Electronics and Computer Science

Faculty of Engineering and Physical Sciences

University of Southampton

Nantheesan Raveenthiran

April 29th 2025

Using Machine Learning Techniques For Early Identification Of Paper Retraction

Project supervisor: **Dr Markus Brede**

Project second supervisor: **Dr Mike Wald**

A project report submitted for the award of

BSc Computer Science

Acknowledgments

A massive thank you, to Dr Markus Brede for providing me with support, guidance and wisdom throughout this project.
Thank you, Dr Mike Wald for your help as my second supervisor

Abstract

With the number of retractions per year increasing rapidly and the time-to-retract period being too long to stop the broader impact of a fraudulent paper's retraction and subsequently failing to mitigate its damage to the research field, the current retraction mechanisms must be revised. In this research paper, we leverage scientometrics and machine learning to try and build models to investigate if we could use these techniques to detect if a paper is going to be retracted. We investigated using traditional metadata, citation metadata, citation networks, and a novel network called "Meta-Net", which is a network constructed from metadata to try and capture semantic meaning in network form. "Meta-Net" was inspired by the citation network's positive results. We found out that there were some problematic metrics in our dataset. Still, excluding them, the techniques did harbor meaningful information that we used to train different models with high predictive power. Our novel approach harbored a small predictive signal, but we concluded that much research could be done on this approach to possibly obtain better, more powerful results. Our results for metadata, citation metadata, and citation network were very positive, resulting in powerful models. Our research concluded that utilizing machine learning for paper retraction is very viable, with richer, more complete data. We could aggregate our independent positive models into one possibly stronger model, which can be used for early detection.

Statement of Originality
- I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.
- I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.
- I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I have acknowledged all sources, and identified any content taken from elsewhere.
I used open source databases to gather data and have acknowledged all sources:

The sources I used were:

- https://pubmed.ncbi.nlm.nih.gov/ -- PubMed Database
- https://opencitations.net/ --- OpenCitation Database
- https://retractionwatch.com/ -- Retraction Watch Database

The libraries I used were:

- Pandas -- https://pandas.pydata.org/
- Numpy -- https://numpy.org/
- SciKit -- https://scikit-learn.org/stable/
- Nltk --- https://www.nltk.org/
- Gensim -- https://pypi.org/project/gensim/
- Tqdm -- https://tqdm.github.io/
- Sbert -- https://sbert.net/

I did all the work myself, or with my allocated group, and have not helped anyone else.
The material in the report is genuine, and I have included all my data/code/designs.
I have not submitted any part of this work for another assessment.
My work did not involve human participants, their cells or data, or animals.

ECS Statement of Originality Template, updated August 2018, Alex Weddell aiofficer@ecs.soton.ac.uk

Chapter 1

Introduction

1.1 Problem

The foundation of scientific trust is built on the integrity of research published by the scientific community. However, this integrity is on the brink of collapse and subsequently breaking the trust of science; with the rise of fraudulent research coupled with the long time-to-retract period, the cascading effect of this fraudulent research is plaguing the scientific body with papers citing fraudulent research before their retraction. This paper aims to research the use of scientometrics and machine learning to allow for early detection of potentially retractable papers, reducing the ”time-to-retract” [1] period.

1.2 Goal

This project aims to explore the possibility of utilizing machine learning techniques, data features, citation network features and a novel network called ”Meta-Net” to find out if we can build a model that can accurately predict the likelihood of a research paper being retracted. Thus allowing for the early detection of possible retractable papers, decreasing the mean ”time-to-retract”[1].

1.3 Scope

This project's scope would be developing a model predicting the likelihood of a research paper being retracted. The machine learning techniques utilized in this research project would be constrained to only supervised learning techniques and exclude other machine learning techniques. In addition, we will explore the relationships and patterns of citation and MetaNet networks and use these metrics as data features. The dataset is restricted to the available information on retracted papers on the significant publication databases at the time of this project; there may be papers that are yet to be retracted but are fraudulent [1].

Chapter 2

Literature Review

This chapter will include research on the retractions in research, network analyses and machine learning research in Scientometrics

2.1 Review

The number of research papers being published is increasing rapidly, with 2016 about 1.92 million papers being indexed by Scopus and Web of Science publication databases and that number growing to about 2.82 million papers in 2022 [2]. A new worry arises: science itself is a constantly evolving multi-scale network [3], with a large number of papers being published each year, there is a growing concern that some of these papers are based on fraudulent foundations by citing fraudulent papers that have not yet been retracted, with the number of retractions per year rapidly increasing [1] and in 2023 the number of papers being retracted being around 10000 [4] the concern is not without a strong base.

The problem we are tackling with this research paper is trying to decrease the time taken for the retraction of fraudulent research papers so that we can mitigate their damage to the research world. The number of research papers that have been retracted recently has increased drastically, and the mean time to retract a fraudulent paper is 33.81 months (±35.63 SD), [1]. This means that in the time-to-retract period, propagation of errors could occur where subsequent research could be built on fraudulent findings, causing large amounts of effort to be wasted on resolving the damage caused by fraudulent publications. Scrubbing the literature to remove their influence is very lengthy [1], and speeding up the retraction process will decrease the influence a fraudulent paper has on the research space. Using this model as an initial detection mechanism would decrease the sample space of papers that need to be investigated for fraud or other errors, decreasing the mean time to retract and helping preserve academic integrity.

2.1.1 Causes of retractions

The primary cause of retraction can be split into two main sub-topics, intentional misconduct, and honest errors; research also shows that the proportion of intentional misconduct is about $\approx 60\%$ and honest errors is $\approx 40\%$ [5] - [6]. Whether the paper is retracted because of intentional misconduct or honest errors, the specific reason for retraction is only a tiny subset of issues, where intentional misconduct primarily entails two main reasons: authorship fraud, research misconduct [7] and honest errors were mainly due to oversight and negligence by the authors [8]. Due to the actual reason for retraction only being a small subset of reasons, there will be hidden characteristics that we can leverage when training our model.

Paper mills produce and sell manuscripts that resemble genuine research [9]. Research conducted by Brundy [10] shows that investigation into one such paper mill led up to 11,300 papers being retracted and that 400,000 papers show textual similarity to that of papers known to be published by paper mills [11]. With new paper mills being discovered rapidly, the number of papers published by paper mills may be much larger than these figures, with researchers discussing how these paper mills may also threaten trust in science. Papers published by these ”paper mills” usually have very similar characteristics, including reusing text, falsified images, and other similarities. They also include abnormal citation patterns where paper mill articles cite each other. Additionally, paper mills offer services like increasing citation count, which would create unique citation patterns. [12]. Our research aims to exploit these abnormal citation patterns as our model is going to be trained using citation patterns, which should yield promising results.

2.1.2 Impact of Retractions

Additionally, the impact of retractions was revealed by Lu's research [13], which shows the extent of the propagation of reputational damage caused when a paper is retracted. It stated that ”a single retraction triggers citation losses through an author's prior body of work” , where citations fall by an average of 6.9% compared to a control author. In addition to reputational damage caused to the author, negative consequences can be found up to 4 degrees of separation away from the author's citation network. Retractions not only effect the author's credibility but also the author's contributions to the wider research field. With about 25 % of authors leaving their scientific publishing careers shortly after a retraction [14]

The negative effects are not only contained in the scientific world, as mass retractions of papers risk destroying public trust in science. [15]

The negative impact of a paper is not effectively mitigated by the paper's retraction, as by the time of retraction, the fraudulent paper has already garnered most of its public attention, spreading misinformation. [16] This is primarily due to the time of retraction being too long to effectively mitigate the effects of fraudulent research.

One cause of significant concern regarding the integrity of research is the large number of citations that retracted papers have in unretracted papers. While some of these citations may be due to showing an example of fraudulent research, some inevitably are not, and the outcomes of these unretracted papers might have an unconscious bias based on the fraudulent research they cited and detecting and retracting these types of papers and mitigating the possible wider damage research field is much more complicated than tradition retracted papers [17].

2.1.3 Current Retraction Mechanisms

Research conducted by Wittau and Seifert shows that currently, retractions detection approaches are split up into two main sections: content-related approaches and non-content-related approaches[18].

Content-related approaches usually analyze the data within the paper, techniques such as asking for the original data at the time of submission as fraudulent papers would struggle to produce this data ([18]-[19]), or analyzing images to detect if they have been manipulated ([18]-[20]) [21].

Non-content-related techniques include analyzing per-review patterns where techniques are used to detect if fake favorable reviews are being submitted ([18]-[22]), using AI detection systems to check the writing style and language of papers to check if ai generated these texts([18]-[23]), comparing papers with already proven fake papers as a strong textual similarity means that the is a likelihood that they were produced by the same paper mill and detecting anomalies and metadata features as these might be an indicator of fake paper([18]-[11]) [21].

With the rapid increase of paper retraction and the rise of paper mills, publications have had the pressure to uphold academic integrity. Springer Nature has unveiled two new AI tools, Geppetto, which is used to check for AI-generated fake content, and SnappShot, which is an image integrity analysis tool [24].

There are gaps in the current retraction mechanisms used in science. The current techniques employed are not utilizing the power of scientometrics and are opting for more traditional techniques. Our research aims to utilize the power of scientometrics and machine learning to create prediction model.

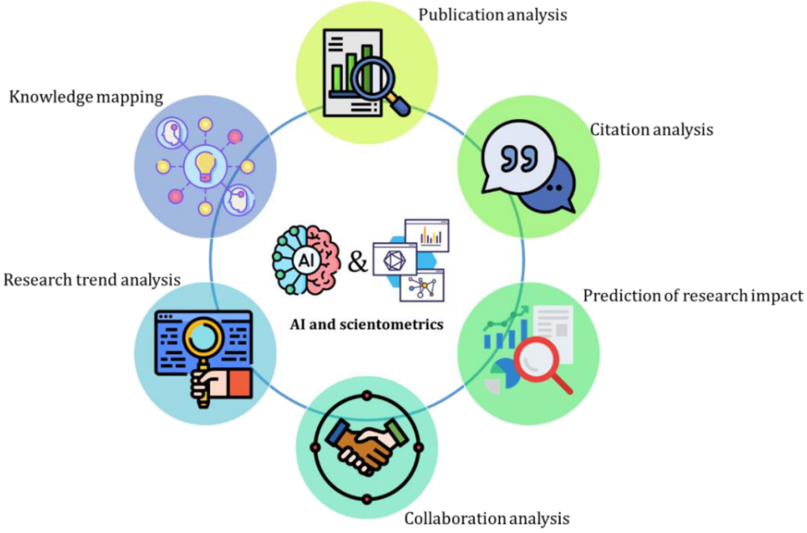


Figure 2.1: Six specific benefits AI can provide to scientometrics [26]

2.1.4 Scientometrics

Scientometrics is the quantitative measure of science. It allows us to quantify and study research and its development. [25]

The synergy of using scientometrics in AI has been observed in the research world, with researchers gaining greater insight into specific fields due to the power leveraged by AI. Specifically, researchers have used citation networks to predict the impact and influence of individual pieces of scientific research and the relationship between these papers. In addition, researchers have used co-authorship networks to understand research networks, allowing them to gain a better insight into collaborative networks and the influence of collaborators. [26]

With the clear use of AI and scientometrics proving to yield promising results, there is an area with potential in research space where the use of AI and scientometrics in detecting paper retraction is still unexplored. By networks and training our machine learning model based on information gathered from these networks, similar to research previously conducted using scientometrics and AI, we can hopefully garner useful results.

Citation Networks

Our research in this paper will involve constructing citation networks and utilizing them to train our model for paper retraction detection. According to research conducted by Goldberg [27], Citation networks are networks where the nodes are documents, and there are directed edges that point to other documents they cite. They further explain how this notion of a network captures information about the ”flow of innovation” and help us understand ”large scale patterns” within the data.

The hidden information that citation networks can reveal means that they have been previously used as part of research in the scientific research world. Kusumastuti leveraged citation networks to conduct research into literature defining aging, using quantitative analysis of citation network [28], and Greenberg went a step further in researching the effectiveness of studying citation networks to understand the evolution of a scientific belief. He concluded that citation networks are a powerful tool in understanding social communication and observing cascading information [29]. Our research attempts to exploit this underlying information that citation networks can reveal to evaluate the effectiveness of using this information to detect the likelihood of a paper being retracted.

Co-Authorship Networks

Fonseca [30] defines a co-authorship/co-publication network as a network where nodes are authors, and there is an edge between authors if they share the authorship of a paper. They further elaborate on the use of co-publication networks in scientific research as analysis of co-publication networks is widely used to reveal and understand underlying collaboration patterns. In their research, they utilize degree centrality from the network to identify the most central organizations, which spread the most influence in the field and further explain how these nodes ”control information and knowledge flow between two separate groups”.

Scientometric utilization in the form of co-authorship networks in research leads to promising insight into interdisciplinary trends, allowing researchers to answer questions surrounding the development of various trends in specific research fields [31]. The outcome of their research is quite promising as the insight into trends they obtained from co-authorship networks could be applicable to detecting a possible trend in papers that are retracted.

In our research, we will leverage the spread of knowledge and influence that a co-publication network represents and, based on that idea, create a new novel network, ’Meta-Net’, to possibly obtain the type of underlying patterns that are revealed in a co-publication network. By capturing these underlying patterns and the cascading influence of retracted/non-retracted papers, we could possibly quantify and use it coupled with a model to get meaningful results.

2.1.5 Previous Research in predictive modeling for paper retractions

Previous research in this field has tried a similar technique to this research paper. A study conducted by Sai Ajay Modukuri [32] attempts to solve this problem by making a machine learning model trained on metadata features and full-text features. They analyzed the metadata and the full text and used various techniques to extract features from these two types of data, which they then used to train their model. Their results show that using the metadata features alone gave an F1 score of 67%, and using full-text features alone gave an F1 score of 63%. Combining both metadata and full-text features gave a combined F1 score of 71%. While using just the metadata and text gave promising results, their work leaves areas unexplored especially the relationship between papers and also the relationship between co-authors. They used citation metrics and semantic scholar to provide the intent behind each citation and used institutional rankings based on Times Higher Education to give the lead author a score, but they failed to capture the patterns and networks created by these research papers. As stated in [3] the entirety of the science is a complex network. In this research project, we plan to construct some of these networks and then extract some relevant network features to train our model, exploiting the relationships between the papers and authors.

Chapter 3

Project Goals

3.1 Objective

This project aims to develop prediction models for paper retractions, trained on the paper's metadata, citations, and ”Meta-Net” network features. The models would provide an initial screening classification for whether or not a paper is likely to be retracted, acting as an initial check that allows for early intervention by a third party.

3.2 Approach

This research approach will follow this structure, with each step building on top of the other. The project will involve experimenting and researching with multiple techniques.

Traditional Metadata ⟶ Citation Metrics ⟶ Citation Network ⟶ Meta-Net

Figure 3.1: High level overview of the workflow.

Each technique will follow the same approach. Traditional metadata and citation metrics will involve using meta data. Citation network and Meta-Net will be network based. The project's process has a clear, step-by-step, logical flow, and each step is further deconstructed into more manageable components.

3.2.1 Data Collection

This step would involve collecting a large quantity of relevant and accurate data needed for network construction and model training.

Steps:

- Find relevant databases that include the necessary information needed in addition to information on the retraction status
- Extract the data from the database, whether through a predefined program or web-scrapper
- Clean and process the data

3.2.2 Data Construction

Once we have collected and processed all the data, we would need to construct/process it into an appropriate form.

Steps:

- For techniques that use meta data (Traditional Metadata/Citation Metrics), data processing will be done.
- For techniques requiring networks (Citation Network/Meta-Net), the data will be constructed into networks.

3.2.3 Select and Fine-Tune Model and Evaluate

This is an iterative step, where we would select a model, fine-tune it, and evaluate it, approximating the most relevant features, model choice, and parameters

Steps:

- Analyse and compute network features, preparing the relevant data features for each entry
- ***Iterative*** **:** {Choose and train a model using the data, then evaluate the model based on various metrics. Then we need to refine the model based on results }

Chapter 4

Data Collection and Analysis

In this research paper, we will move forward, with research techniques that involve traditional meta data and citation meta data. We would then construct networks using this data to try to research if that would yield promising results. We would need to collect this data.

4.1 Traditional Meta Data

4.1.1 Data Collection

The first step of the proposed design for this project involved drafting and selecting a single unique database/search engine for collecting our datasets. The choice behind the single database/search engine is to allow for a streamlined approach to data collection. Using a single point for our data would allow for uniform data with the same features, making the data engineering phase seamless. We drafted the major databases/search engines to compare and finalize one. The draft was composed of PubMed, Web of Science, and Google Scholar. These were the significant sources of publication data at the point of this project. We had to narrow these down further to one based on these factors.

API

API is a crucial component that is actively used when we are collecting our data. Google Scholar does not have native API support. However, there are external open-source libraries that could be used for retrieving information from Google Scholar, such as pygscolar [33] and scholarly [34]. PubMed has native API support for its database. It has various native APIs, but the most relevant for this project would be Entrez Programming Utilities API [35]. Web of Science has a wide selection of APIs that would be relevant to this project. Additionally, Web of Science Expanded API allows for the most extensive range of relevant information to be retrieved for our project, including native citation data, which PubMed does not offer. When using PubMed, we would have to opt for external sources to retrieve citation data, which would not be a problem when dealing with the Web of Science API Expanded as it offers a wide range of citations and other meta-level information that would be relevant further along our project.

Specialiastion

Web of Science and Google Scholar do not have a specific field specialization as they host a wide range of data from various disciplines. However, PubMed is specialized to host life-science and biomedical-related research.

After considering these factors, we decided to use Web of Science as our database. The next step in our approach was to gather data from the database. To do this, we had to apply to the Web of Science API Expanded for an application. During the application process, our application was rejected; the stated reason was that our institution did not have a valid subscription with them. This was brought up to the university in a long transaction of emails, and after a lengthy process, we were told that we would not be provided a web of science api, moving forward we decided that the next best option would be using pubMed.

4.2 Citation Data

4.2.1 Data Collection

To collect citation data due to database limitations, we were not able to collect this citation data; when querying our traditional meta data, we had to use a 3rd party database to query citation information for a given paper. We used opencitations [36] an open public database with citations and references for a given paper. Using this approach was successful in gathering citation information needed for citation metric analysis and citation network analysis, but we could not get enough papers with traditional meta data and citation information, meaning we could not create a model where we aggregated both metrics.

4.3 Data Analysis/Insight

Features
Title
Subject
Institution
Journal
Country
Author
Original Paper Date
Abstract

Table 4.1: Traditional Meta Data

Features
Citations
References

Table 4.2: Citation Meta Data

Note that due to our API limitations, we were not able to obtain a large number of abstracts for retracted papers, with the number of retracted entries going from 12220 to 1779. We will use a smaller dataset for training with abstracts, with 3558 entries.

For every instance of training, there will be the same number of retracted papers as non retracted papers, with the same number of papers for each topic selected in our retraction watch database after filtering for subjects that are related to computer science/technology/engineering here is the subject that is related and the number of paper-positive instances we have:

Subjects	Count
(B/T) Computer Science	5505
(PHY) Engineering - Electrical	609
(B/T) Technology	3391
(B/T) Data Science	1640
(PHY) Mathematics	699
(PHY) Physics	190
(PHY) Engineering - Mechanical	448
(PHY) Engineering - General	511
(PHY) Statistics	54

Table 4.3: Total number of retracted papers in our dataset, with relevant subjects

Analysis into the retracted papers

Looking at the full retraction watch dataset, we can derive some information on the nature of the retraction.

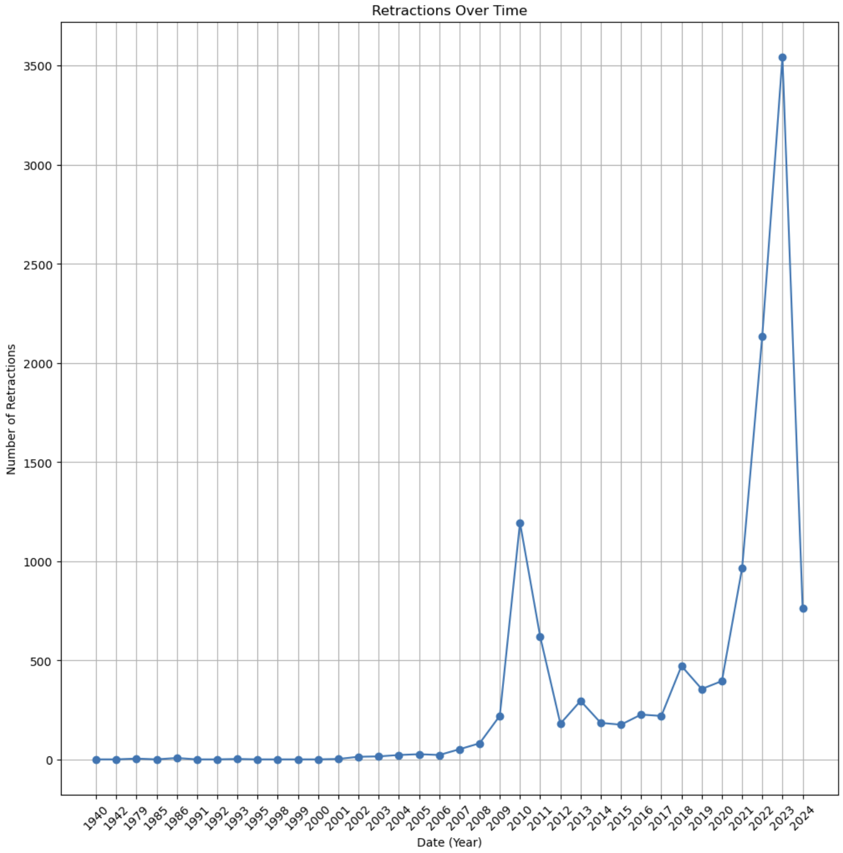


Figure 4.1: Number of Retractions Per Year Graph

Looking at this number of retractions by year graph, it is clear that there is a clear upward trend, with the number of retractions increasing per year. It is also apparent that certain years have an abnormally high number of retractions compared with their surrounding years, serving as anomalies.

Analyzing the dataset for the country of each retracted paper gives us an interesting insight. Certain countries have a much larger number of retracted papers compared with the rest of the dataset.

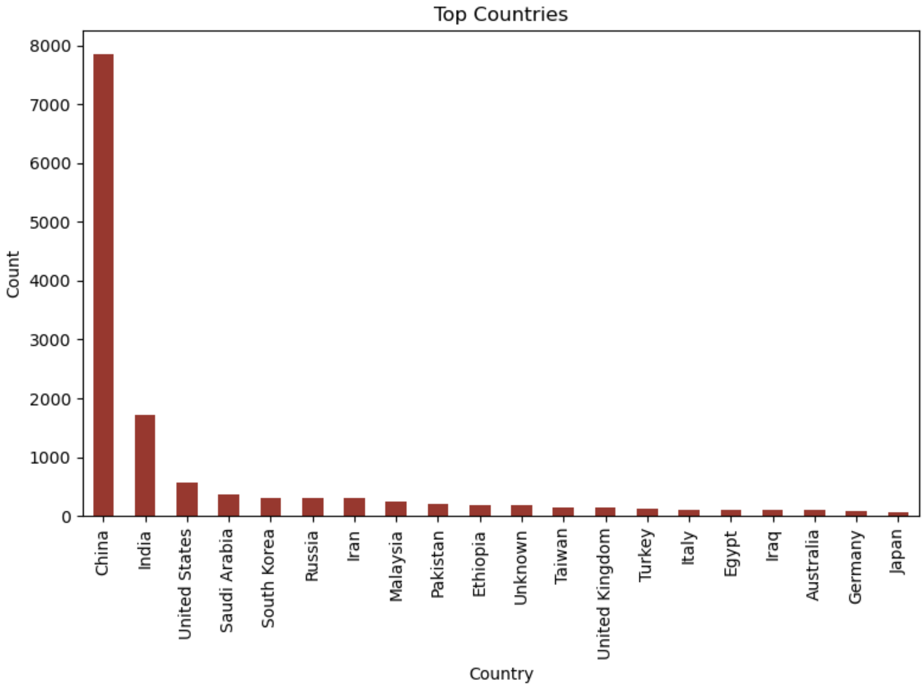


Figure 4.2: Top Countries with the most retractions

China and India are strict outliers 4.3. The data shows us that there is a percentage increase of 198% from the US to India and a percentage increase of 1271.55% from the US to China. India and China account for 65.44% of all the retracted papers in the dataset.

- United States count: $C_{\text{US}} = 573$
- India count: $C_{\text{India}} = 1708$
- China count: $C_{\text{China}} = 7859$

$$\text{Percentage of India and China} = \frac{C_{\text{India}} + C_{\text{China}}}{\sum C} \times 100 = 65.44\% \quad (4.1)$$

$$\text{Percentage increase from US to India} = \frac{C_{\text{India}} - C_{\text{US}}}{C_{\text{US}}} \times 100 = 198\% \quad (4.2)$$

$$\text{Percentage increase from US to China} = \frac{C_{\text{China}} - C_{\text{US}}}{C_{\text{US}}} \times 100 = 1271.55\% \quad (4.3)$$

Figure 4.3: Country Retraction Stats

When analyzing the reasons for retractions annotated in the dataset, even tho we can not use this data in our training, looking at it means we can get a better idea of the context behind the cause of the retractions.

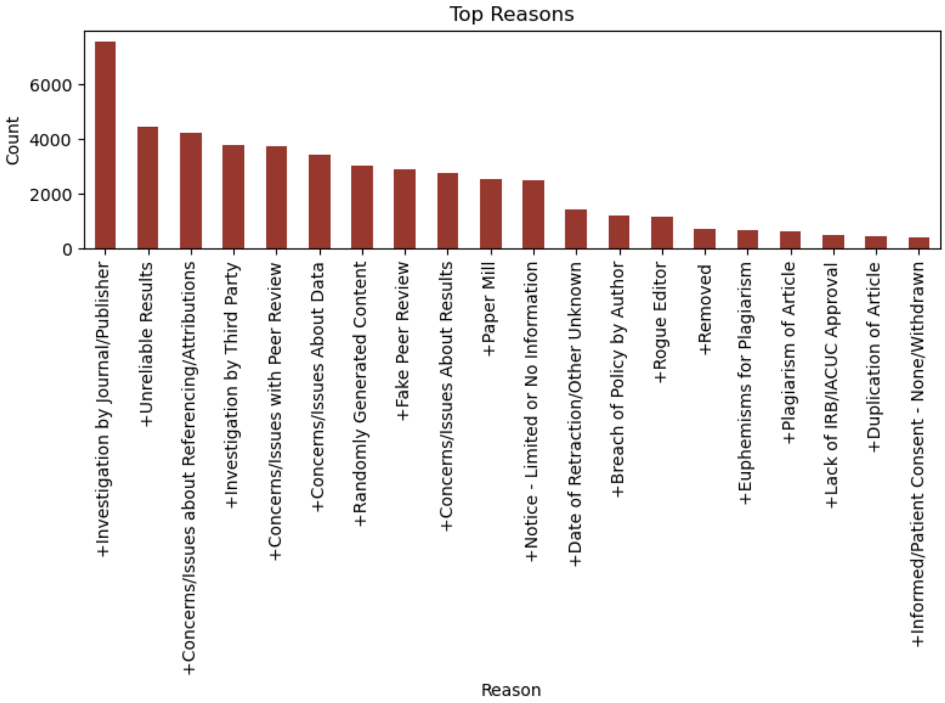


Figure 4.4: Top reasons for retractions

Chapter 5

Data Preprocessing

This section includes the processing done to the raw data collected to prepare it for training our models.

Abstract

First, we need to apply pre-processing. The process would start by lowercasing all the text and removing all the punctuation. From here, we would tokenize the abstract and remove all the stop words (the, and ...). From this point, we would apply lemmatization.
We are focusing on capturing some semantic meaning from the abstract and using that as a potential identifier for its retraction likelihood. Cleaning and normalizing the text are essential for this. The process we are taking is the common principled standard approach taken for language processing.
The abstract in its raw form is qualitative and thus can not be used to train our model. We need to quantify this qualitative data so that it can be used to train our model.

- **Bag-of-Words :** The Bag of Words is a simple approach to use; it treats each abstract as a document, and for each of these documents, it makes a vector. Each dimension in the vector corresponds to a unique word in the abstract; the value contained in this dimension would be the count of this unique word in the abstract [37]. This approach is quite simple and is easily interpretable, but it focuses on the frequency of the words rather than their semantic meaning; the vectors also tend to be sparse, where common words tend to be significant, disregarding potentially important semantically valuable words.

- **TF-IDF :** The Term Frequency-Inverse Document Frequency method is quite similar to the bag of words approach, but rather than solely focusing on the simple frequency of each unique word in each respective document, TF-IDF also takes into account how rare each word is in the entire set of documents. [37]

- **Word2Vec:** This approach uses Neural Networks to represent each word by a vector; it produces low dimensional dense vectors where semantically similar words are placed close to each other in the vector space; there are 2 types of architectures that you can take when using Word2Vec. [38]
  - **Continuous Bag of Words Model**
  - **Skip-Gram**

Categoric Meta-Data

The categoric data will be quantified using One Hot Encoding.

- **Subject**
- **Institution**
- **Journal**
- **Publisher**
- **Country**

One-Hot Encoding: This is an approach for quantifying categoric data. In One-Hot encoding, each unique category is represented by a binary vector, the length of which is the number of categories in that respective category. All categories that are false are set to 0, and the categories that are true are set to 1.

Author

- **Author Count:** For each respective entry, count the number of authors that have co-authored the paper.

Initially, for this section, we decided to include more author metrics, where the metrics included total citations, h-index, and i10-index. However, due to API limitations, we couldn't move forward as we were not able to obtain these metrics.

Paper Publication Date

- **Publication Year:** To capture yearly trends that could possibly lead to a correlation with retraction probability. Possibly identifying periods where the retraction rate is higher could be a useful metric.
- **Age of Paper:** The age of a paper could also be a useful metric when coupled with citation count or other metrics, as there may be patterns that emerge that correlate to the retraction probability. The age of the paper will be calculated using this equation

$$Age(paper) = \frac{2025 \ - \ \text{Year of Publication}}{365}$$

References

- **Reference Count:** Count the number of times the paper has referenced another paper. This might possibly reveal a correlation where the reference count is linked to the quality of the paper or the legitimacy of the paper.

Citations

- **Total Citation Count:** Count the number of citations the paper has. The total citation count will reveal its recognition and impact within the scientific field.

- **Citation Rate:**

$$CitationRate(paper) = \frac{\text{Total Citation Count}}{\text{Age(paper)}}$$

  The Citation Rate is mainly there to mitigate the fact that older papers have a high likelihood of having more citations as they have more time, but using the citation rate, we can see their relative importance regardless of the period of time they have been available.

Normalization

Categoric Data: Our approach of using Hot-One Encoding inherently will normalize the data; thus, further normalization is not needed.
Continuous Data: By nature, due to continuous data having different scaling, normalization is needed. The following data will be normalized using Z-Score normalization.

- **Author Count**

- **Publication Year**

- **Age of Paper**

- **Reference Count**

- **Total Citation Count**

- **Citation Rate**

Stabilize Data: Data whose internal data distribution is skewed needs to be transformed before applying z-score normalization. Data where the internal data distribution is skewed would lead to values that are much higher than the rest of the data, disproportionally affecting the mean and variance.

The following data are likely to have a skewed data distribution, and a variance stabilizing transformation would be applied:

- **Reference Count:** There might be specific papers that have a disproportionally higher number of references.

- **Total Citation Count:** The number of citations per paper is likely to be skewed, where there will be a subset of papers that are highly cited. To prevent our model from over-emphasizing these dominant entries, we need to stabilize the data before normalizing.
- **Citation Rate:** Citation Rate is the same as Total Citation Count, where the internal data distribution is likely to be skewed.

The techniques that we will be using would be:

- **Z-Score Transformation:**

  $$z = \frac{x - \mu}{\sigma}$$

  z-score transformation would transform every data point x to data point z, where μ is the mean of the data, and σ is the standard deviation of the data. This transformation makes it so that the newly transformed data would have a mean of 0 and a standard deviation of 1, as this would be applied across all the continuous data types mentioned above; this would make the data directly comparative irrespectively of the different scales initially, and allow our ML algorithms to converge quicker.

- **Log Transformation:** The stabilizing transformation we will be using is Log Transformation.

  $$LogTransform(x) = Log(1 + x)$$

  Where we apply this equation to every data point, compress the range of the data set, and reducing the influence of outliers, in our data set, stabilizing our variance.

Chapter 6

Model Selection And Evaluation Metrics

6.1 Machine Learning Model Selection

- **Logistic Regression:** It is a simple, interpretable linear model. That predicts which binary class an input falls into using the logistic function. It is best used to provide baseline models that can be used to leverage more powerful models. However, the main drawback of this model is that it works best when there are linear relationships between the features, and when the features are non-linear, it often performs poorly.

- **Random Forests:** This technique is an ensemble learning technique, where multiple independent decision trees are trained, where each tree is trained on a random subset of features, and together this forest of trees' decisions are aggregated together by majority vote to give a prediction for an input. This technique handles non-linear relationships well and is very effective at reducing variance and improving robustness; however, due to the nature of the algorithm, it is less interpretable than logistic regression, and it is more computationally intensive when compared with logistic regression.

- **Support Vector Machines:** This algorithm works by finding a hyperplane using kernels, which separates the 2 classes. It handles non-linear separations by having a variety of kernels. It is effective in high-dimensional spaces and performs well with a robustness to overfitting. However, it is sensitive to feature scaling and parameter tuning, in addition to being less interpretable. SVMs are quite computationally intensive.

- **Gradient Boosting:** This is a powerful ensemble learning technique that is built on a series of interdependent sequential decision trees, where each tree accounts for the errors made by the previous trees; there are various implementations of this algorithm like AdaBoost and XGBoost. It provides high accuracy and predictive capabilities. Due to the nature of the algorithm being interdependent, it captures the complex interactions between features well. However, it is not very interpretable, and the computational expense is low.

When selecting which algorithm is going to be used for our model, we need to consider a few points. First is the nature of the data we are working with. Before any processing, the initial state of the data is varied, with different types of data, such as numerical, textual, and categorical. This would lead us to the possible conclusion that there are going to be complex interactions between our features, where, most likely, the correlation would be non-linear, so our model would need to account for this.
The structure of our approach will involve using 2 primary models and a baseline model.

- **Baseline Model:** This model would provide a baseline a performance to use to benchmark our other algorithms

- **Primary Model:** This would be our strongest, most optimal model, where it should provide the best predictive results.

Choice for Model

- **Gradient Boosting:** This algorithm will be used for one of our primary models. GB model's strengths include being very effective at capturing complex interactions between features and being able to handle non-linear relationships.

- **Random Forest:** This is also going to be our choice for the second primary model. It is well suited to handle complex relationships and is good for reducing variance and being robust to noise.

- **Logistic Regression:** The requirements for our baseline model is that it should be simple, as it is only here to provide a baseline performance, which can be used to gauge how much more powerful the other models are. Logistic Regression fits all our requirements, making it a good choice.

6.2 Training and Validation

6.2.1 Data Splits

We are going to be going with the approach of 2 different sets.

- **Training Set (70% of data)** – This set will be used to train our model and tune our hyperparameters

- **Test Set (30% of data)** – This will be used to evaluate our trained model's performance

6.2.2 Cross Validation

To further evaluate our model's stability and reliability, we will perform k-fold cross-validation, averaging its performance against subsets of our data.

6.2.3 Evaluation Metrics

- **Accuracy:** This is the total number of accurate predictions over the total number of predictions.

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  It gives a good overview on how the model performs overall, across every class. It is quite a simple metric, and does not account for class imbalances, in our case we made sure to have a 1 to 1 of positive and negative cases, so this drawback of this metric should not effect us.

- **Precision:** The proportion of actual positive cases classified as positive over the total number of positive classifications by our model.

  $$\frac{TP}{TP + FP}$$

  It gives us the proportion of the model's positive classifications that are actually positive.

- **Recall:** The equation correctly identifies the number of positives over the total number of positives

  $$\text{Recall} = \frac{TP}{TP + FN}$$

  This metric gives us a good understanding of the proportion of positive cases caught by our model.

- **ROC-AUC:** Reveals the model's ability to distinguish our classes accurately across different thresholds. It is a good indication of the general predictive power of our model.

- **F1-Score** The harmonic mean of Precision and Recall is a good metric when it reflects how well it does on both false positives and false negatives.

Chapter 7

Methodology

7.1 Traditional Meta Data

For training with traditional meta data, we will use the preprocessed metrics discussed in 5.

Key

LR = Logisitic Regression RF = Random Forest GB = Gradient Boost

$$Classes = \begin{cases} 1, & \text{Paper is retracted,} \\ 0, & \text{Paper is not retracted.} \end{cases}$$

All results are rounded to 2 d.p

7.1.1 Abstract

TF-IDF

<table>
<tr><td colspan="5">TF-IDF (LR)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.74</td><td>0.77</td><td>0.76</td><td>0.75</td></tr>
<tr><td><b>1</b></td><td>0.76</td><td>0.73</td><td>0.75</td><td>0.75</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.75</b></td></tr>
<tr><td colspan="5">TF-IDF (RF)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.77</td><td>0.82</td><td>0.79</td><td>0.78</td></tr>
<tr><td><b>1</b></td><td>0.81</td><td>0.75</td><td>0.78</td><td>0.78</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.78</b></td></tr>
<tr><td colspan="5">TF-IDF (GB)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.72</td><td>0.79</td><td>0.76</td><td>0.74</td></tr>
<tr><td><b>1</b></td><td>0.77</td><td>0.70</td><td>0.73</td><td>0.74</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.74</b></td></tr>
</table>

Table 7.1: Cross validation results for TF-IDF with different ML models.

Bag of words

<table>
<tr><td colspan="5">Bag of words (LR)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.79</td><td>0.86</td><td>0.82</td><td>0.81</td></tr>
<tr><td><b>1</b></td><td>0.85</td><td>0.77</td><td>0.80</td><td>0.81</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.81</b></td></tr>
<tr><td colspan="5">Bag of words (RF)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.78</td><td>0.85</td><td>0.82</td><td>0.81</td></tr>
<tr><td><b>1</b></td><td>0.84</td><td>0.76</td><td>0.80</td><td>0.81</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.81</b></td></tr>
<tr><td colspan="5">Bag of words (GB)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.75</td><td>0.75</td><td>0.75</td><td>0.75</td></tr>
<tr><td><b>1</b></td><td>0.75</td><td>0.75</td><td>0.75</td><td>0.75</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.75</b></td></tr>
</table>

Table 7.2: Cross validation results for Bag of words with different ML models.

Word2Vec

<table>
<tr><td colspan="5">Word2Vec (LR)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.70</td><td>0.71</td><td>0.71</td><td>0.70</td></tr>
<tr><td><b>1</b></td><td>0.71</td><td>0.70</td><td>0.70</td><td>0.70</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.70</b></td></tr>
<tr><td colspan="5">Word2Vec (RF)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.72</td><td>0.78</td><td>0.75</td><td>0.74</td></tr>
<tr><td><b>1</b></td><td>0.76</td><td>0.70</td><td>0.73</td><td>0.74</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.74</b></td></tr>
<tr><td colspan="5">Word2Vec (GB)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.69</td><td>0.78</td><td>0.73</td><td>0.71</td></tr>
<tr><td><b>1</b></td><td>0.75</td><td>0.65</td><td>0.69</td><td>0.71</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.71</b></td></tr>
</table>

Table 7.3: Cross validation results for Word2Vec with different ML models.

7.1.2 Country

<table>
<tr><td colspan="5">(LR)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.74</td><td>0.99</td><td>0.85</td><td>0.82</td></tr>
<tr><td><b>1</b></td><td>0.98</td><td>0.66</td><td>0.79</td><td>0.82</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.82</b></td></tr>
<tr><td colspan="5">(RF)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.87</td><td>0.96</td><td>0.91</td><td>0.91</td></tr>
<tr><td><b>1</b></td><td>0.95</td><td>0.86</td><td>0.91</td><td>0.91</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.91</b></td></tr>
<tr><td colspan="5">(GB)</td></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.74</td><td>0.99</td><td>0.85</td><td>0.82</td></tr>
<tr><td><b>1</b></td><td>0.98</td><td>0.66</td><td>0.79</td><td>0.82</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.82</b></td></tr>
</table>

Table 7.4: Cross validation results for Country with different ML models.

7.1.3 Subject

		(LR)		
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.48	0.51	0.49	0.48
1	0.48	0.45	0.46	0.48
ROC-AUC Score: 0.48				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.48	0.40	0.44	0.48
1	0.48	0.56	0.52	0.48
ROC-AUC Score: 0.48				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.48	0.58	0.53	0.48
1	0.47	0.37	0.42	0.48
ROC-AUC Score: 0.48				

Table 7.5: Cross validation results for Subject with different ML models.

7.1.4 Journal

		(LR)		
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.72	1.00	0.84	0.81
1	1.00	0.61	0.76	0.81
ROC-AUC Score: 0.81				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.84	1.00	0.91	0.91
1	1.00	0.81	0.90	0.91
ROC-AUC Score: 0.91				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.75	1.00	0.85	0.83
1	1.00	0.66	0.79	0.83
ROC-AUC Score: 0.83				

Table 7.6: Cross validation results for Journal with different ML models.

7.1.5 Author

	(LR)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.74	0.57	0.64	0.69
1	0.65	0.80	0.72	0.69
	ROC-AUC Score: 0.69			
	(RF)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.67	0.82	0.74	0.71
1	0.77	0.60	0.67	0.71
	ROC-AUC Score: 0.71			
	(GB)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.67	0.82	0.74	0.71
1	0.77	0.60	0.67	0.71
	ROC-AUC Score: 0.71			

Table 7.7: Cross validation results for Author with different ML models.

7.1.6 Institution

	(LR)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.51	1.00	0.67	0.52
1	1.00	0.03	0.07	0.52
	ROC-AUC Score: 0.52			
	(RF)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	1.00	0.01	0.02	0.50
1	0.50	1.00	0.67	0.50
	ROC-AUC Score: 0.50			
	(GB)			
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.50	1.00	0.67	0.51
1	1.00	0.01	0.02	0.51
	ROC-AUC Score: 0.51			

Table 7.8: Cross validation results for Institution with different ML models.

7.1.7 Date

<table>
<tr><th colspan="5">(LR)</th></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.83</td><td>0.96</td><td>0.89</td><td>0.88</td></tr>
<tr><td><b>1</b></td><td>0.95</td><td>0.80</td><td>0.87</td><td>0.88</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.88</b></td></tr>
<tr><th colspan="5">(RF)</th></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.83</td><td>0.96</td><td>0.89</td><td>0.88</td></tr>
<tr><td><b>1</b></td><td>0.95</td><td>0.80</td><td>0.87</td><td>0.88</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.88</b></td></tr>
<tr><th colspan="5">(GB)</th></tr>
<tr><th></th><th>Precision</th><th>Recall</th><th>F1-Score</th><th>Accuracy</th></tr>
<tr><td><b>0</b></td><td>0.83</td><td>0.96</td><td>0.89</td><td>0.88</td></tr>
<tr><td><b>1</b></td><td>0.95</td><td>0.80</td><td>0.87</td><td>0.88</td></tr>
<tr><td colspan="5"><b>ROC-AUC Score: 0.88</b></td></tr>
</table>

Table 7.9: Cross validation results for Date with different ML models.

7.2 Analysis For Traditional Meta Data

Analysis of each feature:

7.2.1 Abstract

Performance: Using textual features, in the form of abstracts, for training shows us that there is a clear, strong predictive power in these features. From our testing, Bag of Words has the best performance, where LR and RF both have an accuracy of 81% and ROC-AUC Score of 0.81. TF-IDF has the second-best performance, with RF giving us the best results with an accuracy of 78% and a ROC-AUC Score of 0.78. This is then followed by Word2Vec having the worst performance out of all the algorithms, where RF outperforms all the models, having an accuracy of 74% and a ROC-AUC Score of 0.74. Precision, Recall, and F1-Score are quite balanced across class 0 and class 1

Interpretation: Bag of words outperforming all the other algorithms suggests that capturing simple word frequencies is more effective than using a technique that tries to capture the importance of words (TF-IDF) or their semantic meaning (Word2Vec). However, this could be due to the inability to capture meaningful semantic relationships or capture the importance of words across documents, because of the size of the abstract, compared to full text analysis of the paper, abstracts are smaller, this may be a key indicator to why BOW is outperforming, the other algorithms. Using full text analysis may yield better results for TF-IDF and Word2Vec. A possible reason why BOW outperforms the other algorithms and shows strong predictive power may lie in the paper mills[9]. These mills might have a possible bias correlation towards the word frequency count when producing papers, and as a result, irrespective of the topic, which may influence the semantics using a comparatively basic algorithm, may reveal this hidden trait, yielding strong predictive power.

7.2.2 Country

Performance : Country shows very high predictive power, where the best performing model is RF with an accuracy of 91% and ROC-AUC Score of 0.91, with LR and GB also showing a strong predictive power, with LR having an accuracy of 82% and a ROC-AUC Score of 0.82, and GB having an accuracy of 82% and ROC-AUC Score of 0.82.

Interpretation : Looking deeper into the results, this very strong predictive power showcased for countries may be a source of concern. One possible reason for this outcome could be the variation of research integrity, standards, and cultures across different countries, which causes countries such as China, India, and South Korea to have abnormally higher retraction ratios compared to countries like the US, EU, and Japan. [39] From the analysis conducted on our dataset, China and India were strict outliers 4.2, whereas previously discussed, there was a 198% percentage increase from the US to India, and there is a percentage increase of 1271.55% from the US to China, with India and China accounting for 65.44% of all the retracted papers in the dataset 4.3. There is a possibility that the dataset used for training and testing could have been heavily skewed, with a large majority of class 1 instances (retracted papers) being from countries like China and India, coupled with these countries being in the vast minority for class 0 instances (not-retracted papers), causing over and under-representation respectfully in the dataset, making it more likely that our model is not learning general patterns of papers that are retracted, but rather creating direct links such that if paper was from a country like China that means it is a retracted paper.

This is further evidenced by the class imbalance present in our results. LR and GB have extremely high precision (near perfect) of 98% for class 1, meaning that when the model predicts retraction, it is most likely correct, but it has a low recall of 66%, meaning it misses quite a lot of actual retracted papers. This is evidence because if our class 1 dataset is overrepresented by instances from China and India, and our class 0 dataset is underrepresented by those countries, this would align with these results. The high precision could be the fact that our model has developed a bias for predicting papers from China and India as retracted, and because our dataset is skewed whenever it predicts these papers are retracted, it is correct, but also our model does not fare well with retractions from countries that are not from these countries, thus the comparatively low recall of 66% as it is missing a lot of retractions from these other countries.

Ethical concern : Compared to training on other features, training on the country does have a heavier ethical concern to take into consideration. Having the country of the paper being used disregards the paper's individual research quality but rather causes the paper to be aggregated with the country of origin metric, for which that paper nor the other papers released by those authors have any meaningful impact. This unfair flagging puts good, legitimate research coming from certain countries at a disadvantage.

7.2.3 Subject

Performance : Subject has around 48% Accuracy and 0.48 ROC-AUC across all the models.

Interpretation : This means that the Subject feature has virtually no predicting power, with the model being close to randomly guessing. This, however, does reflect our dataset, as we purposefully chose the same subjects and distribution across our class 0 and 1 datasets.

7.2.4 Journal

Performance : Journal showcases a very high predictive power, with RF having an accuracy of 91% and ROC-AUC of 0.91. The other models follow RF with weaker but still very high predictive power, with LR and GB having an accuracy of 81% and 83% and having a ROC-AUC of 0.81 and 0.83 respectfully.

Interpretation : Specific journals could have a very high retraction rate compared to the norm due to numerous factors; some journals may have a much stricter and more rigorous review process. Additionally, some journals may be in a field with a high number of retractions. [40]. These abnormally high predictive results may be due to our dataset containing a bias for the class 1 instances, where many of the journals in the class 1 instance are victims of paper mills [9]. Our model could possibly be linking the cause of retraction directly to this subset of journals rather than learning a generalized pattern. The likelihood of the dataset being biased is further evidenced by the precision and recall scores. All 3 models, have a class 1 precision of -100% and class 0 recall of -100%, meaning when they flag a paper as retracted they are almost always correct and they almost perfectly assign all the class 0 or un-retracted papers as such, but they fail correctly assign a large number of actually retracted papers as retracted, this is particularly evident in our LR and GB model where the recall for class 1 is 61% and 66% respectfully, meaning they are missing 34-39% of actual retracted papers, meaning the model is associating retraction with a specific subset of journals while failing to generalize.

7.2.5 Author

Performance : Moderate predictive power, with accuracy being 69-71% and ROC-AUC score being 0.69-0.71 across all models.
Interpretation : There is a clear sign of some predictive signal within the author count, as it is not making random guesses; this metric would provide a meaningful impact in a larger model.

7.2.6 Institution

Performance : 0 predictive power, with accuracy being around 50% and ROC-AUC being around 0.5.
Interpretation : It provides virtually no predictive power, with the model virtually randomly guessing. The LR and GB model predicts almost every instance as class 0, which is evident in our class 0 recall being 100% and our class 1 recall being 0.3%, and our RF model predicts every instance as class 1, where class 1 recall is 100% and class 0 recall is 0.1%. This is most likely due to our dataset having inconsistent naming and having no unified encoding. Our model is not generalizing and does not provide information.

7.2.7 Date

Performance : Very strong predicting power, with accuracy being 88% and ROC-AUC being 0.88 across all the models.
Interpretation : The features derived from the publication date show us that the model trained on that data produces a very powerful model. From earlier analysis 4.1, it is clear that the number of retractions per year is increasing rapidly; this is partly due to technological advances in detection tools, public awareness, and institutional change in what criteria is a retraction required [1]. Our analysis into our retraction database shows us that there is a skew, where a large majority of the database is from newer papers. Due to our random sampling of non-retracted papers (class 0), it is possible that our training dataset contains a bias, with a large majority of class 1 instances being newer and class 0 instances being older, making our model learn that newer papers mean its likely to be retracted, instead of generalizing. Additionally, with our dataset time frame containing the period of

time where the rate of retraction has gone up drastically, our model is identifying that this time-period is a strong indication of retraction probability.

7.2.8 Generalized Metadata Model

Combining all the nonproblematic features, we created a model to give us the best generalized results. We combined all 3 text feature encoding methods on the abstract and the data derived from subject and author. We chose these features because our previous analysis showed them to perform without problems.

(LR)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.80	0.85	0.82	0.82
1	0.84	0.78	0.81	0.82
ROC-AUC Score: 0.92				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.80	0.87	0.84	0.83
1	0.86	0.79	0.82	0.83
ROC-AUC Score: 0.83				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.77	0.83	0.80	0.79
1	0.81	0.75	0.78	0.79
ROC-AUC Score: 0.79				

Table 7.10: Cross validation results for optimized general model with different ML models.

Our combined model outperforms every other model that is trained on a single feature. All 3 models perform well, but the best performing model is RF, with an accuracy of 83% and a ROC-AUC score of 0.83. It also shows very good scores on Recall and Precision, without showing significant skewing or bias towards any class.

7.3 Citation Meta Data

For training with citation meta data, we will use the preprocessed metrics discussed in 5, specifically training the model on the processed data of references and citations.

(LR)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.81	0.83	0.82	0.77
1	0.73	0.70	0.71	0.77
ROC-AUC Score: 0.76				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.83	0.83	0.83	0.80
1	0.74	0.74	0.74	0.80
ROC-AUC Score: 0.79				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.82	0.84	0.83	0.79
1	0.75	0.73	0.74	0.79
ROC-AUC Score: 0.78				

Table 7.11: Cross validation results for citation meta-data with different ML models.

7.3.1 Citation Meta-Data Analysis

Performance : All models performed very well, showcasing the good predictive power of citation meta-data. The most powerful model was RF with an Accuracy of 80% and ROC-AUC score of 0.79, this is followed by GB and LR with an accuracies of 79-77% and ROC-AUC scores of 0.78-0.76, respectfully.

Interpretation : Due to the substantial difference between the ensemble models (RF,GB) performance compared to LR, it is possible that the metrics derived from references and citations interact in not only a linear way but may also contain non-linear patterns, which are captured by the ensemble methods. All methods are very effective in capturing non-retracted papers, with class 0 F1-Score being around 0.82-0.83 for all the models. However, the models' proficiency in capturing retracted papers is lower, with the best performing models being the ensemble models, with an F1-Score of 0.74 compared to the F1-Score of 0.71 for LR. This result still indicates good predictive power for capturing class 1 cases and good overall predictive power. Our models are leveraging citation information to get a strong indication of the retraction status, with the citation based models outperforming all the ”non-problematic” features (used on our general model 7.10) but Bag-Words model and being around the same as the TF-IDF model. This gives us evidence that this metric can be researched further by finding out if citation networks can give us further insights and reveal patterns that can be used for our models to possibly give us additional predictive power.

7.4 Citation Network

Construction

Our citation network is a directed graph-based network, where nodes represent a paper, and directed edges represent citations 7.1.

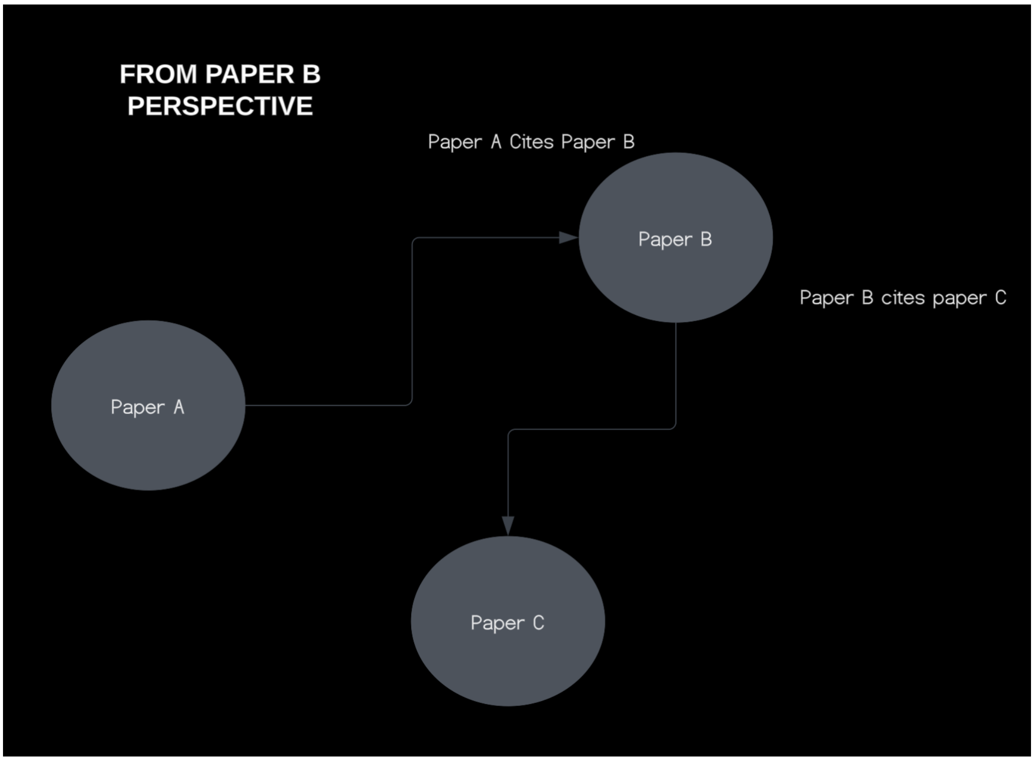


Figure 7.1: Segment of the citation network, POV from a single node

Interconnected-ness

An important issue to consider when constructing our citation network is interconnectedness. A disconnected network will involve smaller disconnected citation networks independent from each other 7.2, which could negatively impact the meaningful information we can derive from the network.

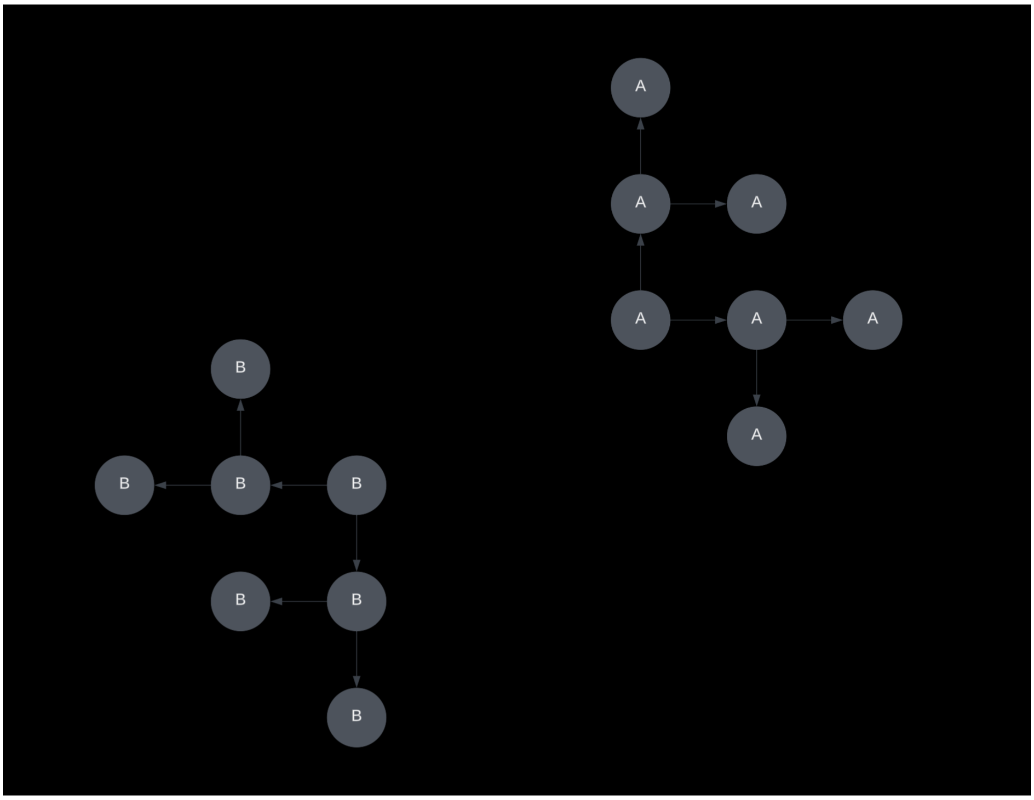


Figure 7.2: Small network showing an disconnected network

Our approach to improve interconnectedness was a brute force approach, where we collected first and second level references and citations. This approach was chosen after filtering out numerous proposed approaches, including a stochastic-based approach, where we randomly selected between 1-5 papers from the subset of references and citations. The stochastic-based approach would have had higher depth but would have suffered from a sparser network and possibly missing critical connections. Due to the exponential nature of collecting and constructing the network, the brute force is much more computationally expensive and very time-consuming due to API limitations, but the brute force approach will have much richer data, where it may find hidden connections to papers of a certain class as well as, span a wider more complete picture the citation patterns for each paper. Using the brute force approach, our final graph had **9843654 nodes** and **18130275 edges**.

7.4.1 Metrics

- **Degree :** The number of connections or incidents on a node.
  In-Degree : The number of edges pointing towards the node
  Out-Degree : The number of edges pointing out of the node [41]

- **Betweeness Centrality :** It is a metric on how often a node lies on the shortest path between other nodes. This metric is described as having "a potential for control of communication". [42]

- **PageRank :** It is a stochastic iterative algorithm. This metric is described as capturing endorsement and influence quality of nodes.[43]

7.4.2 Results

(LR)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.84	0.74	0.79	0.80
1	0.77	0.85	0.81	0.80
ROC-AUC Score: 0.8				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.82	0.80	0.81	0.81
1	0.80	0.83	0.81	0.81
ROC-AUC Score: 0.81				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.83	0.76	0.80	0.81
1	0.78	0.85	0.81	0.81
ROC-AUC Score: 0.81				

Table 7.12: Cross validation results for citation network metrics with different ML models.

Analysis

Performance : All 3 models reveal to us a strong predictive power, where the models have an accuracy between 80%-81% and a ROC-AUC score between 0.8-0.81, with balanced classes in all 3 models. The model trained on citation network data outperforms our citation meta-data model, with the citation network model being the strongest nonproblematic model.
Interpretation : Our strong results are reflective of the conducted research, where results show that non-retracted papers are often influential and are well distributed within a network, compared with retracted papers. Additionally, retracted papers' networks are more hierarchical and have a strong correlation to centrality and degree measures [44]. Authors who already have one retracted paper will try to expand their network, but these networks will be weaker networks where co-authors will also be less productive [45].

7.5 Meta-Net

The information held within networks is clear, based on our analysis of citation networks. The idea of Meta-Net is to research, if we can harness that information held within a network but through constructing that network through meta-data instead of traditional means like citation or co-authorship.
First, the meta-data we will be using is the data that we deemed non-problematic in our analysis in the meta-data section and used in our general model, which are 'Abstract', 'Subject', 'Author Count'.

7.5.1 Construction

The idea will initially involve using a transformer, specifically the sentence transformer SBERT [46]. We will use this transformer to embed our abstracts, turning them into dense vectors. By using cosine similarity on the abstract embedding, we can use k nearest neighbors to compute the **n** nearest papers to that paper. We are gonna be using this as our neighborhood for that paper or node.7.3

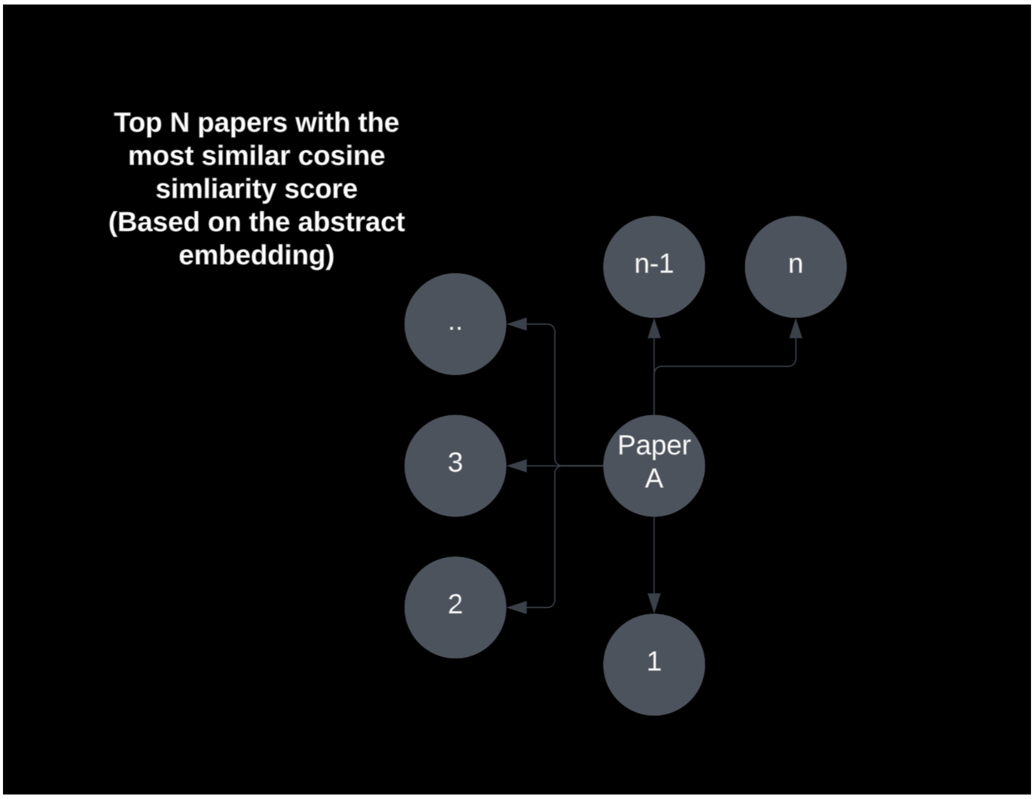

Top N papers with the most similar cosine simliarity score (Based on the abstract embedding)
n-1
n
..
3
Paper A
2
1

Figure 7.3: Example of a neighborhood for a node

From our neighborhood, we are gonna add weighted edges based on our selected metrics *'Abstract', 'Subject', 'Author Count'.* s.t for abstract, it will be a weight based on the cosine similarity between the 2 nodes, and subject and author count will be treated as a binary, meaning if they are the same, their respective weight will be added.

$$\text{where } w_a = \text{abstract weight}, \quad w_s = \text{subject weight}, \quad w_a c = \text{author count weight}$$

$$EdgeWeight(i, j) = cosineSim(i, j) {*} w_a + isSameSubject(i, j) {*} w_s + isSameAuthorCount(i, j) {*} w_a c$$

Once we have constructed the undirected weighted graph, we will follow the same pipeline as our citation network models.

7.5.2 Results

(LR)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.54	0.42	0.47	0.53
1	0.52	0.64	0.58	0.53
ROC-AUC Score: 0.53				
(RF)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.60	0.56	0.58	0.59
1	0.59	0.62	0.60	0.59
ROC-AUC Score: 0.6				
(GB)				
	Precision	**Recall**	**F1-Score**	**Accuracy**
0	0.59	0.47	0.52	0.57
1	0.56	0.67	0.61	0.57
ROC-AUC Score: 0.57				

Table 7.13: Cross validation results for Meta-Net network metrics with different ML models.

Analysis

Performance : All 3 models perform poorly, with the best performing model being RF with an accuracy of 59% and a ROC-AUC score of 0.59, followed by GB and then LR.

Interpretation : There is a signal, even if weak, that is present in the Meta-Net. This huge drop off in performance, compared to citation networks, is possibly due to our implementation of using abstract embedding and calculating the nearest neighbors with cosine similarity. Not creating semantically meaningful neighbors that are present in citation networks. Our weighted edges also are not carrying forward much semantic meaning, resulting in our poor performance.

Chapter 8

Discussion and Future Work

One of the major difficulties in our research was accessing data; due to not having Web of Science and relying on 3 separate databases to provide information on meta-data, citations, and abstracts, we were unable to collect sufficient data where we possessed all 3 of these. As a result, we were unable to aggregate our most powerful models to show us what the most predictive model in our research would be. In future work conducted, having access to more complete data would show us what the best performing model would look like.

Our research evolved as we progressed. The various ways we experimented did show us that there was a strong signal present within the methods we used. Our Meta-Net approach did not yield strong results, trying to artificially create a network based on meta-data, and the similarity of the the abstracts proved to not replicate the same level of information present in a citation network. This could be due to the innate network structure present in citations/references that yielded strong semantic meaning. Our approach of using abstract similarity to create the neighborhoods and the network did not contain that level of semantic meaning. However, there is more work that could be done with regards to artificially creating a network; our approach of using a transformer for embedding abstract and then using cosine similarity created broad, meaningless links; instead of using this approach to create neighborhoods, we could possibly utilize the research conducted by Cohan [47], where they introduce SPECTER which is a new method to generate document embedding that outperforms traditional transformer based embedding techniques like we used, when it came to capturing accurate representation of the documents. The idea of constructing artificial networks, we think, is an area where there is a possibility of making a breakthrough; even our network still did possess some predictive power, using more semantically meaningful or accurate embedding approaches on the abstract to create neighborhoods or using a completely different approach to neighborhood construction and using more metrics when it came to creating the weighted edges, is an area of research that needs to be utilized in order to make better prediction models, that can be coupled with existing methods that we showed that worked, to create a model with high predictive power that can be utilized in early detection.

With regards to citation network, our approach did show us that there is strong predictive power present within networks, but there is more work that could be done; currently, our citation network is static; using temporal citation patterns could be an avenue of research, with retracted papers rapidly declining in the number of citations received post retractions and older well respected papers receiving a steady number of citations, with research indicating that temporal features outperform static network [48]. The research conducted by Chang Zong [49] introduced an exciting framework called CTPIR, which can capture the influence and citation trajectory of new and existing papers by being able to predict future citation trajectories and, thus, citation patterns. We can couple this with a citation network or use it standalone to detect anomalies, which can be used to identify problematic papers early.

Our research shows us that utilizing machine learning for paper retraction is viable and that many of the techniques we discussed in this paper can be used as an initial screen mechanism that can aid in early detection; with a richer dataset and researching some of the techniques is discussed in the future work, there is a possibility that an aggregated model, combining all the successful techniques could provide a very strong model that can be used a part of the paper review process or other areas in which we need to check for the legitimacy of a paper, to aid in early detection.

Chapter 9

Project Management

This chapter contains the plan of the project in the form of GANTT charts and additionally contains a Risk Assessment that identifies potential risks and mitigation strategies.

9.1 Methodology

The research project was managed using an approach inspired by traditional software engineering management techniques. Initially, I tried to implement an agile-style approach by introducing deliverables. However, the area I was researching was changing as I learned more about the topic, meaning I had to implement flexible, broad deliverables; our final GANTT chart is quite different from our original one. I had a broad, high-level plan to keep track of the project, which I turned into deliverables, with deadlines throughout this project.

9.2 GANTT Charts

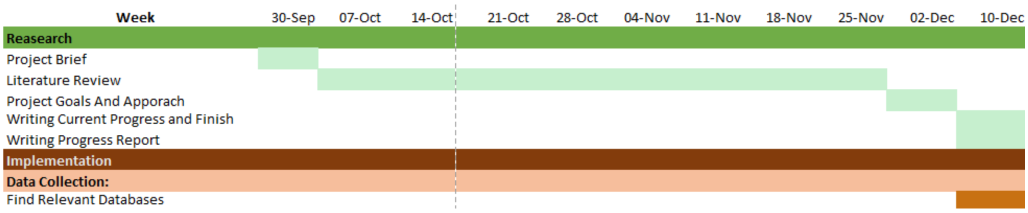


Figure 9.1: GANTT chart for Phase 1 of the Project

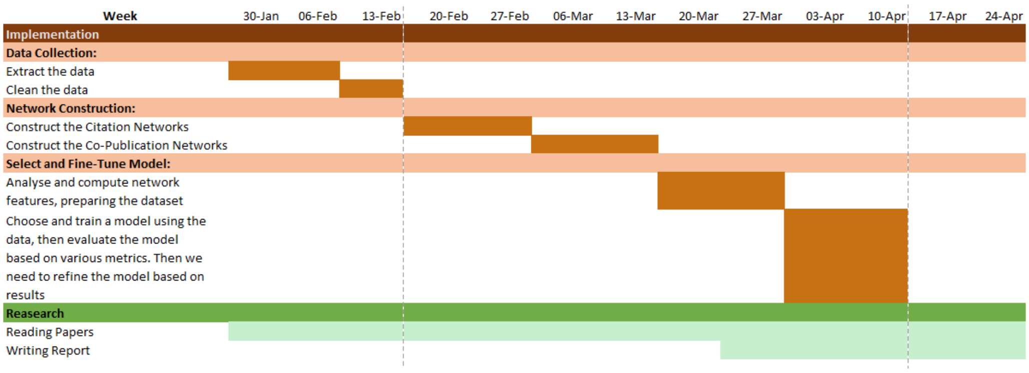


Figure 9.2: GANTT chart for Phase 2 of the Project (PREDICTION)

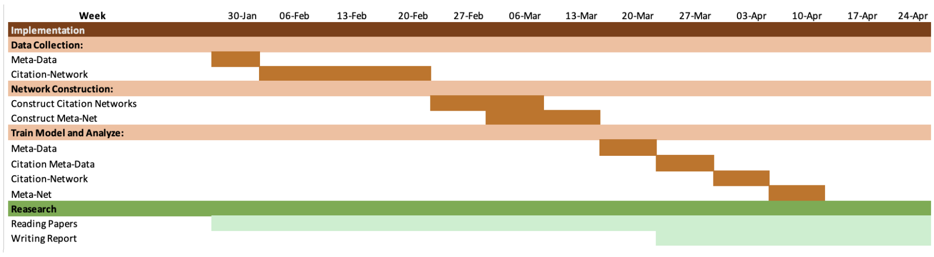
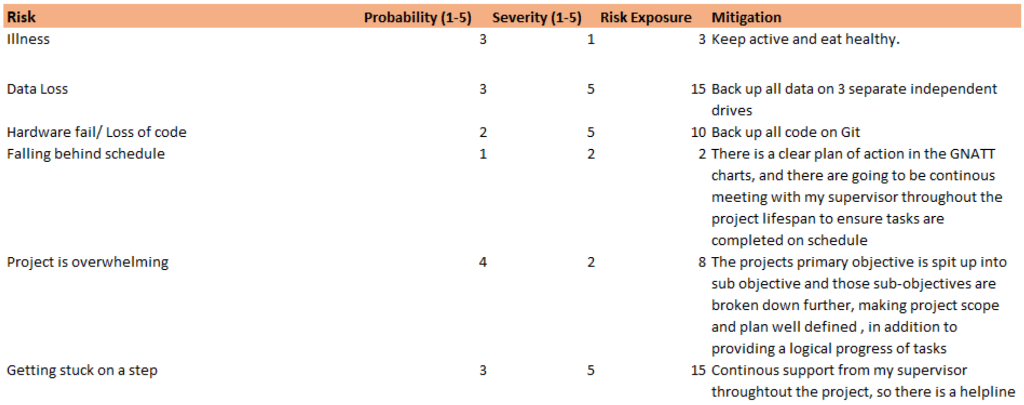


Figure 9.3: GANTT chart for Phase 2 of the Project (ACTUAL)

9.3 Risk Assessments



Risk	Probability (1-5)	Severity (1-5)	Risk Exposure	Mitigation
Illness	3	1	3	Keep active and eat healthy.
Data Loss	3	5	15	Back up all data on 3 separate independent drives
Hardware fail/ Loss of code	2	5	10	Back up all code on Git
Falling behind schedule	1	2	2	There is a clear plan of action in the GNATT charts, and there are going to be continous meeting with my supervisor throughout the project lifespan to ensure tasks are completed on schedule
Project is overwhelming	4	2	8	The projects primary objective is spit up into sub objective and those sub-objectives are broken down further, making project scope and plan well defined , in addition to providing a logical progress of tasks
Getting stuck on a step	3	5	15	Continous support from my supervisor throughtout the project, so there is a helpline

Figure 9.4: Risk Assessment for the Project

9.4 Technologies

Tools	**Description**
Python	Programming language for this project
Zotero	Reference manager to keep all research papers orgainised
Latex	Editor for writing report
Microsoft Teams	Software used to communicate with my supervisor
GitHub	Project management
Lucid Chart	Visuals and Diagrams

Table 9.1: Technologies and Descriptions

Bibliography

[1] R. G. Steen, A. Casadevall, and F. C. Fang, "Why Has the Number of Scientific Retractions Increased?" *PLoS ONE*, vol. 8, no. 7, G. E. Derrick, Ed., e68397, Jul. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0068397. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0068397 (visited on 10/26/2024).

[2] C. Wilcox, *ScienceAdviser: Scientists are publishing too many papers—and that's bad for science*, Nov. 2023. DOI: 10.1126/science.adm9969. [Online]. Available: https://www.science.org/content/article/scienceadviser-scientists-are-publishing-too-many-papers-and-s-bad-science (visited on 10/26/2024).

[3] S. Fortunato, C. T. Bergstrom, K. Börner, *et al.*, "Science of science," *Science*, vol. 359, no. 6379, eaao0185, Mar. 2018, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aao0185. [Online]. Available: https://www.science.org/doi/10.1126/science.aao0185 (visited on 10/26/2024).

[4] R. Van Noorden, "More than 10,000 research papers were retracted in 2023 — a new record," *Nature*, vol. 624, no. 7992, pp. 479–481, Dec. 2023, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/d41586-023-03974-8. [Online]. Available: https://www.nature.com/articles/d41586-023-03974-8 (visited on 10/26/2024).

[5] M. Kovacs, M. A. Varga, D. Dominik, R. Poldrack, and B. Aczel, *Opening the black box of article retractions: Exploring the causes and consequences of data management errors*, Jun. 2024. DOI: 10.31222/osf.io/5t4xg. [Online]. Available: https://osf.io/5t4xg (visited on 11/13/2024).

[6] I. Campos-Varela and A. Ruano-Raviña, "Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors," *Gaceta Sanitaria*, vol. 33, no. 4, pp. 356–360, Jul. 2019, ISSN: 02139111. DOI: 10.1016/j.gaceta.2018.01.009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0213911118300724 (visited on 10/26/2024).

[7] H. Else, “Biomedical paper retractions have quadrupled in 20 years — why?” en, *Nature*, vol. 630, no. 8016, pp. 280–281, Jun. 2024, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/d41586-024-01609-0. [Online]. Available: https://www.nature.com/articles/d41586-024-01609-0 (visited on 11/14/2024).

[8] C. Lievore, P. Rubbo, C. B. Dos Santos, C. T. Picinin, and L. A. Pilatti, “Research ethics: A profile of retractions from world class universities,” en, *Scientometrics*, vol. 126, no. 8, pp. 6871–6889, Aug. 2021, ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-021-03987-y. [Online]. Available: https://link.springer.com/10.1007/s11192-021-03987-y (visited on 11/14/2024).

[9] J. Nash, *Paper Mils-The Dark Side of the Academic Publishing Industry*, en, MDPI, May 2022. [Online]. Available: https://blog.mdpi.com/2022/05/09/paper-mills/.

[10] C. Brundy and J. B. Thornton, “The paper mill crisis is a five-alarm fire for science: What can librarians do about it?” en, *Insights the UKSG journal*, vol. 37, p. 11, Jul. 2024, ISSN: 2048-7754. DOI: 10.1629/uksg.659. [Online]. Available: http://insights.uksg.org/articles/10.1629/uksg.659/ (visited on 11/13/2024).

[11] R. Van Noorden, “How big is science’s fake-paper problem?” en, *Nature*, vol. 623, no. 7987, pp. 466–467, Nov. 2023, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/d41586-023-03464-x. [Online]. Available: https://www.nature.com/articles/d41586-023-03464-x (visited on 11/13/2024).

[12] P. Singh Deo and P. Hangsing, “Preserving Academic Integrity: Combating the Proliferation of Paper Mills in Scholarly Publishing,” en, *Journal of Electronic Resources in Medical Libraries*, vol. 21, no. 3, pp. 117–124, Jul. 2024, ISSN: 1542-4065, 1542-4073. DOI: 10.1080/15424065.2024.2388039. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15424065.2024.2388039 (visited on 11/13/2024).

[13] S. F. Lu, G. Z. Jin, B. Uzzi, and B. Jones, “The Retraction Penalty: Evidence from the Web of Science,” *Scientific Reports*, vol. 3, no. 1, p. 3146, Nov. 2013, ISSN: 2045-2322. DOI: 10.1038/srep03146. [Online]. Available: https://www.nature.com/articles/srep03146 (visited on 10/26/2024).

[14] S. A. Memon, K. Makovi, and B. AlShebli, *Characterizing the effect of retractions on scientific careers*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2306.06710. [Online]. Available: https://arxiv.org/abs/2306.06710 (visited on 11/14/2024).

[15] The Lancet, "Rethinking research and generative artificial intelligence," en, *The Lancet*, vol. 404, no. 10447, p. 1, Jul. 2024, ISSN: 01406736. DOI: 10.1016/S0140-6736(24)01394-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0140673624013941 (visited on 11/13/2024).

[16] H. Peng, D. M. Romero, and E.-Á. Horvát, "Dynamics of Cross-Platform Attention to Retracted Papers," *Proceedings of the National Academy of Sciences*, vol. 119, no. 25, e2119086119, Jun. 2022, arXiv:2110.07798 [cs], ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2119086119. [Online]. Available: http://arxiv.org/abs/2110.07798 (visited on 11/14/2024).

[17] S. Fahimifar, A. Ghorbi, and M. Ausloos, *Are We Standing on Unreliable Shoulders? The Effect of Retracted Papers Citations on Previous and Subsequent Published Papers: A Study of the Web of Science Database*, arXiv:2201.09090 [cs], Jan. 2022. [Online]. Available: http://arxiv.org/abs/2201.09090 (visited on 11/14/2024).

[18] J. Wittau and R. Seifert, "How to fight fake papers: A review on important information sources and steps towards solution of the problem," en, *Naunyn-Schmiedeberg's Archives of Pharmacology*, Jul. 2024, ISSN: 0028-1298, 1432-1912. DOI: 10.1007/s00210-024-03272-8. [Online]. Available: https://link.springer.com/10.1007/s00210-024-03272-8 (visited on 10/26/2024).

[19] T. Miyakawa, "No raw data, no science: Another possible source of the reproducibility crisis," en, *Molecular Brain*, vol. 13, no. 1, 24, s13041–020–0552–2, Dec. 2020, ISSN: 1756-6606. DOI: 10.1186/s13041-020-0552-2. [Online]. Available: https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-020-0552-2 (visited on 11/13/2024).

[20] J. A. Byrne and J. Christopher, "Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills?" en, *FEBS Letters*, vol. 594, no. 4, pp. 583–589, Feb. 2020, ISSN: 0014-5793, 1873-3468. DOI: 10.1002/1873-3468.13747. [Online]. Available: https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.13747 (visited on 11/13/2024).

[21] Wiley, "Wiley announces pilot of new AI-powered Papermill Detection service," en, *Wiley*, Mar. 2024. [Online]. Available: https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-announces-pilot-of-new-AI-powered-Papermill-Detection-service/default.aspx (visited on 11/14/2024).

[22] D. Bishop and A. Abalkina, “Paper mills: A novel form of publishing malpractice affecting psychology,” *Meta-Psychology*, vol. 7, Dec. 2023, ISSN: 2003-2714. DOI: 10.15626/MP.2022.3422. [Online]. Available: https://open.lnu.se/index.php/metapsychology/article/view/3422 (visited on 11/13/2024).

[23] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, and D. Hua, “Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools,” en, *Cell Reports Physical Science*, vol. 4, no. 6, p. 101 426, Jun. 2023, ISSN: 26663864. DOI: 10.1016/j.xcrp.2023.101426. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S266638642300200X (visited on 11/13/2024).

[24] Springer, “Springer Nature unveils two new AI tools to protect research integrity,” en, Jun. 2024. [Online]. Available: https://group.springernature.com/fr/group/media/press-releases/new-research-integrity-tools-using-ai/27200740 (visited on 11/13/2024).

[25] J. Mingers and L. Leydesdorff, “A review of theory and practice in scientometrics,” en, *European Journal of Operational Research*, vol. 246, no. 1, pp. 1–19, Oct. 2015, ISSN: 03772217. DOI: 10.1016/j.ejor.2015.04.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S037722171500274X (visited on 11/18/2024).

[26] H. R. Saeidnia, E. Hosseini, S. Abdoli, and M. Ausloos, “Unleashing the Power of AI. A Systematic Review of Cutting-Edge Techniques in AI-Enhanced Scientometrics, Webometrics, and Bibliometrics,” *Library Hi Tech*, Feb. 2024, arXiv:2403.18838 [cs], ISSN: 0737-8831. DOI: 10.1108/LHT-10-2023-0514. [Online]. Available: http://arxiv.org/abs/2403.18838 (visited on 11/18/2024).

[27] S. R. Goldberg, H. Anthony, and T. S. Evans, “Modelling citation networks,” en, *Scientometrics*, vol. 105, no. 3, pp. 1577–1604, Dec. 2015, ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-015-1737-9. [Online]. Available: http://link.springer.com/10.1007/s11192-015-1737-9 (visited on 11/13/2024).

[28] S. Kusumastuti, M. G. Derks, S. Tellier, *et al.*, “Successful ageing: A study of the literature using citation network analysis,” en, *Maturitas*, vol. 93, pp. 4–12, Nov. 2016, ISSN: 03785122. DOI: 10.1016/j.maturitas.2016.04.010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378512216300822 (visited on 11/14/2024).

[29] S. A. Greenberg, “How citation distortions create unfounded authority: Analysis of a citation network,” en, *BMJ*, vol. 339, no. jul20 3, b2680–b2680, Jul. 2009, ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.b2680. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.b2680 (visited on 11/14/2024).

[30] B. D. P. F. E. Fonseca, R. B. Sampaio, M. V. D. A. Fonseca, and F. Zicker, "Co-authorship network analysis in health research: Method and potential use," en, *Health Research Policy and Systems*, vol. 14, no. 1, p. 34, Dec. 2016, ISSN: 1478-4505. DOI: 10.1186/s12961-016-0104-5. [Online]. Available: http://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-016-0104-5 (visited on 11/13/2024).

[31] M. Ullah, A. Shahid, I. U. Din, *et al.*, "Analyzing Interdisciplinary Research Using Co-Authorship Networks," en, *Complexity*, vol. 2022, no. 1, S. Sarfraz, Ed., p. 2 524 491, Jan. 2022, ISSN: 1076-2787, 1099-0526. DOI: 10.1155/2022/2524491. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2022/2524491 (visited on 11/18/2024).

[32] S. A. Modukuri, "Understanding and Predicting Retractions of Published Works," M.S. thesis, Penn State University, Jun. 2021. [Online]. Available: https://etda.libraries.psu.edu/catalog/23957svm6277.

[33] H. Finsberg, *Pygscholar*, Aug. 2024. [Online]. Available: https://pypi.org/project/pygscholar/.

[34] S. Cholewiak, P. Ipeirotis, V. Silva, and A. Kannawadi, *Scholarly*, Jan. 2023. [Online]. Available: https://pypi.org/project/scholarly/.

[35] *Entrez Programming Utilities*, Bethesda (MD), 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25501/.

[36] *OpenCitations*. [Online]. Available: https://opencitations.net/.

[37] *Sci-Kit Learn Feature Extraction*, Documentation for the scikit-learn library. [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781 [cs], Sep. 2013. DOI: 10.48550/arXiv.1301.3781. [Online]. Available: http://arxiv.org/abs/1301.3781 (visited on 03/16/2025).

[39] M. L. Grieneisen and M. Zhang, "A Comprehensive Survey of Retracted Articles from the Scholarly Literature," en, *PLoS ONE*, vol. 7, no. 10, E. Von Elm, Ed., e44118, Oct. 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0044118. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0044118 (visited on 04/18/2025).

[40] D. Fanelli, "Why Growing Retractions Are (Mostly) a Good Sign," en, *PLoS Medicine*, vol. 10, no. 12, e1001563, Dec. 2013, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001563. [Online]. Available: https://dx.plos.org/10.1371/journal.pmed.1001563 (visited on 04/20/2025).

[41] L. C. Freeman, “Centrality in social networks conceptual clarification,” en, *Social Networks*, vol. 1, no. 3, pp. 215–239, Jan. 1978, ISSN: 03788733. DOI: 10.1016/0378-8733(78)90021-7. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0378873378900217 (visited on 04/25/2025).

[42] L. C. Freeman, “A Set of Measures of Centrality Based on Betweenness,” *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977, ISSN: 00380431. DOI: 10.2307/3033543. [Online]. Available: https://www.jstor.org/stable/3033543?origin=crossref (visited on 04/25/2025).

[43] Sergey Brin and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, Elsevier, vol. 30, no. 1-7, pp. 107–117, 1998.

[44] K. Sharma, A. Sharma, J. Jose, V. Saini, R. Sobti, and Z. Uddin, *Exploring Structural Dynamics in Retracted and Non-Retracted Author’s Collaboration Networks: A Quantitative Analysis*, Version Number: 1, 2024. DOI: 10.48550/ARXIV.2411.17447. [Online]. Available: https://arxiv.org/abs/2411.17447 (visited on 04/27/2025).

[45] S. A. Memon, K. Makovi, and B. AlShebli, “Characterizing the effect of retractions on publishing careers,” en, *Nature Human Behaviour*, Apr. 2025, ISSN: 2397-3374. DOI: 10.1038/s41562-025-02154-0. [Online]. Available: https://www.nature.com/articles/s41562-025-02154-0 (visited on 04/27/2025).

[46] *SBert*. [Online]. Available: https://sbert.net/.

[47] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, “SPECTER: Document-level Representation Learning using Citation-informed Transformers,” en, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 2270–2282. DOI: 10.18653/v1/2020.acl-main.207. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.207 (visited on 04/28/2025).

[48] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman, “Time-Aware Ranking in Dynamic Citation Networks,” in *2011 IEEE 11th International Conference on Data Mining Workshops*, Vancouver, BC, Canada: IEEE, Dec. 2011, pp. 373–380, ISBN: 978-1-4673-0005-6 978-0-7695-4409-0. DOI: 10.1109/ICDMW.2011.183. [Online]. Available: http://ieeexplore.ieee.org/document/6137404/ (visited on 04/28/2025).

[49] C. Zong, Y. Zhuang, W. Lu, J. Shao, and S. Tang, *Citation Trajectory Prediction via Publication Influence Representation Using Temporal Knowledge

Graph, arXiv:2210.00450 [cs], Oct. 2022. DOI: 10.48550/arXiv.2210.00450. [Online]. Available: http://arxiv.org/abs/2210.00450 (visited on 04/28/2025).

Appendix A

Project Brief

A.1 Problem

The number of research papers being published is increasing rapidly with in 2016 about 1.92 million papers being indexed by Scopus and Web of Science publication databases and that number growing to about 2.82 million papers in 2022 [2]. A new worry arises, science itself is a constantly evolving multi-scale network [3] , with a large number of papers being published each year, there is a growing concern that some of these papers are based on fraudulent foundations by citing papers that have not yet been retracted, with the number of retractions per year rapidly increasing [1] and in 2023 the number of papers being retracted being around 10000 [4] the concern is not without a strong base. The problem we are tackling with this research paper is trying to decrease the time taken for the retraction of fraudulent research papers so that we can mitigate their damage to the research world. The number of research papers that have been retracted in recent times has increased drastically and the mean time to retract a fraudulent paper is 33.81 months (±35.63 SD), [1]. This means that in the time-to-retract period, propagation of errors could occur where subsequent research could be built on fraudulent findings, causing large amounts of effort to be wasted on resolving the damage caused by fraudulent publications. The process of scrubbing the literature to remove their influence is very lengthy [1], speeding up the process of retraction will decrease the influence a fraudulent paper has on the research space. Using this model as an initial detection mechanism would decrease the sample space of papers that need to be investigated for fraud or other types of errors decreasing the mean time to retract and helping preserve academic integrity.

A.2 Literature Review

Previous research in this field has tried a similar technique to this research paper. A study conducted by Sai Ajay Modukuri [32], attempts to solve this problem by making a machine learning model trained on metadata features and full-text features. They analysed the metadata and the full text and used various techniques to extract features from these two types of data and then used this to train their model. Their results show that using the metadata features alone gave an F1 score of 67% and using full-text features alone gave an F1 score of 63%. Combining both metadata and full-text features gave a combined F1 score of 71%. While using just the metadata and text gave promising results, their work leaves areas unexplored especially the relationship between papers and also the relationship between co-authors. They used citation metrics and semantic scholar to provide the intent behind each citation and used institutional rankings based on Times Higher Education to give the lead author a score but they failed to capture the patterns and networks created by these research papers. As stated in [3] the entirety of the science is a complex network, in this research project we plan to construct some of these networks and then extract some relevant network features to train our model exploiting the relationships between the papers and authors.

A.3 Goals

The goal of this project is to explore the possibility of utilizing machine learning techniques, data features, citation, and social co-publication network features to find out if we can build a model that can accurately predict the likelihood of a research paper being retracted. Thus allowing for the early detection of possible retractable papers, decreasing the mean "time-to-retract"[1].

A.4 Scope

This project's scope would be the entire development of a model that would predict the likelihood of a research paper being retracted. The machine learning techniques utilized in this research project would be constrained to only supervised learning techniques and exclude other machine learning techniques. In addition, we will explore the relationships and patterns of citation and social co-publication networks and use these metrics as data features. The dataset is restricted to the available information on retracted papers on the major publication databases at the time of this project, there may be papers that are yet to be retracted but are fraudulent [1].

Appendix A

Archive

In the archive, there is the source code of our experiments which is in .txt format, we had to convert them into .txt to save space as our original submission was too large.

There is all the datasets we gathered and used, but we could not include our citation network as it was too large.

There is also a zipped file containing our latex source code