full news article - it might be that the scientific paper was mentioned in passing - in which case the answer might be a “no” rather than a “bad example”.
- An example is a bad example if the newspaper article and scientific paper are clearly unrelated to each other after closer examination.

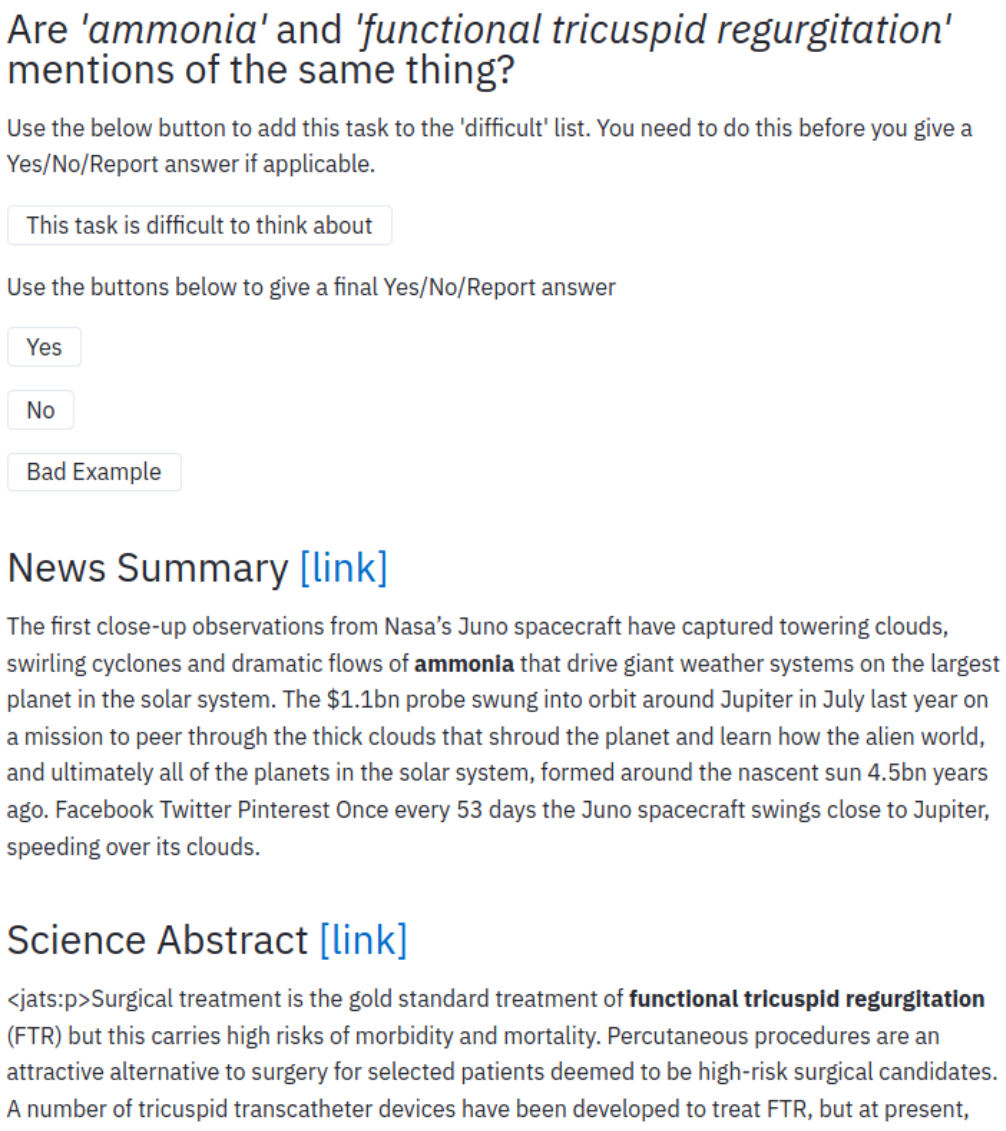


Figure 3: Unrelated scientific paper and newspaper articles - the news article is about discoveries made by the Juno spacecraft, the scientific paper is about a specific surgical treatment.

One or both of the **key phrases** is gobbledygook/unreadable

- If one of the key phrases is missing a few characters but you can still understand what it means from context then please treat the example as "good" and answer with yes or no.
- If the key phrase is completely unreadable even with context from the summary documents then mark the example as "bad"
- Please see "Formatting Issues" below for further guidance

One or both of the **summary documents** is gobbledygook/unreadable

- If either summary document is missing a few words at the beginning or end but you can still understand the general context then please treat the example as “good” and answer with yes or no.
- If there are words missing from the summary, please use the links to the full documents to help you to understand the summary. If this helps then please treat the example as “good” and answer with yes or no.
- If after further examination you cannot understand one or both of the summaries then mark the example as “Bad Example”
- Please see “Formatting Issues” below for further guidance

Formatting issues

Nonsensical word placement

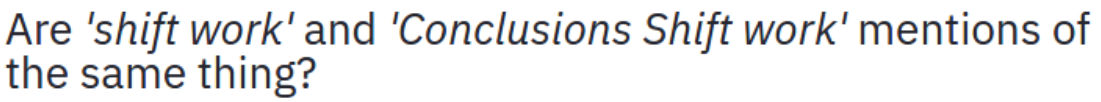


Figure 3: An example where bad formatting has seeped into the entity text (Conclusions is the heading/title of the section that the mention was pulled from)

Sometimes the automatic process that creates the examples will make a mistake and the formatting of the task may be confusing or incorrect. **Figure 3** shows an example where the section heading ‘Conclusions’ has been included in the mention from the scientific work.

Since these issues are a side effect of the way that the text has been processed and are unrelated to whether or not the two mentions refer to the same thing, you should mark these examples as **Bad Example**

Mangled/incorrect numbers

As part of the automated process that creates the tasks, numbers are sometimes mangled or broken. In these cases we are still interested in whether or not the mentions refer to the same concept regardless of the numbers.

In **Figure 4** below the news mention is ‘000 runners’ and the scientific mention is ‘Participants 3913 of 7048 participants in the Bonn marathon 2010’ - by reading the two summaries it is possible to deduce that both phrases refer to the same. Therefore this example results in a “Yes” answer.

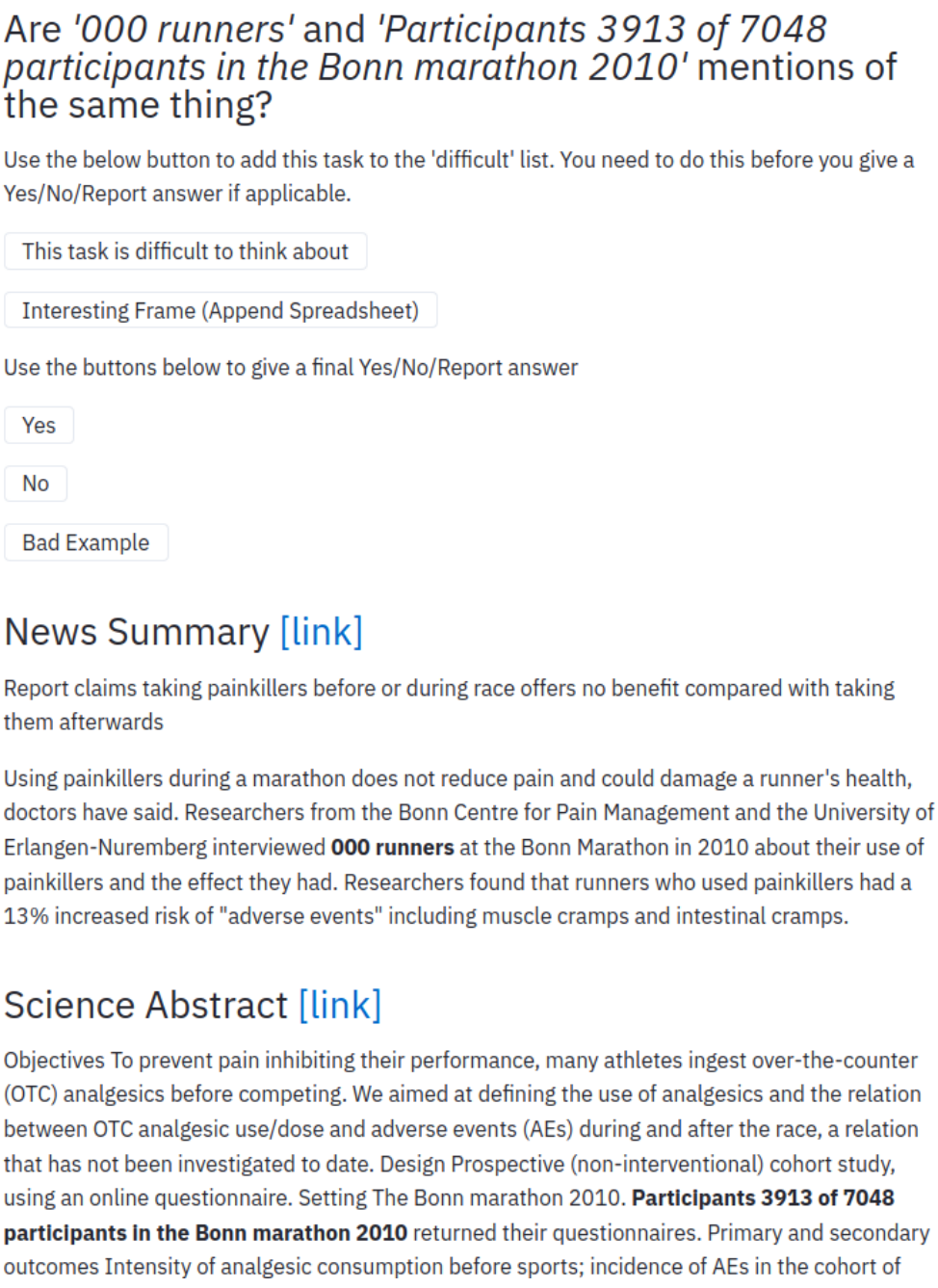


Figure 4: The number of runners in the news mention is mangled but it is clear that the author is referring to the 3913 runners who participated in the study.

Complicated cases of ‘yes’ or ‘no’

The following instructions should help you to resolve cases where the example is more difficult.

You can click the “This example is difficult to think about” button to alert the research team to particularly challenging examples. If another user has already reported an example then a message to this effect will appear instead of the button.

One of the key phrases is less specific than the other

- Generally speaking journalists use less precise wording to appear to a more general/less specialist audience.

- Please check whether, to the best of your judgement, the news article is referring to the same concept as the scientific work despite linguistic differences. Use the summary context or full document link to help in this judgement. Below are some examples to help:

    a. A news article states that “Sheep able to distinguish pictures of celebrities from unfamiliar faces with near-human accuracy… the connection emerges from work on face recognition skills of a welsh mountain breed...”

    AND

    The related scientific paper says “we trained eight sheep (ovis aries) to recognize the faces of four celebrities…”

    The key phrases are “a welsh mountain breed” and “eight sheep (ovis aries)”.

    In this case we can infer from context that both “a welsh mountain breed” and “eight sheep (ovis aries)” refer to the same set of sheep that were trained on the facial recognition task so this example would result in a **YES** answer.

    b. A news article says “the humble shark, in particular, the great white shark is perhaps the most feared…”,

    AND

    A related scientific article says “great white sharks (carcharodon carcharias) less likely to attack humans…”

    The key phrases are “the humble shark” and “great white sharks (Carcharodon carcharias)”.

    In this case we can infer that the journalist was aware of the difference between sharks in general and great white sharks and that “the humble shark” is in fact different to “great white sharks (Carcharodon carcharias)” resulting in a **NO** answer.

‘Technical’ differences between the mentions

Please bear in mind that it is the intent of the journalist that is important rather than any specific technical differences between the mentions. Please keep an open mind when you compare example mentions and give the benefit of the doubt to journalists using less precise language or modifiers.

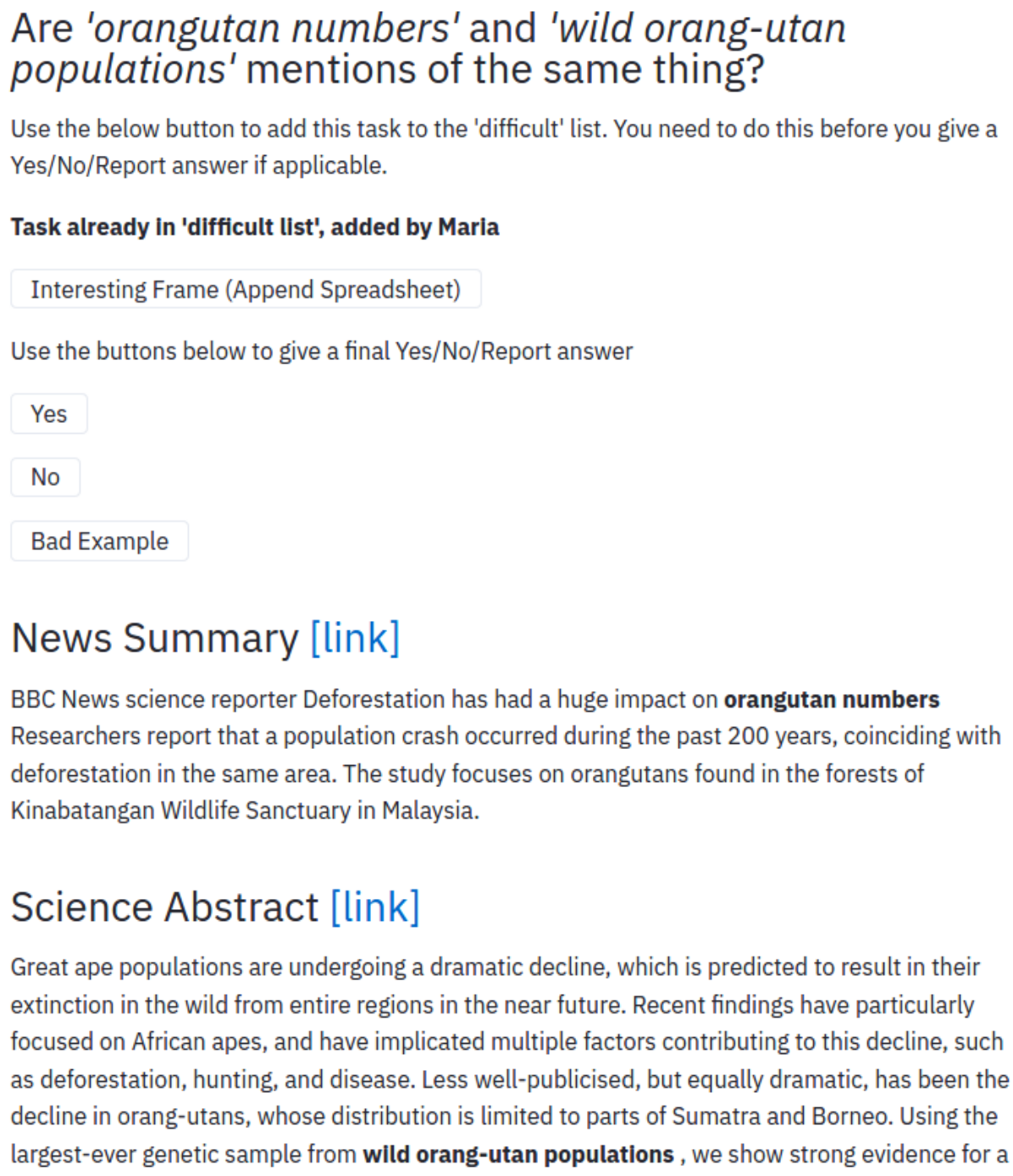


Figure 5: In this example we compare orangutan numbers and wild orangutan populations. The mentions have different geographical contexts (Malaysia and Sumatra and Borneo respectively) but could be considered the same.

In **Figure 5** above we compare “orangutan numbers” and “wild orangutan populations”. The news summary seems to discuss a wildlife sanctuary in Malaysia whereas the scientific work centres on wild orangutan populations in Sumatra and Borneo. It is clear from reading the full news article that the journalist is talking specifically about the study despite the geographical distinction.

Given that our primary focus is journalistic intent and the journalist is adamantly talking about orangutan populations discussed in the study, we would resolve this example to be a “yes”.

By reading the full scientific paper (not required as part of the annotation process) it is possible to understand the full context: the orangutans originated in Sumatra and Borneo but were relocated to a malaysian wildlife sanctuary. This explains the inconsistency between the journalist and scientific work and confirms that “yes” was the correct answer.

Bibliography

Geneva Abdul. 2022. R2-D-Chew: robot chef imitates human eating process to create tastier food. *The Guardian*.

Euan Adie and William Roe. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1):11–17.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Nikolaos Aletras and Mark Stevenson. 2014. Measuring the Similarity between Automatically Generated Topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 22–27, Gothenburg, Sweden. Association for Computational Linguistics.

Claudio Gil Soares de Araújo. 2014. Detailing the Writing of Scientific Manuscripts: 25-30 Paragraphs. *Arquivos Brasileiros de Cardiologia*, 102(2):e21–e23.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*, page 16.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. Writing Strategies for Science Communication: Data and Computational Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv*. ArXiv:1607.06450 [cs, stat].

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL '98/COLING '98*, page 79–85, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv*. Number: arXiv:1409.0473 arXiv:1409.0473 [cs, stat].

Sumit Kumar Banshal, Vivek Kumar Singh, Pranab K. Muhuri, and Philipp Mayr. 2019. Disciplinary Variations in Altmetric Coverage of Scholarly Articles.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189.

Christoph Bartneck and Servaas Kokkelmans. 2011. Detecting h-index manipulation through self-citation analysis. *Scientometrics*, 87(1):85–98.

Marco T. Bastos. 2016. Digital Journalism And Tabloid Journalism. In Bob Franklin and Scott A. Eldridge, editors, *The Routledge Companion to Digital Journalism Studies*, 1 edition, pages 217–225. Routledge, London ; New York : Routledge, 2017.

Martin Becker and Jakob Egetenmeyer. 2018. A prominence-based account of temporal discourse structure. *Lingua*, 214:28 – 58.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. Number: arXiv:2004.05150 arXiv:2004.05150 [cs].

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. volume 5, pages 157–166. Conference Name: IEEE Transactions on Neural Networks.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M Blei. 2012. Introduction to Probabilistic Topic Models. page 16.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Branimir Boguraev and Christopher Kennedy. 1997. Salience-based content characterisation of documents. In *Proceedings of the ACL/EACL Workshop in Intelligent Scalable Text Summarization*, pages 3–9.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Filip Boltužić and Jan Šnajder. 2015. Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 [cs].

Lutz Bornmann. 2014. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4):895—-903.

Lutz Bornmann and Robin Haunschild. 2016. To what extent does the Leiden manifesto also apply to altmetrics? A discussion of the manifesto against the background of research into altmetrics. *Online Information Review*, 40(4):529–543.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23329.

Léon Bottou. 2012. Stochastic Gradient Descent Tricks. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the*

Trade, volume 7700, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Stevo Bozinovski. 2020. Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. *Informatica*, 44(3).

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2021. Sequential Sentence Classification in Research Papers using Cross-Domain Multi-Task Learning. *arXiv:2102.06008 [cs]*. ArXiv: 2102.06008.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

T Bubela et al. 2009. Science communication reconsidered. *Nature Biotechnology*, 27(6):514–518.

D Butler. 2012. Scientists: your number is up. *Nature*, 485(7400):564–564.

Paul Cairney. 2021. The UK government's COVID-19 policy: assessing evidence-informed policy analysis in real time. *British Politics*, 16(1):90–116.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling.

Arie Cattan and Sophie Johnson. 2021. SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. In *Proceedings of Automated Knowledge Base Construction (AKBC)*.

Timothy Caulfield and Ubaka Ogbogu. 2015. The commercialization of university-based research: Balancing risks and benefits. *BMC Medical Ethics*, 16(1):70.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. UNIREX: A Unified Learning Framework for Language Model Rationale Extraction. *undefined*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. Number: arXiv:1312.3005 arXiv:1312.3005 [cs].

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Number: arXiv:1406.1078 arXiv:1406.1078 [cs, stat].

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. page 8.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (almost) from Scratch. ArXiv: arXiv:1103.0398.

Alexis Conneau, Guillaume Lample, Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

P Conrad. 1999. Uses of expertise: sources, quotes, and voice in the reporting of genetics in the news. *Public Understanding of Science*, 8:285–302.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. 2015. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019.

I G Councill et al. 2008. ParsCit: An open-source CRF Reference String Parsing Package. *LREC '08: Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008(3):661–667.

Steven P Crain, Shuang-Hong Yang, Hongyuan Zha, and Yu Jiao. 2010. Dialect Topic Modeling for Improved Consumer Medical Search. page 5.

Michael Crawshaw. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. ArXiv:2009.09796 [cs, stat].

I. Csiszar. 1975. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158. Publisher: Institute of Mathematical Statistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Lecture Notes in Computer Science, pages 177–190, Berlin, Heidelberg. Springer.

Ralph B D'Agostino. 1971. An Omnibus Test of Normality for Moderate and Large Size Samples. *Biometrika*, 58:341–348.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Harris Drucker, Christopher J C Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1996. Support Vector Regression Machines. *Advances in Neural Information Processing Systems*, 9:7.

Daniel Duma, Maria Liakata, Amanda Clare, James Ravenscroft, and Ewan Klein. 2016. Rhetorical Classification of Anchor Text for Citation Recommendation. *D-Lib Magazine*, 22(9/10).

Sourav Dutta and Gerhard Weikum. 2015. Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment. *Transactions of the Association for Computational Linguistics*, 3:15–28.

Marc A. Edwards and Siddhartha Roy. 2016. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, page ees.2016.0223.

Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

European Association of Science Editors. 2007. Ease statement on inappropriate use of impact factors.

European Commission. Joint Research Centre. 2017. *Final report on the collection of patents and business indicators by economic sector: societal grand challenges and Key Enabling Technologies : collection and analysis of private R&D investment and patent data in different sectors, thematic areas and societal challenges*. Publications Office, LU.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2020. A Brief Review of Domain Adaptation. ArXiv:2010.03978 [cs].

R. A. Fisher. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Rebecca Fuoco. 2021. How to get media coverage and boost your science's impact. *Nature*.

Philip Gage. 1994. A New Algorithm for Data Compression. page 14.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. *arXiv:2012.15723 [cs]*. ArXiv: 2012.15723.

Eugene Garfield. 2006. History and Meaning of the Journal Impact Factor. *JAMA*, 295(1):90–93.

Linda Geddes. 2022. Scientists create tomatoes genetically edited to bolster vitamin D levels. *The Guardian*.

Damien P. George and Robert Knegjens. 2014. Paperscape. `http://paperscape.org`.

Vicente P. Gerrero-Bote, Rodrigo Sánchez-Jiménez, and Félix De-Moya-Anegón. 2019. The citation from patents to scientific output revisited: a new approach to the matching Patstat / Scopus. *El Profesional de la Información*, 28(4).

Tirthankar Ghosal, Piyush Tiwary, Robert Patton, and Christopher Stahl. 2022. Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies*, 2(4):1511–1528.

Danilo Giampiccolo, Bernardo Magnini, Elena Cabrio, Hoa Trang Dang, Ido Dagan, and Bill Dolan. The Fourth PASCAL Recognizing Textual Entailment Challenge. page 11.

Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-Supervised Document Similarity Ranking via Contextualized Language Models and Hierarchical Inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098. ArXiv:2106.01186 [cs].

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. page 9.

Yoav Goldberg. 2015. A Primer on Neural Network Models for Natural Language Processing. Number: arXiv:1510.00726 arXiv:1510.00726 [cs].

Jack Goodwin and Eugene Garfield. 1980. Citation Indexing-Its Theory and Application in Science, Technology, and Humanities. *Technology and Culture*, 21(4):714.

Andrew Gregory. 2022. Robot-assisted surgery can cut blood clot risk and speed recovery, study finds. *The Guardian*.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6019–6024.

Laurel L. Haak, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. Orcid: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated Identification of Media Bias by Word Choice and Labeling in News Articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 196–205, Champaign, IL, USA. IEEE.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Abdalraouf Hassan and Ausif Mahmood. 2018. Convolutional Recurrent Deep Learning Model for Sentence Classification. *IEEE Access*, 6:13949–13957. Conference Name: IEEE Access.

Stefanie Haustein, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. 2016. Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1):232–238.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [cs].

HEFCE. 2012. Assessment framework and guidance on submissions.

HEFCE. 2017. Initial decisions on the Research Excellence Framework 2021. Technical report, HEFCE.

HEFCE. 2019. Guidance on submissions (2019/01) - REF 2021.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying Discourse Coherence Theory to Cross Document Coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David J. Hess and Professor David J. Hess. 1997. *Science Studies: An Advanced Introduction*. NYU Press. Google-Books-ID: QAwDUOdHE1kC.

Tim C Hesterberg. 2015. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *American Statistician*, 69(4):371–386.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. The leiden manifesto for research metrics. *Nature*, 520(7548):429.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. rmsprop: Divide the gradient by a running average of its recent magnitude.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. Number: arXiv:1503.02531 arXiv:1503.02531 [cs, stat].

Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572.

Jorge E Hirsch. 2020. Superconductivity, What the H? The Emperor Has No Clothes. 49:6.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Darrin Hodgetts, Kerry Chamberlain, Margaret Scammell, Rolinda Karapu, and Linda Waimarie Nikora. 2008. Constructing health news: possibilities for a civic-oriented journalism. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, 12(1):43–66.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. Number: arXiv:1902.00751 arXiv:1902.00751 [cs, stat].

J Huang and S Ertekin. 2006. Fast Author Name Disambiguation in CiteSeer.

Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2014. Refseer: A citation recommendation system. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 371–374.

John Hutchins. 1997. From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. A Chronology. *Machine Translation*, 12(3):195–252.

W. John Hutchins. 2004. The Georgetown-IBM Experiment Demonstrated in January 1954. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Robert E. Frederking, and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research*, volume 3265, pages 102–114. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proceedings of the 2010 Text Analysis Conference*.

Jyun-Yu Jiang, Chenyan Xiong, Chia-Jung Lee, and Wei Wang. 2020. Long Document Ranking with Query-Directed Sparse Transformer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4594–4605, Online. Association for Computational Linguistics.

JISC and the Internet Archive. 2013. JISC UK Web Domain Dataset (1996-2013).

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In Jaime G. Carbonell, Jörg Siekmann, G. Goos, J. Hartmanis, J. van Leeuwen, Claire Nédellec, and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398, pages 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed edition. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J. OCLC: 213375806.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. page 9.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 Task 3: Cross-Level Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation*, 50(1):5–33.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Bradley Karas, Sue Qu, Yanji Xu, and Qian Zhu. 2022. Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Frontiers in Artificial Intelligence*, 5.

Amirhossein Kazemnejad. 2019. Transformer architecture: The positional encoding. *kazemnejad.com*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [cs].

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. Technical report, University of Oslo.

John D Lafferty and David M Blei. 2006. Correlated Topic Models. page 8.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv:1909.11942 [cs].

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159—174.

J Lane and S Bertuzzi. 2010. The STAR METRICS project: current and future uses for S&E workforce data. In *Science of Science Measurement Workshop, held Washington DC*.

Mark A. Largent and Julia I. Lane. 2012. Star metrics and the science of science policy. *Review of Policy Research*, 29(3):431–438.

Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. Number: arXiv:1607.05368 arXiv:1607.05368 [cs].

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. *Computing Research Repository*, arXiv:1909.02339.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *Proceedings of the 31 st International Conference on Machine Learning*, page 9.

League of European Research Universities. 2015. Not everything that can be counted counts ...

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Ha. 1998a. Gradient-Based Learning Applied to Document Recognition. page 46.

Yann LeCun, Leon Bottou, Genevieve B. Orr, and Klaus Robert Müller. 1998b. Efficient BackProp. In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pages 9–50. Springer, Berlin, Heidelberg.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Lillian Lee. 1999. Measures of Distributional Similarity. page 8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Loet Leydesdorff and Staša Milojević. 2013. Scientometrics. ArXiv:1208.4566 [cs].

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8320–8331.

Maria Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4. Association for Computational Linguistics.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin R Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A Discourse-Driven Content Model for Summarising Scientific Articles Evaluated in a Complex Question Answering Task. In *EMNLP*, pages 747–757. Association for Computational Linguistics.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012a. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012b. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *LREC 2010, 7th International Conference on Language Resources and Evaluation*.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: 1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Annie Louis and Ani Nenkova. 2013a. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2013b. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

Yushuang Lyu, Muqi Yin, Fangjie Xi, and Xiaojun Hu. 2022. Progress and Knowledge Transfer from Science to Technology in the Research Frontier of CRISPR Based on the LDA Model. *Journal of Data and Information Science*, 7(1):1–19.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ansel MacLaughlin, John Wihbey, and David A. Smith. 2018. Predicting News Coverage of Scientific Articles. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, pages 191–200.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo Maria Ponti, and Anna Korhonen. 2021. Verb Knowledge Injection for Multilingual Event Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6952–6969, Online. Association for Computational Linguistics.

Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic Textual Similarity Methods, Tools, and Applications: A Survey. *Computación y Sistemas*, 20(4).

Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. 2009. Introduction to Information Retrieval. page 581.

Christopher D Manning and Hinrich Schiitze. 1998. Foundations of Statistical Natural Language Processing. page 704.

M Marelli, S Menini, M Baroni, L Bentivogli, R Bernardi, and R Zamparelli. 2014a. A SICK cure for the evaluation of compositional distributional semantic models. page 8.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014b. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8.

Frank J Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78.

Paul R. Mcallister, Francis Narin, and James G. Corrigan. 1983. Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management*, EM-30(4):205–211.

Kathy McKeown, Hal Daume, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid O'Seaghdha, Dragomir Radev, Clay Templeton, and Simone Teufel. 2016. Predicting the Impact of Scientific Concepts Using Full-Text Features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.

Lokman I. Meho. 2007. The rise and rise of citation analysis. *Physics World*, 20(1):32–36.

Antonio V. Menéndez Alarcón. 2012. Newspapers Coverage of Spain and the United States: A Comparative Analysis. *Sociology Mind*, 02(01):67–74.

K Merton, Robert. 1968. The Matthew Effect in Science. 159(3810):56–63.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. *Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, Barcelona, Spain*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1–9.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient Estimation of Word Representations in Vector Space. Number: arXiv:1301.3781 arXiv:1301.3781 [cs].

Tomáš Mikolov. 2012. *STATISTICAL LANGUAGE MODELS BASED ON NEURAL NETWORKS*. Ph.D. thesis.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, volume 2, page 880, Singapore. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. *arXiv:2104.08773 [cs]*. ArXiv: 2104.08773.

David Moher, Lex Bouter, Sabine Kleinert, Paul Glasziou, Mai Har Sham, Virginia Barbour, Anne-Marie Coriat, Nicole Foeger, and Ulrich Dirnagl. 2020. The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, 18(7):e3000737. Publisher: Public Library of Science.

Katarzyna Molek-Kozakowska. 2017. Stylistic analysis of headlines in science journalism: A case study of New Scientist. *Public Understanding of Science*, 26(8):894–907. Publisher: SAGE Publications Ltd.

Christopher E. Moody. 2016. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. Number: arXiv:1605.02019 arXiv:1605.02019 [cs].

E Moreau et al. 2008. Robust Similarity Measures for Named Entities Matching. *COLING*, pages 593–600.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2019. How we do things with words: Analyzing text as social and cultural data. ArXiv:1907.01468 [cs].

Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, Melbourne, Australia. Association for Computational Linguistics.

Mary Jo Nye. 2016. Speaking in Tongues. *Distillations*.

Korawit Orkphol and Wu Yang. 2019. Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet. *Future Internet*, 11(5):114. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

José Luis Ortega. 2018. Disciplinary differences of the impact of altmetric. *FEMS Microbiology Letters*, 365(7).

Bert Overlaet. 2022. A Pathway towards Multidimensional Academic Careers. Technical report.

Lawrence Page, Sergey Brin, Ravi H. Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University.

Harshit Parikh, Harsh Sawant, Bhautik Parmar, Rahul Shah, Santosh Chapaneri, and Deepak Jayaswal. 2020. Encoder-Decoder Architecture for Image Caption Generation. In *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*, pages 174–179.

Sarah Parks, Becky Ioppolo, Martin Stepanek, and Salil Gunashekar. 2018. *Guidance for standardising quantitative indicators of impact within REF case studies*. RAND Corporation.

N Paskin. 2015. The digital object identifier: From ad hoc to national to international. In *The Critical Component: Standards in the Information Exchange Environment*. ALCTS (American Library Association Publishing).

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, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jiaxin Pei and David Jurgens. 2021. Measuring Sentence-Level and Aspect-Level (Un)certainty in Science Communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014b. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proc. of NAACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 186, Honolulu, Hawaii. Association for Computational Linguistics.

Heather Piwowar. 2013. Altmetrics: Value all research products. *Nature*, 493(7431):159.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

David Pride and Petr Knoth. 2017. Incidental or influential? – A decade of using text-mining for citation function classification. Wuhan. Num Pages: 12.

David Pride and Petr Knoth. 2018. Peer Review and Citation Data in Predicting University Rankings, a Large-Scale Analysis. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, *Digital Libraries for Open Knowledge*, volume 11057, pages 195–207. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

J. Priem, D. Taraborelli, P. Groth, and C. Neylon. 2010. Altmetrics: A manifesto. http://altmetrics.org/manifesto.

Anthony F. J. van Raan. 2004. Sleeping Beauties in science. *Scientometrics*, 59(3):467–472.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jonathan Raiman and Olivier Raiman. 2018. DeepType: Multilingual entity linking by neural type system evolution. In *AAAI Conference on Artificial Intelligence*.

Delip Rao, Paul McNamee, and Mark Dredze. 2010. Streaming cross document entity coreference resolution. In *Coling 2010: Posters*, pages 1050–1058, Beijing, China. Coling 2010 Organizing Committee.

James Ravenscroft, Maria Liakata, and Amanda Clare. 2013. Partridge: An Effective System for the Automatic Cassification of the Types of Academic Papers. In *Research and Development in Intelligent Systems XXX*, pages 351–358. Springer.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

REF 2014. 2011. Assessment framework and guidance on submissions.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Mike Rossner, Heather Van Epps, and Emma Hill. 2007. Show me the data. *The Journal of Cell Biology*, 179(6):1091–1092.

Stuart J. Russell, Peter Norvig, and Ernest Davis. 2010. *Artificial intelligence: a modern approach*, 3rd ed edition. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River.

Jill Rutter, Edward Marshall, and Sam Sims. 2012. *The "S" Factors: Lessons from IFG's Policy Success Reunions*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv:2110.08207 [cs]*. ArXiv: 2110.08207.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. *arXiv:2001.07676 [cs]*. ArXiv: 2001.07676.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean Voice Search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

David Steingard, Marcello Balduccini, and Akanksha Sinha. 2022. Applying AI for social good: Aligning academic journal ratings with the United Nations Sustainable Development Goals (SDGs). *AI & SOCIETY*.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher Joseph Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations (ICLR)*.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A Simple Framework for Opinion Summarization. *arXiv:2005.01901 [cs]*. ArXiv: 2005.01901.

Hui-Yun Sung, Chun-Chieh Wang, Dar-Zen Chen, and Mu-Hsuan Huang. 2014. A comparative study of patent counts by the inventor country and the assignee country. *Scientometrics*, 100(2):577–593.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. page 9.

Martin Szomszor and Euan Adie. 2022. Overton – A bibliometric database of policy document citations. Number: arXiv:2201.07643 arXiv:2201.07643 [cs].

Ole Henning Sørensen, Jakob Bjørner, Andreas Holtermann, Johnny Dyreborg, Jorid Birkelund Sørli, Jesper Kristiansen, and Steffen Bohni Nielsen. 2022. Measuring societal impact of research—Developing and validating an impact instrument for occupational health and safety. *Research Evaluation*, 31(1):118–131.

Xiangru Tang, Alexander R. Fabbri, Ziming Mao, Griffin Adams, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries. *arXiv:2109.09195 [cs]*. ArXiv: 2109.09195.

Technical Committee ISO/TC 46. 2017. Information and documentation — WARC file format. Standard, International Organization for Standardization, Geneva, CH.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Simone Teufel. 2006. Argumentative Zoning for Improved Citation Indexing. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 159–169. Springer-Verlag, Berlin/Heidelberg.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Center for the Study of Language and Inf.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. Number: arXiv:2104.08663 arXiv:2104.08663 [cs].

Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. 2013. Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*.

Michael E Tipping and Christopher M Bishop. 1999. Mixtures of Probabilistic Principal Component Analysers. page 30.

Dana Topousis, Rick Weiss, and NIH Office of Communications. 2010. STAR METRICS: New Way to Measure the Impact of Federally Funded Research.

Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. Number: arXiv:1511.06388 arXiv:1511.06388 [cs].

U.S. Department of Commerce and National Institute of Standards and Technology. 2012. *Secure Hash Standard - SHS: Federal Information Processing Standards Publication 180-4*. CreateSpace Independent Publishing Platform, North Charleston, SC, USA.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015a. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 21–26.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015b. Identifying Meaningful Citations. page 6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52.

Henning Wachsmuth, Nona Naderi, Graeme Hirst, Benno Stein, Vinodkumar Prabhakaran, Yonatan Bilu, Yufang Hou, and Tim Alberdingk Thijm. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 176–187.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *Computing Research Repository*, arXiv:2004.14974.

Ludo Waltman. 2015. A review of the literature on citation impact indicators. *Clinica Terapeutica*, 168(1):14–22.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2021. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53(3):1–34.

Christian Wartena. 2013. Distributional Similarity of Words with Different Frequencies. *DIR*, page 4.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Jacob O. Wobbrock, Lara Hattatoglu, Anya K. Hsu, Marijn A. Burger, and Michael J. Magee. 2021. The Goldilocks zone: young adults' credibility perceptions of online news articles based on visual appearance. *New Review of Hypermedia and Multimedia*, 27(1-2):51–96. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13614568.2021.1889690.

Jenny Wooldridge and Mike B. King. 2019. Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology*, 70(3):271–282. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24122.

Dustin Wright and Isabelle Augenstein. 2021. Semi-Supervised Exaggeration Detection of Health Science Press Releases. *arXiv:2108.13493 [cs]*. ArXiv: 2108.13493.

Jian Wu, Kyle Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Tuarob, Alexander Ororbia, Douglas Jordan, Prasenjit Mitra, and C. Lee Giles. 2015. Citeseerx: Ai in a digital library search engine. *AI Mag.*, 36:35–48.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Number: arXiv:1906.08237 arXiv:1906.08237 [cs].

Xiaoyao Yin, Yangchen Huang, Bin Zhou, Aiping Li, Long Lan, and Yan Jia. 2019. Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access*, 7:169434–169445.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 545–550.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. Number: arXiv:2007.14062 arXiv:2007.14062 [cs, stat].

Yu Zhang and Qiang Yang. 2021. A Survey on Multi-Task Learning. ArXiv:1707.08114 [cs].

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Yujia Zhou, Zhicheng Dou, Huaying Yuan, and Zhengyi Ma. 2022. Socialformer: Social Network Inspired Long Document Modeling for Document Ranking. Number: arXiv:2202.10870 arXiv:2202.10870 [cs].