GRADUATE PROGRAMS INSTITUTE

MACHINE LEARNING BASED EVALUATION OF SCIENTIFIC RETRACTIONS

Ehtisham Kiyani

MASTER'S THESIS

İstanbul, February 2022

MACHINE LEARNING BASED EVALUATION OF SCIENTIFIC RETRACTIONS

Ehtisham KIYANI

MASTER'S THESIS

COMPUTER ENGINEERING DEPARTMENT

COMPUTER ENGINEERING MASTER'S PROGRAM WITH THESIS

MASTER'S THESIS ADVISOR

Asst. Prof. Dr. Özlem Feyza Erkan

MASTER'S THESIS JURY MEMBERS

Dr. Burcin Kulahcioglu

Dr. Selcuk Kiran

PREFACE

First of all, I would like to thank my thesis supervisor, Dr. Ö. Feyza Erkan. She was always there for me when I needed help and advice, and she was always happy to answer my many questions I had about researching and writing my thesis. I have to say that my first research would not have turned out this way without her help.

I dedicate my thesis to my parents because of their support and encouragement, I was able to do such a fantastic work. Finally, I'd want to express my gratitude to my siblings and friends because without their continuous love and support, writing my thesis would not have been easy.

CONTENTS

ABSTRACT

MACHINE LEARNING BASED EVALUATION OF SCIENTIFIC RETRACTIONS

Researchers produce a tremendous number of publications based on their findings every single day. The frequency of retracted papers published in scientific journals has sharply increased during the last 20 years, outpacing the gradual growth in the overall number of publications. Retraction is a mechanism for correcting the literature by informing the readers about papers that include information or data that is seriously faulty or inaccurate that makes the researcher impossible to rely on the findings and conclusions. Questionable research may propagate disinformation and put people in danger, particularly in the medical literature. As of May 2020, the Retraction Watch Database shows 23,000 retracted papers and 800,000 articles that directly cite them. In this thesis, we address the problem of determining which publications are close to retraction based on their association with the retracted articles and when we need to reconsider papers that cite them. We proposed a method in which we can predict the paper that is close to retraction using different machine learning algorithms like Logistic Regression, Naïve Bayes, SVM and Random Forest. Moreover, we introduce an annotation scheme for citation functions with the help of a subject matter expert. According to this scheme, we have found that there are 221 papers out of 664 that need reconsideration. The trained model achieved approximately 82% accuracy in identifying possible retractions. In this way, we can save our resources to check some percentage of potentially harmful literature instead of reviewing the complete database. In the context of healthcare, several domains and categories may be taken into consideration for future work to form new data. More insights and trends may be discovered from the collected data. The development of deep learning models for feature engineering would also be useful.

Keywords: citation context analysis, retraction, annotation scheme, machine learning

ÖZET

BİLİMSEL ÇALIŞMALARDAKİ GERİ ÇEKİLMELERİN MAKİNE ÖĞRENİMİ TABANLI DEĞERLENDİRİLMESİ

Araştırmacılar her gün bulgularına dayanarak muazzam sayıda yayın üretmektediler. Bilimsel dergilerde yayınlanan geri çekilmiş makalelerin sıklığı, son 20 yılda, genel yayın sayısındaki kademeli büyümeyi geride bırakarak keskin bir şekilde artmıştır. Yayınların geri çekilmesi, araştırmacının bulgulara ve sonuçlara güvenmesini imkansız kılan ciddi şekilde hatalı veya yanlış bilgi veya veriler içeren makaleler hakkında okuyucuları bilgilendirerek literatürü düzeltmeye yarayan bir mekanizmadır. Şüpheli araştırmalar, özellikle tıp literatüründe, dezenformasyon yayabilir ve insanları tehlikeye atabilir. Mayıs 2020 itibariyle, Geri Çekme İzleme Veritabanı 23.000 geri çekilmiş makaleyi ve bunlardan doğrudan alıntı yapan 800.000 makaleyi göstermektedir. Bu tezde, geri çekilen makalelerle olan ilişkilerine dayanarak hangi yayınların geri çekilmeye yakın olduğunu belirleme ve bunlardan alıntı yapan makalelerin ne zaman yeniden gözden geçirilmesi gerektiği incelenmektedir. Lojistik Regresyon, Naïve Bayes, Destek Karar Makineleri ve Rastgele Orman gibi farklı makine öğrenimi algoritmalarını kullanarak geri çekilme olasılığı yüksek olan makalenin tahmin edilebileceği bir yöntem önerilmiştir. Ayrıca, konu uzmanının yardımıyla alıntı işlevleri için bir ek açıklama şeması sunulmuştur. Bu şemaya göre, 664 makaleden 221'inin yeniden değerlendirilmesi gerektiği belirlenmiştir. Eğitilmiş model, olası geri çekmeleri belirlemede yaklaşık olarak %82 doğruluk elde etmiştir. Böylece, tüm veritabanını gözden geçirmek yerine potansiyel olarak zararlı literatürün bir kısmını kontrol etmek için kaynaklarımızı koruyabiliriz. Sağlık hizmetleri bağlamında, yeni veriler oluşturmak için gelecekteki çalışmalar için birkaç alan ve kategori dikkate alınabilir. Toplanan verilerden daha fazla bilgi ve eğilim keşfedilebilir. Özellik mühendisliği için derin öğrenme modellerinin geliştirilmesi de yararlı olacaktır.

Anahtar Kelimeler: Alıntı Bağlamı Analizi, Geri Çekme, Ek Açıklama Şeması, Makine Öğrenimi

Tarih:

ABBREVIATIONS

DL : Digital Libraries

NLTK : Natural Language Toolkit

NLP : Natural Language Processing

SA : Sentiment Analysis

SVM : Support Vector Machine

BOW : Bag of words

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Researchers publish an incredible number of publications every day (Usman, Mustafa and Afzal, 2021). Researchers' dedication to the profession and desire to discover the truth are betrayed when they lose their moral sense and engage in unethical or fraudulent behavior in the name of science. In the scientific community, misconduct may take many different forms, including fabricating data and altering results, copying text or images, duplicating image, and interfering with the peer review process.

A research publication is retracted when research misconduct is discovered in it (Jan, Bano and Mehraj, 2018). According to Greitemeyer (2014) a retraction is the worst form of post-publication punishment in the scientific community. The formal statement that a study is so faulty it must be withdrawn from the literature." An amendment notice or, in more serious circumstances, an "expression of concern" may be provided before a piece of writing with a little mistake or inaccurate information is retracted. The growing number of scientific publications being retracted is a concerning trend (Grieneisen and Zhang, 2012). The frequency of retracted papers appearing in scientific journals has sharply increased during the last 20 years, outpacing the gradual growth in the overall number of publications (Fang, Steen and Casadevall, 2012). Retractions averaged 315 each year from 2000 to 2004, and from 2016 to 2020, they increased by 1833, or more than 500% (Retraction Watch, 2021). The majority of retractions are the result of wrongdoing, such as data fabrication, falsification, or plagiarism, duplicate publishing, fraudulent peer review, or flagrant breaches of the rules governing human research (Fang and Steen, 2012). As of May 2020, the Retraction Watch Database has approximately 23,000 articles that have been retracted (Fu and Schneider, 2020). A retracted work is cited in over 800,000 articles (Fu and Schneider, 2020).

In the field of biomedicine, 94% of articles that have been retracted had at least one citation, with an average of 35 (Dinh et al., 2019). Furthermore, not all references to these papers are unfavorable; Wakefield's fraudulent paper asserting a connection between the MMR vaccination and autism had 94 positive citations (Suelzer et al., 2019)

1.1 MOTIVATION

Since it is believed by many research that referencing retracted articles spreads erroneous information, it is maybe acceptable if the author is aware of the paper's retraction status. Pre- and post-retraction citations are the most probable two scenarios. How could someone in the first situation (pre-retraction) be aware that the referenced publication would soon be retracted? In the second scenario (post-retraction), there may be two more situations. In the first scenario, the author still cites the retracted article while being aware of the status of the paper's retraction, which is acceptable. The second scenario is when the authors cite the retracted publication even if they are not aware that it has been retracted. Now, the author's reason for referencing the retracted publication differs depending on the specific research issue. There might be a variety of explanations, such as the author is just discussing the retracted article or the author's work depends on it. Acceptable if it's merely a debate and the conclusions don't depend on that particular study. Citations of questionable research spread misinformation and may be risky, especially in medical literature where questionable research puts patients at risk. Over 800,000 publications directly mention retracted papers, according to statistics. In the field of biomedicine, 94% of articles that have been retracted contain at least one reference, and the average number of citations is 35, which includes both positive and negative citations.

1.2 OBJECTIVE OF THE THESIS

Our goal in this thesis is to set an agenda for knowledge maintenance and predictions about the papers that are close to retraction based on machine learning techniques. We are not using the conventional data but working on the basis of ground truth by mapping the papers which are citing the retracted paper and got retracted with those papers which are citing the same retracted papers but not yet retracted. We have citation context and other metadata which helps us to flag those papers that are close to retraction. Moreover, with the help of subject matter expert an annotation scheme for citation functions has been proposed. We are labeling the text column as potentially harmful or not harmful because using this we are able to find those paper that need to be reconsider.

In this thesis, we will address the following questions:

1) How can we determine the papers that are close to retraction based on their association with the retracted articles?
2) Under which scenario it could be necessary to reconsider the paper that cite the retracted articles?

1.3 ORGANIZATION OF THE THESIS

The rest of the thesis is structured in the following manner:

Chapter 2 presents a comprehensive review on citation context analysis of retracted papers has been discussed.

Chapter 3 describes the methodology for the evaluation of scientific retractions. The steps of data preprocessing are also explained in detail.

Chapter 4 discusses the experimental results and analyzes the performance of the proposed models. Moreover, the statistical analysis results are also explained.

Chapter 5 summarizes the results that we achieved and suggests some future directions for the improvement of evaluation of scientific retractions.

2 LITERATURE REVIEW

In this thesis citations following retractions were analyzed. The literature shows that not all retracted article citations are positive; rather, a questionable citation encourages researchers to be more careful when citing in future work and strengthens the validity of their research by drawing attention to instances of intentional or unintentional scientific fraud. We are dealing with the field of medicine, thus citing questionable research may spread misinformation and be dangerous, particularly in medical where questionable research puts patients at risk. In the discipline of biomedicine, 94% of retractions include at least one reference, and the average retractions have 35 citations, including both positive and negative citations.

2.1 CITATION CONTEXT ANALYSIS

Citation context analysis is a thorough and rigorous method of literature evaluation that goes beyond conventional narrative and systematic reviews to more fully comprehend the significance of prominent authors and foundational works (Anderson and Lemken, 2020). Authors' motives, objectives and intentions for citing works have been examined using citation context analysis (Jha et al., 2017). It has been investigated where the citations appear in full-text publications as well as the text that appears around the citations (Bertin et al., 2016). For instance, Bertin et al. (2016) found that the structure of the citing article is connected to the number of citations seen in full text (i.e., the parts designated as IMRaD, which include the introduction, methodology, results or findings, and discussion). The authors have stated that the introduction and the discussion parts are often included the most citations.

Studies of the text around citations have provided insight into the semantics of citations and the motivations behind citations. For instance, negative citations of articles are uncommon (Teufel, Siddharthan and Tidhar, 2006). Citations may also be used to demonstrate the related work, comparison, and corroboration, among other purposes (Hernandez-Alvarez, Soriano, and Martínez-Barco, 2017). Garfield (1965 16) initially presented a system with 15 reasons for citing publications. In 43 open-access biomedical journals, 1,710 citation contexts have been categorized by (Agarwal, Choubey and Yu, 2010) into eight different groups: background, explanation, contemporary, contrast, of results, evaluation, similarity/consistency, method, and the modality.

On the other hand, the use of citation contexts mentioning retractions is understudied. Existing studies simply looked at whether the retractions were noted in the citation contexts; they didn't look into why. Van Der Vet and Nijveen (2016) have revealed that just two post retraction citations (3.5%) out of 57 citations issued after the year of retraction revealed knowledge of the retraction in the text of a single retracted publication published in Nature. (Schneider et al., 2020) found that just five citations (4.5%) out of the 112 post-retraction citations to a retracted publication referenced the retraction. Budd, Coble and Abritis, (2016) references to 265 publications in the MEDLINE database that were retracted were evaluated during the period between 2001 and 2005. Out of 4,917 citations, only 204 (4.15%) recognized the retraction (Budd, Coble and Abritis, 2016). The frequency and quantity of citations that acknowledge retractions as reported in earlier studies are shown in (Figure 2.1)

Paper	% citations acknowledging retraction	# citations acknowledging retraction in the citations included for analysis	# retracted papers included for citation analysis
Santos-d'Amorim, de Melo, and dos Santos (2021)	37.85	81/214 citations	One retracted Covid-19 paper
Piller (2021)	47.5	95/200 citations	Two retracted Covid-19 papers
Van der Walt, Willems et al. (2020)	21.05	4/19 sampled citations	33 retracted Covid-19 papers
Schneider et al. (2020)	4.5	5/112 citations	One retracted paper
Theis-Mahon and Bakker (2020)	5.4	37/685 citations	81 retracted dentistry papers
Yang et al. (2020)	16.03	21/131 citations	46 retracted psychological papers
Bolboaca, Buhai et al. (2019)	1.07	6/559 citations	54 retracted papers reporting a radiology- imaging diagnostic method
Hamilton (2019)	6.6	27/407 citations	47 retracted radiation oncology papers
Suelzer et al. (2019)	38.2 (2005-2010); 71.7 (2011-2018)	123/322 (2005-2010); 360/502 (2011-2018)	Wakefield's Lancet paper (partly retracted in 2004; fully retracted in 2010)
Budd et al. (2016)	4.15	204/4,917 citations	265 retracted papers in MEDLINE
van der Vet and Nijveen (2016)	3.5b	0/37 citations in 2014; 2/57 citations in 2015	A paper published in Nature and retracted in Feb. 2014.
Budd et al. (2011)	6	14/247 citations in the 2000 sample; 8/144 citations in the 2005 sample	1,112 retracted papers in PubMed
Neale et al. (2010)	2.8	17/603 citations stratified random sampled from 5,393 citations	102 papers affected by scientific misconduct
Budd et al. (1999)	6.4 (AIM); 7.7 (non-AIM)	19/299 citations from AIM journals; 123/1,594 citations from non-AIM journals	235 retracted papers in MEDLINE
Kochan and Budd (1992)	5.7	17/298 citations	John Darsee's papers
Pfeifer and Snodgrass (1990)	2.9	5/178 citations	82 retracted papers identified from journals in Index Medicus

Figure 2.1 Citations in the literature that acknowledge retraction

Retracted papers were still used in other papers, and the papers that used them rarely told their readers that the papers had been retracted. (Bolboacă et al., 2019) 5 of the 9 studies that came out in the last five years (2016–2021) said that less than 7% of citations talked about retractions. Interesting research on famous, well-known retractions (Piller, 2021) discovered increased citation rates for retractions.

According to certain research, each citation context's tone (positive, neutral, negative) was investigated (Bar-Ilan and Halevi, 2017). Bar-Ilan and Halevi, (2017) explained how fifteen retracted articles were still cited in 238 other sources after they had been removed.

They found that the retracted articles were often discussed positively, and that the retraction status of the referenced article was seldom addressed in subsequent works. Yang, Qi and Diao, (2020) Examined 131 post-retraction citations to 46 retracted psychology studies and discovered approximately 119 which means (90.84%) positive references. According to Theis-Mahon and Bakker, 2020, out of the 685 post-retraction citations to dental publications that were retracted, 475 (69.34%) were considered to be positive. One retracted Covid-19 publication had 214 post retraction citations, and (Santos-d'Amorim et al., 2021) examined them. The authors found 64 citations that were positive (30%), 81 that were negative (38%) and 69 that were neutral (32%). Studies that focused at citation contexts in more depth only looked at a single retracted paper and the articles that cited it (Van Der Vet and Nijveen, 2016). (Suelzer et al., 2019) Proposed the most fine-grained scheme for dividing citation circumstances into distinct groups. Wakefield's infamous Lancet study linking the MMR vaccine to autism was cited 1,153 times in the study. The authors used eight categories (positive, conceptual, assumptive, contrastive, perfunctory, methodological, negative, and persuasive) to mark up the references (Suelzer et al., 2019).

2.1.1 Citation to Retracted Papers

The context of the references to retracted articles was investigated by (Gabehart, 2005). Only 5 of the 137 referrals to publications in Medline's database, which spanned the years 1990 to 2000, were negative, and 30% of the citations to those papers were post-retraction.

Based on the post-retraction citation context, 15 retracted publications were researched in Elsevier (Bar-Ilan and Halevi, 2017). From the 238 articles that made reference to it, 198 were positive (83%), 28 were neutral (13%), and just 12 were negative (5%).

Redman et al. (2008) examined 315 papers that had been retracted from PubMed between the years 1995 and 2004. They discovered that these articles had been referenced a total of 3942 times prior to the retraction, but 4501 times after it.

This study specified a much greater incidence of retraction compared to the prior research on the retractions that was conducted by (Budd et al., 1998).

Secondly, there was a decline in the amount of time between the publishing and the retraction. The research showed that the studies that had a high number of citations before the retractions also had a high number of citations after the retractions, with the studies published in journals with a higher impact having an even greater number of citations after the retractions.

Charisse, Madlock Brown, and Eichmann, (2014) Classified retracted articles in a more thorough manner than the prior published work in order to study the lack of influence of retractions on the citation network. The Medline database was used to list all of the retraction announcements made between the years of 2003 and 2010, and then these announcements were sorted into nine categories. Cokol, Ozbay, and Esteban (2008) found a positive correlation between pre- and post-retraction citations. They also underlined that scientific error and misconduct are leading to more publication retractions (Wager and Williams, 2011).

Da Silva and Cimenti, (2017) discussed the issue of post-retracted citations of papers and followed up on other studies that found that citation patterns for articles remained mostly unchanged after retractions (Budd, Sievert and Scoville, 1999). Furthermore, the citation after the retraction of a scientific publication is a result of a lack of knowledge about the literature that has been retracted. The authors also stated that journal editors should retract or rectify the dubious literature.

Da Silva and Dobranszki, (2017) analyzed and compared the pre- and post-citation values of 10 of the most widely referenced retracted publications that were listed by the creator of retraction watch, Oransky, in the year 2015. The year and reason for the retraction were provided in these articles. Nonetheless, despite being retracted, the articles still to be cited in the respective disciplines in which they were published. It is conceivable that some researchers think that the technique, results, or conclusion remain to be legitimate despite the publications being retracted.

Budd et al. (1999) examined the causes for the retraction of 235 publications by MEDLINE, the Science Citation Index, and an abbreviated version of the Medicus using the Index. The results of the study clearly imply that even when an article's retractions are obvious, researchers will still use the content and cite it long after the publication has been removed.

Wager and Williams, (2011) investigated the factors that lead to the retraction of previously published papers by using a methodology that was similar to that of (BuddSievert, Schultz, and Scoville, 1999). They gathered all of the retracted articles from Medline between the years 2005 and 2008 and analyzed at a total of 312 articles. From the beginning of the 1980s to 2008, the number of retractions has increased by a factor of ten. This includes both errors and misconducts. They discovered that the rules of the journals are not standard, which means that some of the journals don’t reference the reasons for retractions and obviously fail to discern between dishonest mistake and wrongdoing.

Grieneisen and Zhang, (2012) examined at 42 of the most extensive bibliographic databases for important scholarly subjects as well as publisher websites in order to investigate the range of retracted publications and the features of those articles across the broad spectrum of academic fields. The findings indicated that the number of retractions that occurred as a consequence of apparent publishing misconduct (47%) was higher than the number that occurred as a result of apparent research misconduct (20%). The research concluded that the articles which are retracted can be found across the entire spectrum of academic fields, that the vast majority of retracted articles don’t contain erroneous data, Moreover, the majority of the articles that are retracted' authors have not been accused of doing unethical research.

In another work, the extensive post-retraction editing process and the use of various plagiarism detecting technologies, such as iThenticate, may be the true reason for the growing number of retraction announcements (Sheth and Thaker, 2014). The paper goes on to highlight that one of the consequences of scientific retractions is the incorrect citation of works that have been retracted. The current procedures of notice are shown to be inadequate by the continuous citation of material that has since been retracted.

Establishments need to foster a culture of liability and give writers with assistance in order to produce quality papers. Additional announcements of retractions have to be disseminated in both the printed and the digital media.

Nath, Markus, and Druss, (2006) conducted research on both intentional and unintentional retraction, with the emphasis being placed on inadvertent errors. Through the use of Medline, all of the retractions that occurred between the years 1982 and 2002 were gathered, and the concern for their withdrawal were divided into two categories: misconduct, which included cases of fabrication, plagiarism, and falsification; and the other type, which encompassed unintentional mistakes in data processing and sampling. Only 107 of the 395 papers were considered to be examples of unethical behavior in the scientific community, but the vast majority of the articles, 244, were considered to be examples of accidental errors. As a result of these findings, it was concluded that scientific misconduct was a far less common cause of retraction in the biomedical literature than inadvertent errors (which accounted for double the percentage).

On May 3, 2012, PubMed indexed 2047 biomedical and life-science research papers that were subsequently determined to have been retracted. Fang, Steen, and Casadevall (2012) analyzed these retraction notices in depth. If numbers are to be believed, just 21.3% of all notices of retraction were received in mistake. Misconduct, fraud, or even suspected fraud accounted for 43.4% of the remaining 67.4%, while duplicate publishing (14.2%) and plagiarism (9.8%) were the root causes of the other two cases. In addition, the study discovered that retractions due to fraud took longer to be issued than those due to other reasons, and that the majority of publications retracted for fraud are connected with high-impact journals and published in nations with strong research infrastructure. Similar to how nations with a deficient research infrastructure and ties to low-impact publications are more prone to create duplicate or plagiarized works of literature.

Decullier et al. (2013) examined at 244 retractions were published in Medline over the course of a year and analyzed their adherence to retraction standards, retractions' motivations, and their geographic distribution. In addition, our investigation uncovered the journal's impact factor and the presence or absence of a retraction notice in the primary article. Mistakes accounted for 28%, plagiarism for 20%, fraud for 14%, and overlap for 11%. Authors suggest keeping the original papers accessible but adding a notice that makes it clear the work has been retracted.

A research was undertaken by Moylyan and Kowalczuk, (2016) to investigate the reasons why papers are retracted from the BioMed Central Journals. Also, to determine whether or not retraction notifications agreed with the rules established by the Committee on Publication Ethics. According to the findings of the study, the majority of the retracted articles were due to misconduct (102 total articles, or 76%). The misconduct was committed in the form of a compromised peer review in 44 of the articles (33%), plagiarism in 22 of the articles (16%), data falsification or fabrication in 10 of the articles (7%), honest error in 17 of the articles (13%), and data fabrication in 10 of the articles (7%).

(Da Silva, 2016) argues that retractions are an indication of failure and makes some assumptions about the reasons why individuals reverse their claims. One of the causes that led to retractions was the cultural differences of the many authors who worked together on a joint project. Another issue consisting of duplications that were formed as a result of poor supervision or advisory bodies was also a contributor to the problem. Retractions may occur when authors and/or editors engage in citation stacking (Heneberg, 2016) to increase their own impact factor, or when major citation holders agree to reference each other in order to increase their impact factor. Citation manipulation by powerful citation holders who agree to cite one other to artificially inflate their impact factors is another potential source of retraction.

Shuai et al. (2017) investigated the effect that retractions have on the academic influence of an article. It was determined that there were around 2,659 papers that had been retracted and the primary reasons for the retractions were broken down into seven categories: plagiarism, scientific misconduct, rule violations, falsification or fabrication, mistakes, and other factors. It was discovered that scientific misconduct was responsible for more than fifty percent of all retraction cases, whereas inadvertent mistakes were responsible for around twenty-four percent. Plagiarism, fabrication, and falsification account for the most of the retractions published as a result of unethical behavior in the scientific community.

Kochan and Budd, (1992) conducted research on the impact and longevity of the works that Darsee and his coauthors had written and published. It has been shown that there have been a large number of apparent mistakes made in their published papers, and that fraud has been committed by inventing a considerable percentage of the research in the case study. Therefore, the findings of the research indicate that even after a significant period of time has passed, some of Darsee's publications are still being mentioned favorably after they have been retracted, and this continues to have a wider negative consequence, especially in the area of cardiology.

Davis (2012) used MEDLINE and Mendeley to determine how many retracted works are available to the public online and in academic libraries. The author came to the conclusion that a sizeable portion of the articles that were retracted could be found on a wide range of websites, such as advocacy websites, commercial websites, educational websites, and websites that are not associated with a publisher. The papers that had been published in the most prominent scientific journals were the ones that were located in the personal libraries the most often. The author made the suggestion that the advantages of having widespread access to scientific journals may be one of the reason for the spread of knowledge that is inaccurate, flawed, or untrustworthy. It's possible that using automated ways to keep readers up to speed on the latest developments can help cut down on the prevalence of errors in the scientific literature.

Using the PubMed database, Casadevall, Steen, and Fang (2014) identified the key causes of inaccuracy in a subset of retractions for which there was no evidence of wrongdoing. They did this by focusing on retractions for which there was no evidence of wrongdoing. In all, 439 articles were discovered and placed into one of the eight categories described below: irreproducibility, analytical error, laboratory error contamination, programming difficulties control issues, control problems, or other. After doing a study on the letters of retraction for these studies, it was shown that the most common reasons for retracting papers due to errors were mistakes made in the laboratory, errors made during the analysis, and results that could not be replicated.

Ven der Vet and Nijveen, (2016) investigated the issue of how mistakes are spread via citations by analyzing the full citation network of a frequently referenced manuscript known as the "Naryana paper." This paper was first published in 2012 in the journal Nature. In 2014, the study was withdrawn when many organizations voiced their dissatisfaction with their inability to replicate its results (Newton et al., 2014). On the other hand, the article was being mentioned directly, and the articles that cited this work in a more indirect manner did not include any evidence of the retracted findings.

Gasparyan, Ayyazyan, Akhazhanov, and Kitas, (2014) conducted an investigation into the errors and unethical behavior that were published in interdisciplinary and specialized biomedical publications. Duplicate publications or had been retracted were obtained with the use of PubMed's errata filters, and the most common categories of duplicate and retracted articles were noted. It was discovered that the duplicate and retracted articles correlated with country-based h-index values and the total number of publications across all countries.

Steen (2011) evaluated and analyzed 788 publications that were retracted from PubMed between the years 2000 and 2010. In a previous study (Nath, Marcus, and Druss, 2006), the researchers came to a conclusion that retracted publications had a relatively small number of coauthors. However, the findings of this research uncovered significant evidence to refute the conclusions of (Nath et al., 2006), particularly in light of the fact that the number of authors reported for each withdrawn paper ranged anywhere from two to twenty-six. A little less than seven percent of the articles that were retracted were published by a single author, while eighteen percent of the papers that were withdrawn had eight or more authors, and six of the papers that were retracted had more than twenty authors. According to the research, retracting a publication because the authors faked data implies a deliberate effort to mislead readers, and fraudulent retracted papers target high-impact journals.

In his study, Gewin (2017) addressed the time-span of retractions and advocated that the retraction of papers should be quick, transparent, and open in order to minimise bad effects on a researcher's career. Gewin's research was published in 2017. In order to rectify the body of work, he separated the approaches into four distinct groups: rectification (erratum), concern statement, partial retraction, and retraction to prevent the use of the work that was retracted in the future . He did this so that the erratum, statement of concern, and partial retraction could be remedied.

Using PubMed, Steen and Fang, (2013) conducted research to investigate the factors behind the rise in the number of scientific retractions. It was determined that the period of time between the publishing and the subsequent retraction was, on average, 32.91 months. When the frequency of retractions was compared to journal impact factors, it was found that journals with higher impact factors removed articles more often because the peer review process was scrutinized more carefully.

According to the findings of the study that was carried out by Cokol et al. (2007), the procedure of retracting any of the papers that should be retracted is never carried out. This is because the quality of a scientific article is reliant on elements such as the amount of work and time that is given to regulating their quality. This is due to the fact that the quality of a scientific article is based on these factors. According to the results of the research, a larger proportion of retraction requests were sent to high impact journals as compared to low impact journals. This shows that high impact journals either have a greater propensity to publish erroneous articles or scrutinise submissions in a substantially more in-depth way than low impact journals do. Alternatively, high impact journals may have both of these characteristics.

Trikalinos et al. (2008) looked at the characteristics and authors of papers that were retracted from high-impact journals due to the use of falsified data. They contrasted the publications that were retracted with identical articles that were published in the same journals but were not retracted. There was no discernible difference between retracted publications and their matched non-retracted counterparts in terms of the number of citations received over the course of the preceding year, the number of authors, the nation, the funding, or the field. However, retracted articles were twice as likely to have authors from multiple countries. It has been hypothesized that retractions brought on by falsification might take a significant length of time, particularly when prominent scholars are involved.

Steen (2011) carried out a study to investigate the influence that erroneous research has on the published medical literature and the effects that this has on the quality of life of patients. According to the findings of the study, retracted publications had garnered a total of 5503 citations, of which 5143 (or 93%) were linked to a research and 851 (or 15%) were prospective clinical trials including patients who were treated. In all, there were 28,783 participants recruited and 9189 patients treated in the 180 main studies that were ultimately discredited. In 851 studies which were secondary used a retracted paper as a source, about 445 064 subjects were enrolled and 70 501 patients were treated. The findings of the study indicated that a significant number of patients are at danger due to defective research, which reflects the fact that ideas disseminated in articles that have since been retracted may have an effect on later research.

Budd, Seivert, and Schultz, (1998) investigated the influence that retracted papers have on the flow of information in the biomedical field by analyzing the features of publications that were subsequently retracted from the scientific literature. The information was obtained by searching MEDLINE database for these papers from 1966 all the way up to August of 1997. It was inferred that retracted papers were consistently referenced by biomedical researchers as valid work in future publications, posing issues for medical science.

Rubric is the name that Bilbrey, Dell, and Creamer, (2014) gave to their approach for evaluating and figuring out the overall quality of retraction notices. They built this system. Each retraction notice was assigned a grade on a scale based on the rubric, and the current retraction quality notices for 171 retracted works that were published in 15 different journals were assessed. Over the last 50 years, retraction notifications had not improved, and they differed between and within journals. The notifications varied by area of research, author, and justification.

Chen et al. (2013) presented a visual analytic assessment of retracted papers in connection to other studies to enhance awareness of their possible risks. Science published their results, new visual analytic methods may help verify citation traces when these works are retracted. The number of retracted papers is rising, and many have hundreds of citations.

As part of their study, Resnik, Wager, and Kissling, (2015) emailed the top 200 scientific publications to enquire about the methods that they use for retracting articles. Nearly 147 (74%) journals provided a response, and 94 of those journals had retraction policies that allowed their editors to remove papers without the author's consent. It was discovered that the journal editor did not get the authors' permission before retracting their work since doing so would have been problematic because not all of the authors would have agreed with the journal's principles.

Foo (2011) drew attention to the fact that the current editing and peer-review procedure may have some potential flaws when it comes to dealing with fraudulent publications. In addition, the study investigated whether or not there was a significant time gap between fraudulent publications that were retracted before the year 2000 and those that were retracted after the year 2000. To do this, a comparison was made between the papers written by a single author and the overall number of journal publications that were retracted. Using the PubMed database, an investigation was conducted into the 303 retracted papers that had been produced by 44 distinct authors. According to the findings, just 6.60% of the papers that were retracted were written by a single author, and the detection time to uncover fraudulent publications dropped from 52.24 months (before the year 2000) to 33.23 months (after the year 2000). It was also shown that detection of fake papers are easier by using public databases that are freely available to the general public, such as PubMed.

3 METHODOLOGY

This section contains the details of the dataset, preprocessing stages of the data, statistical analysis and machine learning models that are utilized in this study to determine the paper that are close to retraction. Moreover, with the help of subject matter expert annotation scheme for annotation functions are proposed to find the papers that need reconsideration.

3.1 DATASET

This part containing the details about the gathering of the data from the retraction watch which is have a very large database of retracted papers. Moreover, it has the details of the attributes or features that we are going to use in this study.

3.1.1 Retraction Watch

The data set was obtained from the website known as Retraction Watch, which focuses on retractions in the scientific community and covers themes relating to scientific article retractions and retractions in general. Adam Marcus, a scientific writer, and Ivan Oransky, a former vice president of Editorial at Medscape, are the people responsible for creating the blog. Oransky and Marcus established Retraction Watch in Aug, 2010 with the intention of improving the transparency of the retractions process. They noted that retractions of articles are often not published, the reasons for retractions are not made known to the public, and other researchers or the general public who are uninformed of the retraction may base choices on false findings. Oransky used a study that was published in the Proceedings of the National Academy of Sciences as an example of a research finding that indicated a certain medication may be useful in the treatment of specific types of breast cancer. In spite of the fact that the paper's conclusions were later retracted, none of the media outlets that had first covered it for its positive findings—the basis on which a company had been founded—reported the retraction of the findings. According to Oransky and Marcus, retraction notices may also "be the source of fascinating narratives that reveal a lot about how research is undertaken." They can also shed light on instances of scientific fraud and give a glimpse into how science self-corrects. As of January 2021, more than 50 separate pieces of research have cited Retraction Watch. This comes at a time when the scientific publishing community is investigating the impact of retracted research. During the COVID-19 pandemic, Retraction Watch kept a separate list of retracted articles that contributed to the spread of misinformation about the pandemic. In addition, additional research was conducted to investigate the subsequent pollutions of further research caused by the citation and use of retracted papers within scholarly research.

According to the findings of Retraction Watch, retractions occur far more often than was previously supposed. When Retraction Watch was just being established, Marcus would say he "wondered if we'd have enough stuff." On a yearly basis, it was estimated that around eighty items were taken down. On the other hand, the website claimed around 200 retractions in its first year of operation. As of October 2019, there are 20,000 items in the Retraction Watch Database (Wikipedia contributors, 2022)

3.1.2 Data Collection

In this thesis, we used a medical dataset which was originally collected from Retraction Watch by scraping (Figure 3.1). The dataset includes scientific papers in the sub-domain of cancer studies.

	retacted_title	citing_title	text	citing_retracted	sharing_author	cited_alone	date	citation_section
0	Knockdown of long noncoding RNA HOTAIR inhibit...	Long Noncoding RNA HOTAIR Contributes to Progr...	HOTAIR is frequently observed in different can...	yes	no	yes	2020	introduction
1	Knockdown of long noncoding RNA HOTAIR inhibit...	The emerging role of lncRNAs in chondrocytes f...	HOTAIR is a human lncRNA, coded by a gene loca...	no	no	yes	2020	The specific role of an important lncRNA, HOTA...
2	Knockdown of long noncoding RNA HOTAIR inhibit...	The Role of Noncoding RNAs in B-Cell Lymphoma	The decreased level of miR-101 predicts poorer...	no	no	yes	2020	ncRNAs Regulate Key Signaling Pathways Associa...
3	Knockdown of long noncoding RNA HOTAIR inhibit...	Dysregulation of miRNA in Leukemia: Exploiting...	Moreover, in lymphoma, the interaction between...	no	no	yes	2021	miRNAs and lncRNAs Interaction
4	Knockdown of long noncoding RNA HOTAIR inhibit...	NRF1-regulated CircNSUN2 promotes lymphoma pro...	In order to study the relevant mechanisms of i...	no	no	yes	2021	background

Figure 3.1 Sample of data set

We are not using conventional data; instead, it is based on the ground truth. For instance, for every retracted paper, there is at least one citing paper that was retracted itself after citing the retracted paper, and there are a few other citing papers that have not yet been retracted despite citing the retracted papers. The data collection contains a total of 664 records. The record of the paper that is referencing the retracted publications is included in the variable "Citing retracted." According to the table below, 465 records have not yet been retracted and are labelled with a No label, whereas the remaining recordings are the retracted records with the Yes label which means they are retracted (Figure 3.2).

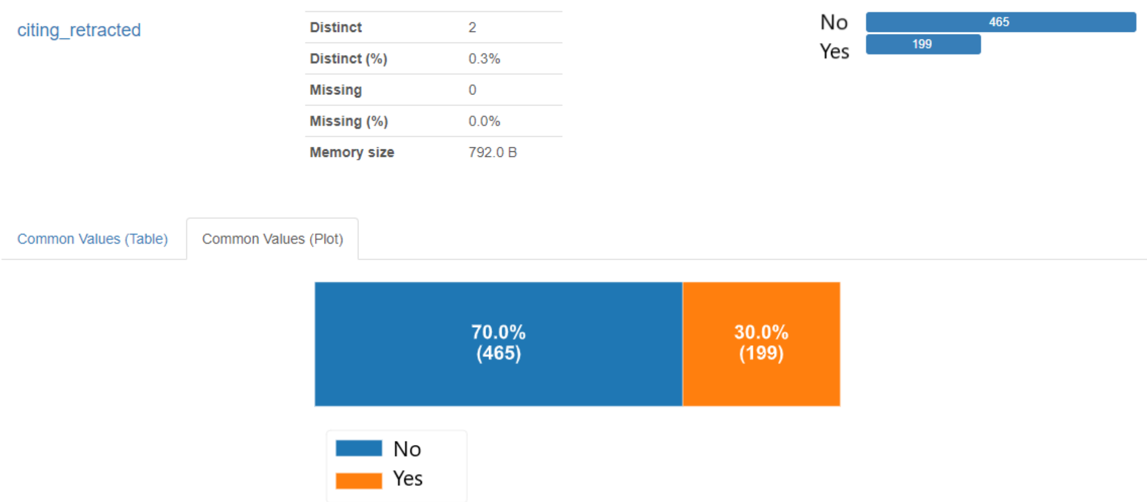


Figure 3.2 Papers retraction status after citing the retracted papers

The dataset has a rich collection of meta-data and the (Table 3.1) lists them along with a short description.

Table 3.1 Attributes with short description

Attributes	Explanations
retacted_title	Title of the retracted paper
citing_title	Title of the paper which is citing the retracted paper
text	Citation context (The part where the citing paper is mentioning the retracted paper plus with one above and below sentence)
citing_retracted	yes or no on the basis if the citing paper got itself retracted after citing the retracted paper
sharing_author	based on whether or not the author of the retracted work and the paper that was cited are the same person, yes or no
cited_alone	yes or no on the basis if the part where the citing paper is mentioning the retracted paper alone or with other references
citation_section	the section in which the retracted paper is discussed or retracted
polarity_of_ columns_text	polarity of the citation context where the citing paper is mentioning the retracted paper
bert_sim_colABo	similarity score of titles which are in the retrated_title and citing_retracted columns using bert

We decided to utilize this data set because mentioning flawed research not only encourages the spread of error but also poses a potential danger, particularly in medical literature where flawed research might put patients' lives at risk.

3.2 Data Annotation

With the help of a subject matter expert, we have used annotation scheme for citation functions in which we labeled our text column having the citation part as potentially harmful or not harmful. This all was done manually and this part is not included in the machine learning part because through this we will be satisfying our second question.

An enhanced annotation system that contains three broad citation function categories was presented by (Xiang Li, 2013). These categories are positive, neutral, and negative respectively. The presence of positive citations indicates agreement, use, or compatibility with the material being mentioned. Citations that are neutral refer to prior information or experience that is relevant to the referenced work, while citations that are negative highlight the flaws of the cited work. They expand these categories by classifying citation functions into more detailed subcategories and then ranking the original categories accordingly. POSITIVE, for instance (see + in the Table 3.2), demonstrates a broad sentiment of agreement among respondents. They separated POSITIVE into the categories Based on, Corroboration, Discover, Positive, Practical, Significant, and Standard (Xiang Li et al., 2013). The contrast and the co-citation both fall into the category of neutral, and the last one is negative category.

Table 3.2 Annotation scheme for citation functions

Citation Functions	Description
Based on (+)	A work is based on the work that is cited.
Corroboration (+	Two works back each other up.
Discover (+)	Recognize a technique's invention
Positive (+)	The work is successful
Practical (+)	The cited work is useful
Significant (+)	The mentioned work is important
Standard (+)	The cited work is widely used
Supply (+)	Recognize the material's producer
Contrast (=)	Neutrally compares two works
Co-citation (=)	Citations that appear closely
Neutral (=)	The work that was cited did not belong to any other function
Negative (−)	The weak points of the work cited are talked about

The (Figure 3.3) is showing the sample of data set after performing the annotations.

	retacted_title	citing_title	text	citing_retracted	sharing_author	cited_alone	date	citation_section	polarity_of_columns_text	sbert_sim_colABo	bert_sim_colABo
0	Knockdown long noncoding RNA HOTAIR inhibits c...	Long Noncoding RNA HOTAIR Contributes Progress...	HOTAIR frequently observed different cancer ty...	yes	no	yes	2020	introduction	-0.084524	0.820839	0.914084
1	Knockdown long noncoding RNA HOTAIR inhibits c...	The emerging role lncRNAs chondrocytes osteoar...	HOTAIR human lncRNA coded gene located chromos...	no	no	yes	2020	The specific role of an important lncRNA, HOTA...	-0.028571	0.495203	0.774413
2	Knockdown long noncoding RNA HOTAIR inhibits c...	The Role Noncoding RNAs B Cell Lymphoma	The decreased level miR 101 predicts poorer pr...	no	no	yes	2020	ncRNAs Regulate Key Signaling Pathways Associa...	-0.253333	0.728686	0.828911
3	Knockdown long noncoding RNA HOTAIR inhibits c...	Dysregulation miRNA Leukemia Exploiting miRNA ...	Moreover lymphoma interaction HOTAIR miR 148b ...	no	no	yes	2021	miRNAs and lncRNAs Interaction	0.000000	0.647600	0.865326
4	Knockdown long noncoding RNA HOTAIR inhibits c...	NRF1 regulated CircNSUN2 promotes lymphoma pro...	In order study relevant mechanism malignant be...	no	no	yes	2021	background	0.187500	0.598853	0.897310

Figure 3.3 Sample dataset after annotation

3.3 Data Preprocessing

The data gathered include certain "text" fields, such as citation context, article titles that have been cited and retracted, citation sections, and these "text" fields may contain noise as well as linguistic data that is either incomplete or erroneous. In order to investigate the linguistic information included in the dataset and provide an explanation for its true meaning, these erroneous data must first be rectified and removed. In light of this circumstance, data preparation becomes vital. The preparation of something before to its use or, in certain situations, feeding is what preprocessing refers to. As the name suggests, preprocessing involves processing. The approach that was developed using the Python programming language, as well as all of the preparation tasks that were performed in order to filter the noise from the data, are described in full in the text that accompanies the technique. The preparation of data is an essential step in every aspect of machine learning as well as in every other problem in which text is the primary component (Munawar, 2022).

3.3.1 Tools

3.3.1.1 Python

The processing of linguistic data was carried out over the course of this investigation using Python version 3.9.1 and the Natural Language Toolkit (nltk), which has been tailored specifically for usage with this specific version of Python. "Jupyter Notebook," which is an environment for interactive computing that is web-based, was used to assist in the implementation of the programming interface. It is a pretty simple process to construct a machine learning model in Python using a notebook. The fact that we can immediately see the outcomes of our computations and execute very little bits of code that handle our data is a really positive development. All of the components necessary for working with data may be assembled, validated, and executed with the assistance of this tool. We also have the option of turning our Notebook into a slideshow or saving it and sharing it online through GitHub. Binder allows you to collaborate on a Notebook without having to install the software required to do so (IBM- Unites states, 2022)

3.3.1.2 Natural Language Processing

Natural language, which may include any language—including English, French, Hindi, and others—is the means of communication that individuals use to communicate with one another for various reasons. Linguistic communication is the term for communication that takes place via language (Bird et al., 2009). This message can be sent orally or in written communication may take many forms, including emails, social media blogs, letters, books, and any other sort of written form that can be typed or handwritten. Verbal communication may take place in a variety of settings, including voice over phone, presenting a lecture, and anything else that can be heard. Additionally, each mode of communication, whether it is oral or written, have its own vocabulary, structure, syntax, and parts of speech. This is true whether the communication is in the form of words or sounds. Therefore, there are two ways to process natural language: first, by thinking rationally and writing it down, and second, by thinking logically and talking about it. Both of these methods are discussed more below. Additionally, the phrase "processing natural language" is used to express the concept of "logically thinking or comprehending." This is something that humans do in their minds in a logical way, while a computer does it by following steps written in programming language and carried out by Central Processing Units (CPU). Natural language processing is a field of study in computer science that is part of artificial intelligence or computational linguistics. It looks at how computers and natural languages made by humans interact.

It also has an impact on the computer programming languages, which enable interaction between programmers and actual physical entities as well as human language processing. These artificial languages, such as Python, C, C++, and Java, have their own grammar, vocabulary, and notation. Natural Language Processing (NLP) or computational linguistics may thus be used to describe the process of processing human language(s) using artificial languages. The term "natural language processing" thus refers to a broad range of techniques that allow the automated creation, manipulation, and analysis of real human languages.

So linguistic communication that we are using in this research is of written English and the form of written communication is research papers. So we are using the set of techniques of Natural language processing like removing stop words, word lemmatization, word tokenization etc. to get the automatic generation, manipulation and analysis of natural language.

3.3.1.3 NLTK

When it comes to executing operations on strings, Python's built-in methods for handling language processing tasks are fairly straightforward. Natural Language Toolkit (NLTK), which is available for Python, is used in order to accomplish the goal of achieving sophisticated capabilities for processing linguistic data. Students are able to study about natural language processing (NLP) and do research using NLTK, which is a collection of modules and corpora distributed under the GPL open-source licence (Bird et al., 2006). Over fifty different corpora and lexical resources, such as WordNet, are combined with language processing libraries, such as work tokenization, classification and stemming, tagging, parsing, and sematic rules, for the purpose of conducting an analysis of a text document. The fact that NLTK is entirely self-sufficient and has received praise from the academic world is the most important advantage it offers (Bird et al., 2009). Additionally, not only does it allow access to techniques and packages for basic NLP tasks, but it also provides the pre-processed and raw versions of standard corpora that are used in NLP books and courses. This is a really useful feature (Bird et al., 2009).

3.3.2 Techniques

3.3.2.1 Data Cleaning

The data collection that we are working with has some text columns, and those text columns could include some undesired information, such as special characters, repeated characters, or some unwanted space. Pre-processing and transformation of the data is required so that the machine learning algorithm may be applied to the dataset and provide reliable results for classification. If this step is completed successfully, then the classification can be performed. As a result, in order to carry out an accurate analysis of this data derived from the text, it is essential that the erroneous data be cleaned up and eliminated. According to Fernández-Gavilanes et al. (2016) the preprocessing of data helps to simplify the language used for assessing the sentiments included within text, normalizes the linguistic data, and gets rid of any noise that may be present. The (Figure 3.4) provides which might be considered the most comprehensive overview of preprocessing.

Figure 3.4 Data pre-processing Overview

Since the information was gathered by manually, it is possible that it contains extra blank spaces at the beginning, in between, or at the context of the text columns, as well as special characters like punctuation and repeated letters. To begin, all unnecessary white space is eliminated by using the built-in function that is provided by Python. Second, all of the pointless and unnecessary special characters that were originally in the text are removed. These characters consist of the following: | [] ;: - + () >?! @ #% *, as well as maybe more. Neither of these characters has an unique, well-defined meaning, nor does either make it clear whether they are used in a positive or negative context. As a result, eliminating them is the greatest option available. Additionally, these unique characters may occasionally be seen appended to words such as "sweetheart!" Words containing special characters (such as an exclamation mark, in this case) are not included in the dictionary, so when you look up two words and compare them, the dictionary won't know what you mean. The (Table 3.3) shows the pseudocode used for the pre-processing of given data set.

Table 3.3 Pseudocode for pre-processing of data

Input: Textual data

Output: For next steps, pre-processed data

For each **text** like retracted_ title, citing_retracted, text columns in **Data set**

Remove all the special characters ([_"\-;%|+&=*%.,!?:#$@\[\]/]) from the textual data. Store the result in text.

Return processed Text.

3.3.2.2 Word Tokenization

After the data set has been cleaned, the unprocessed words may still be present in the sentences. The words on their own have a certain meaning, and together they may convey an opinion or a feeling to the reader. Tokenization is the process or series of processes in natural language processing that consists of breaking sentences into their component words and punctuation marks (Bird et al., 2009). Tokenizing a phrase into words refers to the process of generating a list of terms by splitting the string into individual characters.

3.3.2.3 Word Lemmatization

Word lemmatization is yet another essential step in the process of natural language processing. By removing the affixation from the word or altering the vowel in it, this method may change the structural form of a word such that it corresponds more closely to the base form or the dictionary form of the term. The result that is obtained from the word is referred to as the lemma (Liu et al., 2012). Lemmas are the words that have a root meaning of the phrase that was sought, and the words that have been lemmatized are the gateway to WordNet (Bhattacharyya et al., 2014). In light of this, applying the lemmatization algorithm to the word will result in the production of a lemma, which will then be submitted to the WordNet dictionary. This will result in the extraction of the word's sense as well as its sense number, which is the goal for achieving a higher sentiment score for the word.

3.3.2.4 Removing Stop Words

Stop words (such as "and," "the," "am," and "is") have a high incidence in the text but little emotional value, therefore they do not alter the polarity score when applied to lexical resources during natural language processing. Examples of such words include "and," "the," "am," and "is." As a result of this, it is standard procedure for many researchers to remove stop words from the text before doing content analysis in the relevant area. According to Saif et al. (2012), they conducted an experiment in which they compared the results produced by filtering out stop words from the text to those obtained by leaving them in and found that the former yielded much higher accuracy in sentiment classification. Research by Firmino Alves et al. (2014) shows that deleting stop words from the text may be used as a strategy for classification, which in turn produces fascinating findings and accurate calculations of sentiment polarity. In the book written by Bird et al. (2009) the authors indicate that stop words could have very little lexical information. Furthermore, the authors propose that filtering stop words is an essential step before carrying out any additional processing activities.

3.4 Statistical Analysis

We are doing statistical analysis in this section because statistical analysis is the process of collecting data to identify the patterns and trends which help us while decision making. There are many insights which we can get quickly using this process.

3.4.1 Similarity

Bidirectional Encoder Representations from Transformers, or BERT for short, is a strategy for pre-training natural language processing (NLP) systems that was created by Google. BERT is a transformer-based machine learning method. We are using the cosine function to find the similarity of text data.

Calculating similarities between two texts is a fundamental part of a wide variety of text mining applications. For instance, if we had a perfect mechanism for determining the degree to which two sections of text are equivalent to one another, we could construct the perfect information retrieval system. In recent years, many different metrics have been proposed to deal with various information retrieval and natural language processing issues. Some examples of these distance measures include the Euclidean distance, Cosine similarity, Jaccard, index and Dice score, and the Jensen-Shannon divergence based metric. The cosine, which determines the angle that exists between two vectors, is by far the most often used metric that is currently available. In practice, it is determined by taking the dot product of two vectors that have been normalized (Li and Han, 2013). If you have two vectors of N dimensions, v and w , you can determine the cosine similarity between them using:

$$Cosine(\vec{v}, \vec{w}) = \frac{\vec{v} * \vec{w}}{|\vec{v}| * |\vec{w}|} = \frac{\sum_{i=1}^{N} v_i * w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}} \quad (1)$$

In this study, the similarity between the two columns of data set, namely citing title and retracted title has been calculate using (1)

3.4.2 Polarity

The primary focus of the algorithms used in sentiment analysis is on the processing of views, attitudes, and even emoticons that are included inside a large body of text. When compared to one another, the various methods provide quite different results with regard to the specified range of sentiments. Even while a simple analyzer can only describe up to three fundamental polar emotions (positive, negative, and neutral), more complex versions have a far wider range of possibilities.

TextBlob is a Python library that is free and open-source for processing textual data. It provides a straightforward API that may be used to access its functions and carry out fundamental NLP operations. TextBlob is capable of executing a wide range of operations on textual data, including the extraction of noun phrases, the analysis of sentiment, the classification of information, translation, and many more (Figure 3.5).

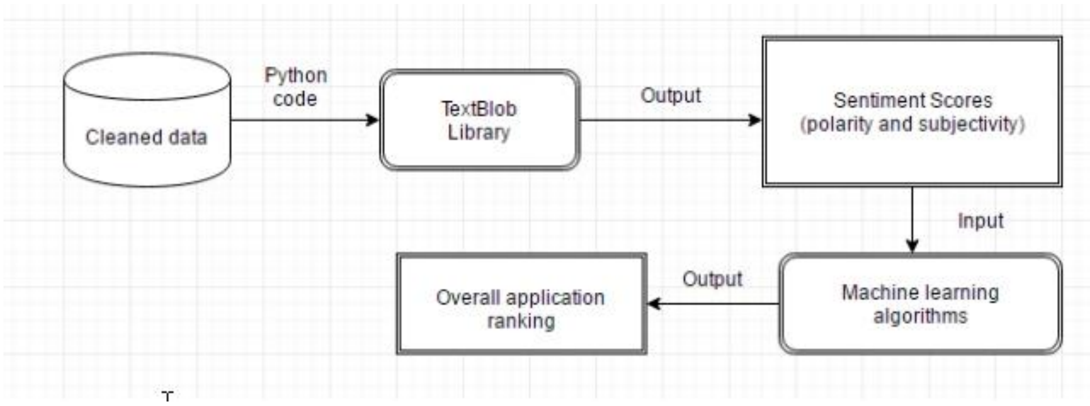


Figure 3.5 Sentiment Analysis in Python with TextBlob

Kuzminykh, N. (2020, October 23). *Sentiment Analysis in Python With TextBlob*. Stack Abuse.

https://stackabuse.com/sentiment-analysis-in-python-with-textblob/

TextBlob's output for a polarity job is a float that is somewhere between the range [-1.0, 1.0]. A value of -1.0 indicates a negative polarity, whereas a value of 1.0 suggests a positive polarity. Because it does not include any words from the training set, this score might also be equal to 0, which indicates that the statement is being evaluated in a manner that is neither positive nor negative. In the event that this score is equal to zero, the assertion will be regarded as being true (Kuzminykh, 2020).

3.4.3 Correlation

The coefficients are known as correlation matrices (Figure 3.6), and they are shown in table format as a correlation matrix. It also includes the values for each variable. The layout of the coefficients in rows and columns provides users with assistance in analyzing the connection between two or more variables and how those variables are dependent on one another. The value of the matrix is somewhere in the range of -1 to 1 (Wallstreetmojo, 2022).

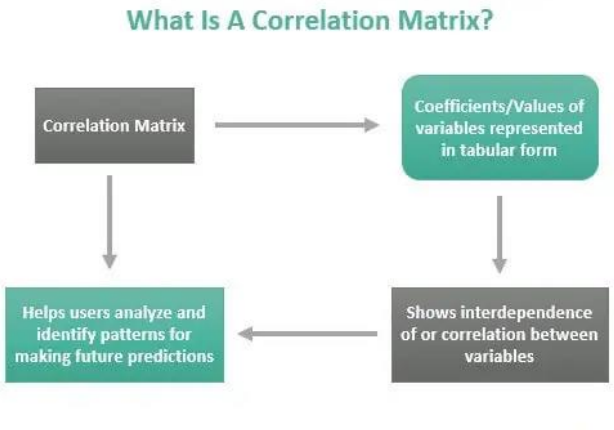


Figure 3.6 About Correlation

Wallstreetmojo, W. (2022, August 30). Correlation Matrix. WallStreetMojo. https://www.wallstreetmojo.com/correlation-matrix/

Key Takeaways

- The correlation matrix is a useful tool for determining the degree of interdependence or interaction that exists between two or more variables.

- Because it is presented in a tabular format, the datasets are simple to read, comprehend, and recognize patterns in so that they may be used to anticipate future occurrences.

- The notion assists in summarizing the facts and arriving at credible conclusions, which enables investors to make more informed judgments about their investments.

- The matrix may be efficiently generated via the use of modern technologies such as SPSS and Python-driven Pandas, or it can be done so by simply utilizing Excel.

3.4.4 Vectorization

The Count Vectorizer is an excellent function that is made available by the scikit-learn module in Python. It is used to convert a given text into a vector based on the frequency (count) of each word that occurs over the whole of the text. This is accomplished by using word counts. The use of frequency analysis allows for the completion of this change. This is beneficial, when we have numerous texts of this kind and we want to transform every word in each text into a vector (for using in **further text analysis).** Let us consider a few samples texts.

- "I like to play football"
- “John and I play tennis”

Each different word is represented by a column in the matrix created by Count Vectorizer, and each text sample from the document is represented by a row in the matrix. A count of all the distinct words in the text is likewise provided by Count Vectorizer. Each cell's value is just the number of words that are present in the relevant section of the text sample. The following illustration may help illustrate this point (Table 3.4):

Table 3.4 illustration of the point

	I	Like	To	Play	Football	John	And	tennis
Sentence1	1	1	1	1	1	0	0	0
Sentence1	1	0	0	1	0	1	1	1

Key Observations:

- The table contains representations of the document's eight distinct words, each of which is a column.
- The document has two different text samples, each of which is displayed as a row in the table.
- Each column has a number, which indicates the total amount of words contained inside that specific piece of text.
- The letters in each word have been changed to lowercase.

These terms are not kept in Count Vectorizer as strings since it does not use them. Instead, a specific index value is assigned to each one of them. In this scenario, the word "I" would be assigned index 0, the word "Like" would be assigned index 1, the word "To" would be assigned index 2, and so on. The real representation is shown in the table that may be seen in (Table 3.5).

Table 3.5 Sparse matrix

0	1	2	3	4	5	6	7
1	1	1	1	1	0	0	0
1	0	0	1	0	1	1	1

This way of representation is called sparse matrix.

3.5 Machine Learning

ML is a subfield of AI that focuses on developing algorithms that are able to learn patterns from data on their own without having to be explicitly taught such patterns by a human. Throughout the years, there has been an increase in the usage of ML in a number of sectors due to the accessibility of large amounts of information in addition to computing resources (Koblah et al., 2022).

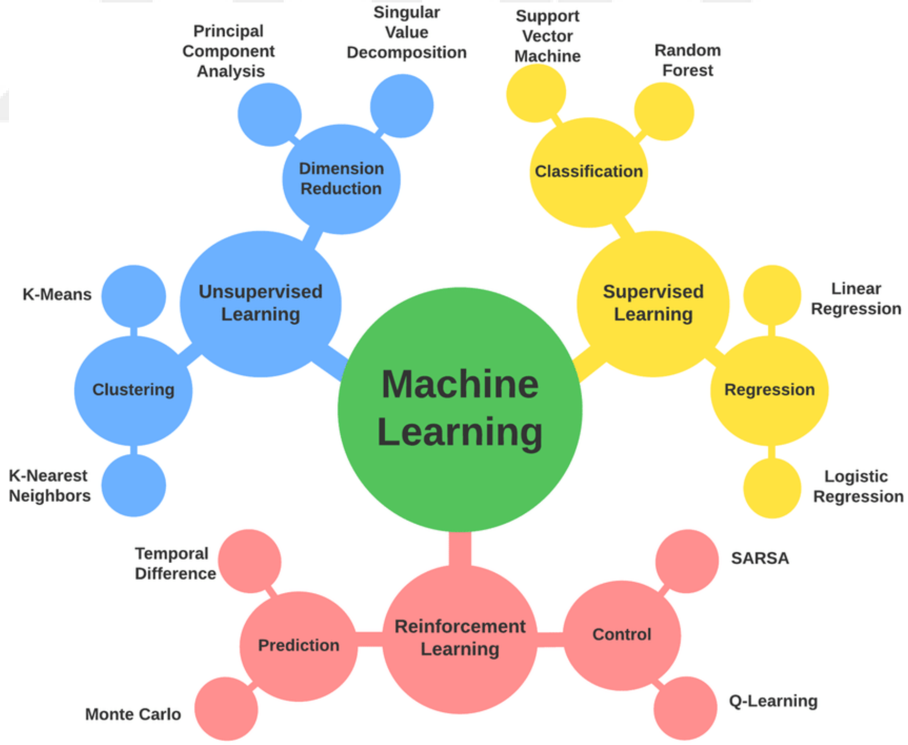


Figure 3.7 A categorization of major machine learning techniques

3.5.1 Types of Machine Learning

Supervised learning, unsupervised learning, and reinforcement learning are the three most common ML algorithms. In the following paragraphs, we will present an overview of these three primary categories and a summary of some example methods (Figure 3.7).

The form of machine learning known as supervised learning is task-driven and makes use of labelled training samples in order to map inputs to outputs. Supervised machine learning is used for a significant portion of the time to complete classification and regression tasks. In classification tasks, the inputs are translated to a discrete value (i.e., a class label), while in regression tasks, the inputs are mapped to a continuous value. Both types of tasks are performed by a computer. Because predicting the probability (which is a continuous value) that an input belongs to each class and providing the class with the greatest likelihood is how all classification problems may be conceived of, regression is the generic form of classification. Support vector machines (SVMs) and random forest are two common supervised machine learning algorithms that are used for classification tasks. On the other hand, linear regression and logistic regression are used for regression tasks (Alpaydin, 2020).

Unsupervised Learning is a kind of data-driven machine learning that does not need labelled training samples in order to discover patterns in incoming data. This type of learning is known as "deep learning." Clustering and dimensionality reduction are two common applications of unsupervised machine learning. Clustering activities entail grouping data based on comparable underlying structures, while dimensionality reduction tasks involve reducing data to its most basic (or prominent) underlying structures. Clustering tasks may be seen as a subset of dimensionality reduction tasks. For clustering, common unsupervised machine learning techniques include k-means and k-nearest neighbour (KNN), while for dimensionality reduction, common unsupervised ML approaches include singular value decomposition (SVD) and principal component analysis (PCA) (Alpaydin, 2020).

Reinforcement Learning, often known as RL, is a kind of feedback-based machine learning that includes learning from one's errors via a process of trial and error. In problem areas where time and event sequences are important, feedback may be delayed, and actions have consequences, RL is often used for prediction and control. In RL, forecasting the performance of some policy is part of the prediction process, while control is the process of deciding the optimum policy to implement in order to achieve the desired level of performance. Despite the fact that most practical applications focus on control, the issue of prediction must often be handled first. When it comes to prediction, popular RL approaches include Monte Carlo and temporal difference (TD) Learning, and when it comes to control, popular RL methods include SARSA and Q learning (Alpaydin, 2020).

It is necessary to properly organize the ML pipeline for each work, and this should be done regardless of the kind of machine learning model that is utilised (Koblah et al., 2022). The next paragraphs provide an overview of a generic machine learning pipeline, which is then summarized in the following sentence: (Figure 3.8).

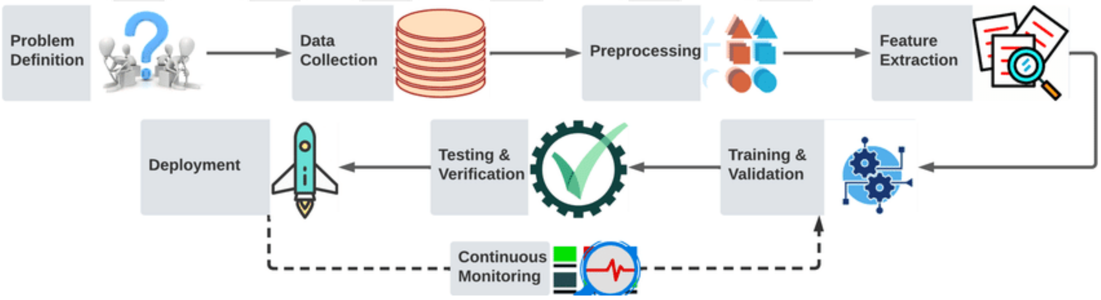


Figure 3.8 General machine learning pipeline.

3.5.2 Approaches used in this Study

Machine learning is a method wherein a computer is taught to do a task by inferring previously unseen patterns and structures in data with little human intervention. Computational techniques are those that make use of the information that is already accessible in order to make accurate predictions utilizing a variety of algorithms. The information may come in the form of data that has already been processed, or it may be compiled in electronic form and then prepared for analysis. In this study, we have used several Natural Language Processing (NLP) methods in order to conduct an analysis of the linguistic data included within the articles. Following the training of the dataset using machine learning classification algorithms, we will proceed with the analysis of the findings. And last, to wrap things up, we will end the work by contrasting the outcomes of several machine learning classification techniques.

3.5.2.1 Logistic Regression

Logistic regression is a classification technique that is used in circumstances in which the value of the target variable is of a categorical nature. This is the case while using the technique. When the data in question has a binary output, which indicates that it either belongs to one class or another, or is either a 0 or a 1, the majority of the time, logistic regression is utilized. This is because binary outputs may be represented by either a 1 or a 0. It is important to keep in mind that, in contrast to regression tasks, classification tasks use discrete categories (Kambria, 2019).

Here, by the idea of using a regression model to solve the classification problem, we rationally raise the question of whether or not we can draw a hypothesis function to fit to the binary dataset. This is brought up as a result of the fact that we are considering using a regression model to solve the problem. Despite the name of the method, which is "Logistic Regression," it is really a classification algorithm and not a regression technique.

The sigmoid function, also known as the logistic function, is a kind of function that, when plotted on a graph, looks like a "S" shaped curve. It takes the numbers that are between 0 and 1 and "squishes" them towards the margins at the top and bottom, categorizing them as either 0 or 1.

The Sigmoid function may be represented by the following equation:

$$y = \frac{1}{(1 + e^{-x})} \tag{2}$$

In this particular case, what does the variable e refer to? A value of about 2.71828 is assigned to the letter e, which stands for the exponential function or the exponential constant. (Figure 3.9) is showing the sigmoid function representation

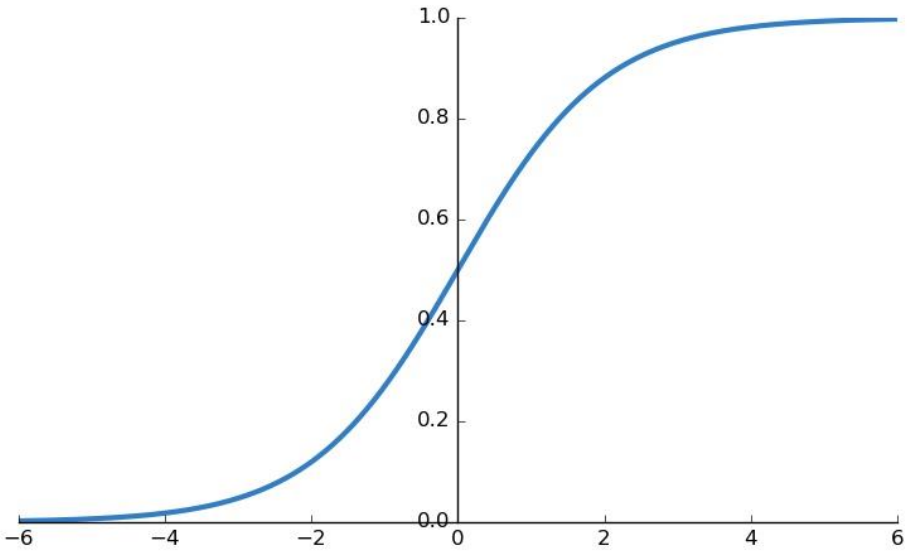


Figure 3.9 Sigmoid Function

(*Google Afbeeldingen Resultaat Voor http://ronny.rest/media/blog/2017/2017_08_10_sigmoid/sigmoid_plot.jpg*, n.d.)

This gives a value y that is very close to 0 if x is a large negative number and close to 1 if x is a large positive number. After the input value has been pushed toward 0 or 1, it can be run through a typical linear function, but the inputs can now be put into different categories. There are approximately three types of logistic regression binomial, multinomial and Ordinal. We are using Binomial type because we have two possible type of the Dependent variables, such as 0 or 1.

3.5.2.2 Naïve Bayes

In Natural Language Processing, the Naive Bayes algorithm is by far the most used classifier (NLP). In their study on sentiment analysis, Pang et al. (2008) compared the NB method with many other MLalgorithms and found that the Naive Bayes approach obtained an accuracy of 90 percent in categorizing the data set. The ease of use of this classifier is the primary benefit, in addition to its ability to accurately forecast the appropriate category for each new occurrence (Murphy 2006). It just multiplies all of the feature values that have been retrieved from each instance in the class (for example, Retracted and Non-Retracted) to create a class from those values. Every single labelled sentiment tag (instance) makes a contribution to the final classification result, and each token in the data set is assigned the same amount of significance in comparison to the others. During the process ML, the citing retracted will be categorized using this classifier, while other aspects will no longer be taken into consideration. Therefore, by taking into account only one attribute—namely, 'citing retracted'—for classification using the Naive Bayes classifier, and then analyzing the results of that classification.

The Naive Bayes classifier works based on the assumption that each characteristic or attribute of an instance (in this case, the "Citing Retracted" attribute) is assessed separately from the other characteristics that go into making up the given class. And as a direct result of this, in order to compute the Bayesian probability, it will multiply all of the elements of the feature vector for the class that has been specified. Here, for the data set that is provided, a particular class is denoted by the letter Y (Retracted, Non-Retracted), where X represents an instance that is defined by a feature vector of the form {X1 , X2,…, Xn } , where n is the number of features that are included in the data set. As a result, the Bayesian probability of the given class Y having an instance X may be determined using the following equation, which was provided by (Murphy et al., 2006):

$$P(Y/X) = \frac{P(^{X}/_{Y})P(Y)}{\sum_{Y}^{C} P\left(\frac{X}{Y}\right) P(Y)} \quad (3)$$

- X: It is instance of a class (Data Labels)
- Y: It is sentiment class (Retracted, Not Retracted)
- P(X|Y): instance happened for each value of Y in a certain class (class-conditional density)
- P(Y) : class prior probability

In Bayesian probability, P (Y|X) represents the likelihood that class Y will be observed in relation to instance X, while P (X|Y) represents the likelihood that class Y will witness instance X. This may be found by using the equation. In this particular scenario, the probabilities of each occurrence (the data labels) fall into a certain category (Retracted and Non-Retracted).

3.5.2.3 Random Forest

Random Forest is a kind of supervised learning that is used in machine learning. It offers solutions to problems involving classification and regression. Ensemble learning brings together a number of different classifiers to solve complex problems and improve overall model performance.

Usman, Shafique and Malik, (2016) with the help of tree bagging and the random subspace approach, we construct decision trees in our model by first drawing a random sample from our training set. By selecting samples at random, we make the forest include a variety of tree types. A classification may be derived from each tree. Then, we choose the output of the trees in the forest that have the highest correlation with one another. As soon as all of the trees in the forest have been constructed, the labelled data will be sent via the trees. The proximities are coming up next, and the closeness between two events will grow by one if both of those events are on the same leaf node.

3.5.2.4 Support Vector Machine:

It is a kind of supervised learning that generates a mapping function by using the data from the training set that is readily accessible (Wang et al., 2005). Support Vector Machines, or SVMs, are often used for problems involving classification and nonlinear regression. These machines can classify both linear and nonlinear data (Wang et al. 2005). The mapping function may also serve as the classification function, which sorts the data in the data set according to their respective labels. SVMs are universal learners, as stated by (Joachims et al., 1998). This means that they may be used to train polynomial classifiers, and their ability to learn is not dependent on the dimensionality of the feature space. According to Joachims et al. (1998) SVM is particularly helpful for concerns relating to text classification by linearly separating them. One of the drawbacks of using SVM is that it is unable to differentiate between words that have different meanings in various phrases. As a result, specific "domain-based lexicons" are unable to be constructed (Joachims et al., 1998).

3.5.3 Evaluation Metrics

Evaluation metrics are used in order to determine the quality of model. In the field of machine learning, understanding how to assess your model is one of the most significant issues. When you are developing your model, it is of the greatest importance to evaluate how correctly it predicts the result that you expect. For each set of machine learning algorithms, we use different ways to measure how well they work. We use classification metrics to evaluate classification models and regression metrics to evaluate regression models (Goyal, 2022).

Classification: It makes a category prediction based on given inputs. In these issues, a data point is attempted to be categorized into a certain category or class. In this case, the desired result will be a discrete or categorical value, such as 1 or 0, Yes/No, spam/not spam, etc.

Regression: It makes a prediction based on certain inputs, usually a number. These problems rely only on training data to predict continuous values utilizing input variables. Here, a number or actual value, such as time-series data, sales figures, heights, weights, etc., will always be the desired result.

The following is a selection of popular classification metrics that are used while evaluating models:

Accuracy: Accuracy is the most straightforward statistic to use when evaluating models. It is the proportion of accurate predictions to the total number of predictions produced for a given dataset.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made} \tag{4}$$

When the target class is properly balanced, accuracy may be helpful, but when the target class is poorly balanced, accuracy is not a suitable option.

Precision or Positive Predictive Value: It is the proportion of True Positives to all positive predictions made by the model. It is beneficial for skewed and imbalanced datasets.

The lesser the precision, the more false positives the model predicts.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall or Sensitivity: It is the number of true positives in your dataset compared to the total number of positives. It measures how well the model can find positive samples. The recall goes down as the number of false negatives predicted by the model goes up.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1-score or F-measure: Precision and recall are combined into one single metric here for your convenience. When the F1 score is greater, it indicates that our model is performing more effectively. The value of F1-score might fall anywhere between 0 and 1. The F1 score is determined by taking the weighted average of both the precision and recall categories. A high F-score for the classifier will only be achieved if both the precision and the recall are very good. Only classifiers that have a comparable level of accuracy and recall are given preference by this measure.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

4 FINDINGS

This section contains the experimental setup of our proposed model. The findings are shown that we achieved using multiple machine learning algorithms using citation context and metadata for the evaluation of scientific retractions. Moreover, the different cases results are shown that we achieved using the citation scheme for annotation functions.

4.1 DATA SET PREPARATION

The one of the major contributions of the thesis is the preparation of the dataset. The data is gathered from Retraction Watch in the raw form. Then we did data annotation using the annotation scheme for citation functions. We labeled the data like:

- retracted title
- citing title
- text
- citing retracted
- sharing author
- date to publish
- citation section
- cited alone

The total number of entries that we collected are 664. There are 199 papers which are retracted after citing the retracted paper and 465 are citing the same retracted paper but not yet retracted.

4.1.1 Data Cleaning

The following procedure have been applied to prepare the dataset before processing stage:

- Removed special characters ({} ( ) \ | [ ] - + < > ; : ? @ # ! % *,) from the text fields
- Removed URL links
- Removed extra spaces

4.1.2 Word Tokenization

In this study, the tokenize package from the Natural Language Toolkit (NLTK) is used in order to tokenize the words. The characteristics of the data you are working on and the language you are using both have a role in the decision of the tokenizer to use (Figure 4.1).

	retacted_title	citing_title	text	citing_retracted	sharing_author	cited_alone	date	citation_section
0	[Knockdown, of, long, noncoding, RNA, HOTAIR, ...	[Long, Noncoding, RNA, HOTAIR, Contributes, to...	[HOTAIR, is, frequently, observed, in, differe...	yes	no	yes	2020	introduction
1	[Knockdown, of, long, noncoding, RNA, HOTAIR, ...	[The, emerging, role, of, lncRNAs, in, chondro...	[HOTAIR, is, a, human, lncRNA, coded, by, a, g...	no	no	yes	2020	The specific role of an important lncRNA, HOTA...
2	[Knockdown, of, long, noncoding, RNA, HOTAIR, ...	[The, Role, of, Noncoding, RNAs, in, B, Cell, ...	[The, decreased, level, of, miR, 101, predicts...	no	no	yes	2020	ncRNAs Regulate Key Signaling Pathways Associa...
3	[Knockdown, of, long, noncoding, RNA, HOTAIR, ...	[Dysregulation, of, miRNA, in, Leukemia, Explo...	[Moreover, in, lymphoma, the, interaction, bet...	no	no	yes	2021	miRNAs and lncRNAs Interaction
4	[Knockdown, of, long, noncoding, RNA, HOTAIR, ...	[NRF1, regulated, CircNSUN2, promotes, lymphom...	[In, order, to, study, the, relevant, mechanis...	no	no	yes	2021	background

Figure 4.1 The result obtained after tokenizing word

4.1.3 Word Lemmatization

Here, using the "WordNetLemmatizer" class from the Python NLTK's stem package's "wordnet" class, we lemmatize words by comparing them character by character. It is an excellent option to utilize in order to extract useful lemmas and vocabulary from the text (Bird et al. 2009). It is crucial that the input word be sent to the "WordNetLemmatizer" algorithm in lower case to ensure accuracy if you want to identify a word's lemma or root meaning. The example of word lemmatization in our data is given below (Table 4.1).

Table 4.1 Example on Word Lemmatizing

Word Tokenization	[‘HOTAIR’, ‘is’, ‘frequently’, ‘observed’, ‘in’, ‘different’, ‘cancer’, ‘types’, ‘and’, ‘closely’, ‘related’, ‘to’, ‘increased’, ‘malignancy’, ‘and’, ‘poor’, ‘prognosis’, ‘due’, ‘to’, ‘the’, ‘acceleration’, ‘of’, ‘tumor’, ‘metastasis’]
Lemmatized Text	[‘HOTAIR’, ‘be’, ‘frequently’, ‘observe’, ‘in’, ‘different’, ‘cancer’, ‘type’, ‘and’, ‘closely’, ‘relate’, ‘to’, ‘increase’, ‘malignancy’, ‘and’, ‘poor’, ‘prognosis’, ‘due’, ‘to’, ‘the’, ‘acceleration’, ‘of’, ‘tumor’, ‘metastasis’]

4.1.4 Removing Stop Words

We applied the stop word function on the three columns which are retracted title, citing title and on the text. Example of the stop words is shown in the (Table 4.2):

Table 4.2 Example with removed stop words

Input Text	“HOTAIR is frequently observed in different cancer types and closely related to increased malignancy and poor prognosis due to the acceleration of tumor metastasis”
Output Text	“HOTAIR frequently observed cancer types closely increased malignancy poor prognosis acceleration tumor metastasis”

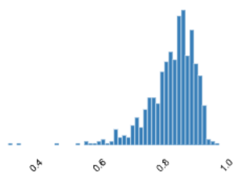
Stop words like "is," "in," "different," "and," "related," "to," "due," "the," and "of" are taken out of the sentence in the result shown above. When used on lexical resources to figure out a word's polarity, the filtered words may still have the same meaning as the original word.

4.2 SIMILARITY AND POLARITY RESULTS

The similarity score between the two columns retracted title and citing retracted were found using the BERT. The polarity score lies between 0 to 1 where 0 shows no similarity and 1 shows 100% similarity results. As we are also checking that which parameter can cause the paper retraction so we are trying to figure out that do we can draw any conclusion for this parameter. The below (Figure 4.2) is showing the statistics of the similarity check between the two mentioned columns using the BERT techniques.

bert_sim_colABo
Real number (ℝ)

Distinct	656	Minimum	0.31063569
Distinct (%)	98.8%	Maximum	0.97608149
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.82734561	Memory size	5.3 KiB



More details

Statistics Histogram Common values Extreme values

Quantile statistics

Minimum	0.31063569
5-th percentile	0.68633075
Q1	0.79178907
median	0.8441909
Q3	0.88135003
95-th percentile	0.91782392
Maximum	0.97608149
Range	0.6654458
Interquartile range (IQR)	0.089560955

Descriptive statistics

Standard deviation	0.077680392
Coefficient of variation (CV)	0.093891103
Kurtosis	6.0827951
Mean	0.82734561
Median Absolute Deviation (MAD)	0.042757094
Skewness	-1.7382303
Sum	549.35748
Variance	0.0060342432
Monotonicity	Not monotonic

Figure 4.2 Similarity of the paper titles using BERT

The minimum similarity score that we got is of 31% and the maximum score is 97%. The average similarity score of the whole data set for the title columns is 82%. These are the stats of the whole documents and if we narrow down this to the only retracted papers then there are approximately 198 papers out of 199 who had similarity score of the title columns more than 60%. These results shows that the retracted papers and the citing retracted papers have quite similar titles. Below (Table 4.3) is showing the

2 results that we got after finding the similarity using Bert.

Table 4.3 Similarity score samples using BERT

Retracted_title	Citing_retracted	Similarity score using Bert
"Apoptosis Induced by the Histone Deacetylase Inhibitor FR901228 in Human T-Cell Leukemia Virus Type 1-Infected T-Cell Lines and Primary Adult T-Cell Leukemia Cells"	"Inhibition of heat shock protein-90 modulates multiple functions required for survival of human T-cell leukemia virus type I-infected T-cell lines and adult T-cell leukemia cells"	0.960903943
"Inhibition of hTERT Gene Expression by Silibinin-Loaded PLGA-PEG-Fe3O4 in T47D Breast Cancer Cell Line"	"Synthesis, characterization, biocompatibility of hydroxyapatite-natural polymers nanocomposites for dentistry applications"	0.624494672

The data set we are using have a column 'text'. Each entry have a small paragraph in this text column having citation context of the paper with one above and below sentences. We found the polarity of the text column which is actually the polarity of the citation context of the paper. The analysis of the polarity of text column is shown in (Figure 4.3).

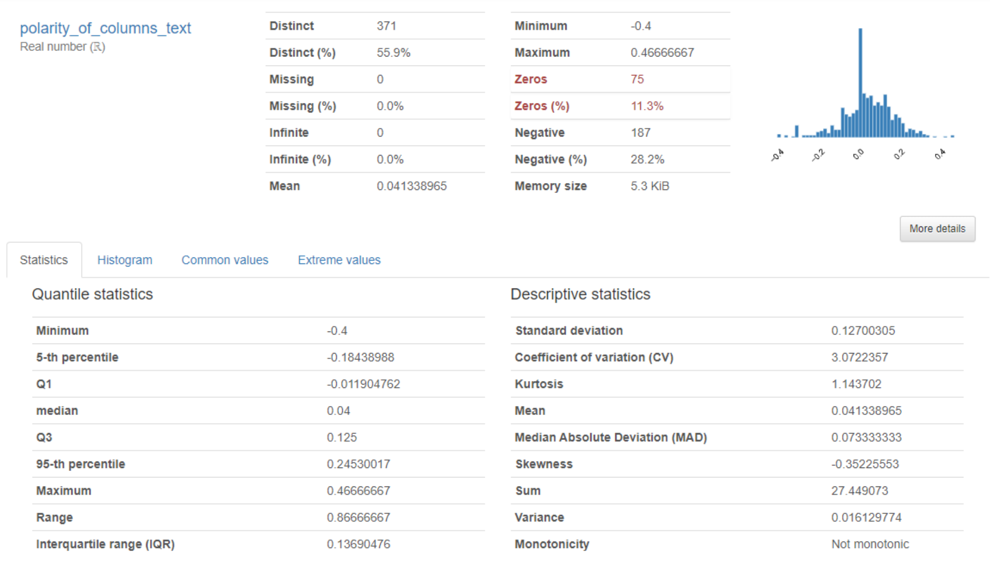


Figure 4.3 Polarity score of column text

The (Figure 4.3) shows that the minimum polarity of the text column used in the data set is -0.4 and the maximum value is 0.428. The average score we got 0.04 for the whole document. Taking into the consideration only the retracted papers then there are approximately 124 papers out of 199 which have positive polarity score. The (Table 4.4) is showing the polarity of two records among retracted papers.

Table 4.4 Polarity status of 2 records

Text	Polarity
“Gastric carcinoma (GC) causes the highest cases of cancer related death in China, largely resulting from cell invasion and early metastases [14]. Great efforts have been made to elucidate the molecular carcinogenesis of GC, as well as to search for an effective molecular therapy to assist surgery and chemotherapy”	0.42857143
“The GST-VDR (DEF) and GAL4-VDR (DEF) [pM-VDR(DEF)] constructs were prepared as described previously.(17,21) The eukaryotic expression vector for full-length human CDP (pCMVSPORT6-CDP) was purchased from Open Bio systems (Image ID 5740343).”	0.01111111

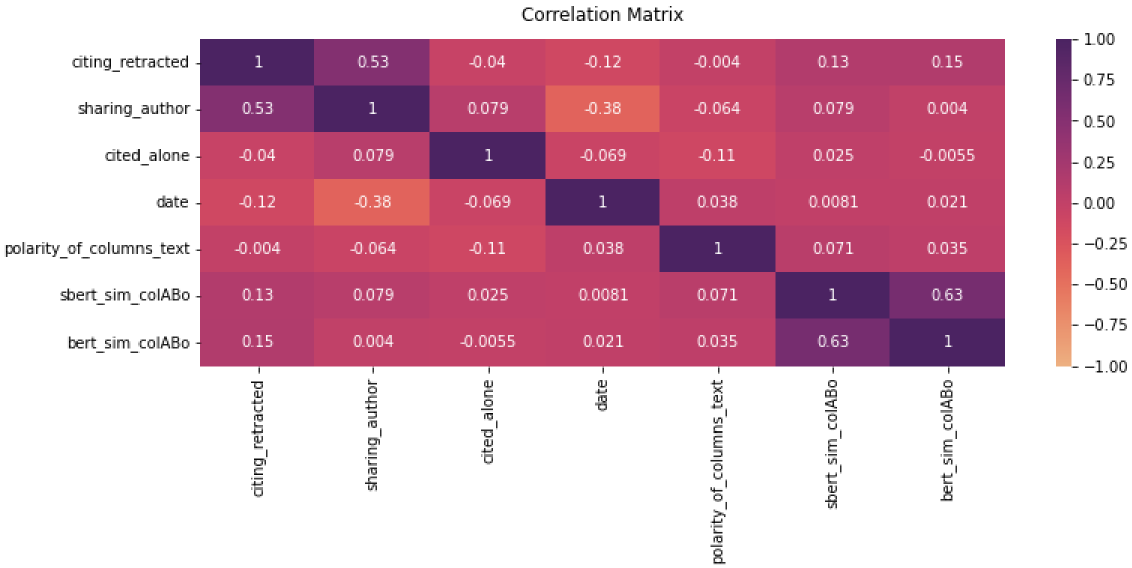


Figure 4.4 Correlation Matrix

The (Figure 4.4) is showing that the sharing author have strong association with the retracted papers and all other features are independent from each other.

4.3 CLASSIFIER RESULTS

There have been four different methods to machine learning used throughout the process of training on a labelled data set, and the results acquired using any of the classifiers have shown to be accurate. The performance assessment, accuracy, sensitivity, and classification results produced via the use of Logistic Regression, Naive Bayes, Support Vector Machines, and Random Forest. The statistics of the performance is show in (Table 4.9).

4.3.1 Performance of Logistic Regression:

For the implementation of logistic regression, we used Python programming language. The same processed data is used as the input. The required libraries are already imported as we performed the previous defined steps including the vectorization in which we have converted our text data into numeric values. Because it is imperative that we do not test on the same data that we trained on, our citing retracted column has been removed from the data set, and the data set has been divided into testing and training sections.

This is because testing on the same data that we trained on would result in extreme bias, and the results would not accurately reflect what the classifier has been able to learn.

For the purpose of data splitting, we made use of the "train test split" function in Scikit-learn. When we did so, we stated that we wanted to reserve 20% of the data for testing purposes. Now we fed the data to the logistic regression classifier and for this we imported this logistic regression model into the code and then created the instance of the classifier and fit it to the training data. After that we created the prediction by running the classifier on the test data set. The classification accuracy of Logistic Regression is (0.82) which is 82%. Here, 1 showing the papers that are retracted and 0 is for those which are not yet retracted. The performance measures stats are given in (Table 4.5)

Table 4.5 The Performance measures using logistic regression

	Precision	**Recall**	**F1-Score**	**Support**
0	0.80	0.99	0.88	89
1	0.96	0.50	0.66	44
Accuracy			0.83	133
Macro avg	0.88	0.74	0.77	133
Weighted avg	0.85	0.83	0.81	133

4.3.2 Performance of Naïve Bayes:

The classification accuracy of Logistic Regression is (0.72) which is 72%. Here, 1 is for retracted and 0 is for not yet retracted papers. The performance measures stats are given in (Table 4.6).

Table 4.6 Performance measure using the Naive Bayes

	Precision	Recall	F1-Score	Support
0	0.75	0.89	0.81	89
1	0.64	0.41	0.50	44
Accuracy			0.73	133
Macro avg	0.70	0.65	0.66	133
Weighted avg	0.72	0.73	0.71	133

4.3.3 Performance of Linear SVC:

While using this strategy, we partitioned the data using the "train test split" function included in Scikit-learn. When doing so, we indicated that twenty percent of the data should be reserved for testing purposes. Now that we had the data ready, we loaded it into the SVM classifier. In order to do this, we first imported the SVM model into the code, then built an instance of the classifier, and last tailored it to the training data. After that, we made the prediction by applying the classifier to the test data set and seeing the results. In the end, we determined that our accuracy was 0.827 percent. Here, 1 showing the papers that are retracted and 0 is for those which are not yet retracted. The performance measures stats are given (Table 4.7).

Table 4.7 Performance measures using SVM

	Precision	Recall	F1-Score	Support
0	0.69	0.12	0.21	89
1	0.33	0.89	0.48	44
Accuracy			0.38	133
Macro avg	0.51	0.50	0.35	133
Weighted avg	0.57	0.38	0.30	133

4.3.4 Performance of Random Forest:

The classification accuracy of Random Forest is (0.82) which is 82%. Here, 1 is for retracted and 0 is for not yet retracted papers. The performance measures stats are given (Table 4.8).

Table 4.8 Performance measures using the Random Forest

	Precision	**Recall**	**F1-Score**	**Support**
0	0.69	0.12	0.21	89
1	0.33	0.89	0.48	44
Accuracy			0.38	1333
Macro avg	0.51	0.50	0.35	133
Weighted avg	0.57	0.38	0.30	133

Now our classification algorithms are trained and logistic regression is giving approximately 82% accuracy so now we implemented the functionality in which if all the feature given in the data set are given as the input, then our model will predict the label like this paper has chance of retraction or not in terms of 1 or 0. 1 is for retracted and 0 is used for not retracted. The Output of the implemented functionality is shown below in which we predicted the output by given all feature values as input and then cross checked with the actual value and we noticed that our model is prediction approximately the same accuracy as we get while classification. The output is show below (Figure 4.5).

ActualValue	Predicted	sharing_author	cited_alone	polarity_of_columns_text	sbert_sim_colABo	bert_sim_colABo
1	1	0	1	-0.084524	0.820839	0.914084
0	0	0	1	-0.028571	0.495203	0.774413
0	0	0	1	-0.253333	0.728686	0.828911
0	0	0	1	0.000000	0.647600	0.865326
0	0	0	1	0.187500	0.598853	0.897310
...	...	...	...	...	...	...
0	0	0	1	0.049286	0.685572	0.812386
0	0	0	0	-0.087500	0.704523	0.846388
1	1	1	0	0.200000	0.754726	0.850312
1	1	1	1	-0.166667	0.796339	0.908443
0	0	0	0	-0.160000	0.778196	0.851769

Figure 4.5 Result of actual and predicted values of retraction

Table 4.9 Comparison of Models

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-Score**

Logistic Regression	0.82	0.88	0.74	0.77
Naïve Bayes	0.72	0.70	0.65	0.66
SVM	0.81	0.51	0.50	0.35
Random Forest	0.82	0.88	0.74	0.77

4.4 HIGHLY IMPORTANT FEATURE

After the statistical analysis we have found some strong association between the features of the data set. The (Table 4.10) is showing the features with impacting percentage percentage in among retracted papers.

Table 4.10 Features with high impact

	Features	**Number of papers among 199 retracted papers**	**Percentage impacting**
1	Sharing author	123	61%
2	Positive polarity of text column	126	63%
3	Similarity of text more than 60%	198	99%
4	Sharing author+ cited alone	64	32%

The above (Table 4.10) is clearly showing that the 61% of the retracted papers have sharing author in our dataset which is huge in number. 63% of the retracted paper have positive polarity of the text column which means they have some positive intent while discussing the retracted papers and this is the polarity of the citation context including one sentence above and below. There are 99% papers among the retracted papers who have Similarity score more than 60% and this similarity score is of the titles of the papers. Moreover, there are 32% of the retracted papers who have sharing author and they are cited alone.

4.5 POTENTIALLY HARMFUL OR NOT HARMFUL

With the help of subject matter expert, we used annotation scheme for citation functions to find the different cases for the retracted papers. We made a new label harmful and then we mapped the citation functions to the text column which have small paragraphs for each entry. After mapping the text, we marked them as harmful or not harmful keeping the sense of citation functions. We have categorized our findings into 4 parts shown in (Table 4.11).

Table 4.11 4 Cases obtained using citation functions

	Harmful	Not Harmful
Retracted	93	105
Not Retracted	221	224

Case 1: Papers got retracted but they were also Harmful

This case involves a total of 93 out of 664 publications being retracted, and those papers were also harmful; hence, this case is acceptable since it is expected that retracted papers would be harmful. It's possible that their findings include some inaccuracies, or that they were retracted because they cited a previously retracted study in a too favorable manner.

Case 2: Papers got retracted but they were not Harmful

In this case, there are a total of 105 out of 664 publications that have been retracted, and because they did not do any harm, this demonstrates that it is not necessary for a work to be retracted just because it cites a paper that has been retracted. There is a possibility that the results or conclusions might be deemed inaccurate, which would result in the paper being retracted.

Case 3: Papers are not yet retracted but it is harmful

In this scenario, there are a total of 221/664 papers that have not been retracted despite the fact that they are harmful. Because of this, we have a primary concern regarding this scenario. These papers have the potential to be flagged as having some serious flaws, which could result in the papers being retracted; however, before the papers are retracted, we may need to reconsider these papers.

Case 4: Papers are not retracted yet but they are also not harmful

In this case there are total 224/664 papers that are not retracted and they are also not harmful. The papers belong to case 4 are not expected to be retracted.

5 CONCLUSION AND FUTURE WORK

A new technique for performing the evaluation of scientific retractions on a data set has been presented in this thesis. The proposed method employs machine learning algorithms and the annotation scheme for citation functions. Using techniques from Natural Language Processing (NLP), the algorithms that were created for cleaning or filtering data and for pre-processing data are illustrated.

In the first part of our research, we focused on the relation between the retracted articles and the articles that are close to retraction. We have noticed that the polarity of the text column, containing the citation context, similarity score among the paper titles and sharing authors are showing very high association with the retracted articles. Moreover, we build machine learning models like Logistic Regression, Naïve Bayes, SVM and Random Forest, in which we are giving the metadata as input including the citation context and our model is predicting that this paper have chances or retraction or not in terms of 0 or 1. Here, 0 means **not** and 1 means **yes** and we have tested our model and we got the 82% accuracy of the model which means there are 82% chances of right prediction. In short if we provide all the metadata that discussed in previous chapters and do all the steps that we did, then we confidently say that based on these parameters we can find the association of the papers that are close to retraction with the retracted articles.

We further addressed the problem of specifying the case where it could be necessary to reconsider the paper that cite the retracted articles? We have labeled our text columns as potentially harmful or not harmful based on the annotation scheme for citation functions. We have identified four cases (discussed in detail in previous chapters) and among those third case is that which the paper is harmful but not retracted yet. In this case, we have 221 entries out of 664 that we flagged which means that these papers need reconsideration. That means we have reduced the amount of paper around 33%. According to our findings, the other portion is find but these 221 papers may have chances of retraction in future and they need to re-examine or reconsider.

As the future work different domains can be considered in the context of healthcare for generation of new dataset. These new datasets will have the potential to provide more insights and patterns. Furthermore, it would be exciting to develop deep learning models for feature engineering.

REFERENCES

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. In AMIA Annual Symposium Proceedings (Vol. 2010, p. 11). American Medical Informatics Association.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Anderson, M. H., & Lemken, R. K. (2020). Citation context analysis as a method for conducting rigorous and impactful literature reviews. Organizational Research Methods, 1094428120969905.

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. Scientometrics, 113(1), 547-565.

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. Scientometrics, 113(1), 547-565.

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Bhattacharyya, P., Bahuguna, A., Talukdar, L., & Phukan, B. (2014, January). Facilitating multi-lingual sense annotation: Human mediated lemmatizer. In Proceedings of the Seventh Global Wordnet Conference (pp. 224-231).

Bilbrey, E., O'Dell, N., & Creamer, J. (2014). A novel rubric for rating the quality of retraction notices. Publications, 2(1), 14-26.

Bird, S. (2006, July). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (pp. 69-72).

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with python O'reilly media Inc.

Bolboacă, S. D., Buhai, D. V., Aluaș, M., & Bulboacă, A. E. (2019). Post retraction citations among manuscripts reporting a radiology-imaging diagnostic method. PLoS One, 14(6), e0217918.

Budd, J. M., Coble, Z., & Abritis, A. (2016). An investigation of retracted articles in the biomedical literature. Proceedings of the Association for Information Science and Technology, 53(1), 1-9.

Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: reasons for retraction and citations to the publications. Jama, 280(3), 296-297.

Budd, J. M., Sievert, M., Schultz, T. R., & Scoville, C. (1999). Effects of article retraction on citation and practice in medicine. Bulletin of the Medical Library Association, 87(4), 437.

Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Sources of error in the retracted scientific literature. The FASEB Journal, 28(9), 3847-3855.

Chen, C., Hu, Z., Milbank, J., & Schultz, T. (2013). A visual analytic study of retracted articles in scientific literature. Journal of the American Society for Information Science and Technology, 64(2), 234-253.

Cokol, M., Iossifov, I., Rodriguez-Esteban, R., & Rzhetsky, A. (2007). How many scientific papers should be retracted?. EMBO reports, 8(5), 422-423.

Couzin, J., & Unger, K. (2006). Cleaning up the paper trail.

da Silva, J. A. T. (2016). Retractions represent failure. Journal of Educational and Social Research, 6(3), 11.

Davis, P. M. (2012). The persistence of error: a study of retracted articles on the Internet and in personal libraries. Journal of the Medical Library Association: JMLA, 100(3), 184.

Decullier, E., Huot, L., Samson, G., & Maisonneuve, H. (2013). Visibility of retractions: a cross-sectional one-year study. BMC research notes, 6(1), 1-6.

Dinh, Ly, Janina Sarol, Yi-Yun Cheng, Tzu-Kun Hsiao, Nikolaus Parulian, and Jodi Schneider. "Systematic examination of pre-and post-retraction citations." Proceedings of the Association for Information Science and Technology 56, no. 1 (2019): 390-394.

Fang FC, Steen RG, & Casadevall A (2012). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences USA, 109(42), 17028–17033.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences, 109(42), 17028-17033.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. Expert Systems with Applications, 58, 57-75.

Firmino Alves, A. L., Baptista, C. D. S., Firmino, A. A., Oliveira, M. G. D., & Paiva, A. C. D. (2014, November). A Comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup. In Proceedings of the 20th Brazilian Symposium on Multimedia and the Web (pp. 123-130).

Foo, J. Y. A. (2011). A retrospective analysis of the trend of retracted publications in the field of biomedical and life sciences. Science and engineering ethics, 17(3), 459-468.

Fu, Y., & Schneider, J. (2020, August). Towards knowledge maintenance in scientific digital libraries with the keystone framework. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 217-226).

Gabehart, M. E. (2005). An analysis of citations to retracted articles in the scientific literature.

Garfield, E. (1965, December). Can citation indexing be automated. In Statistical association methods for mechanized documentation, symposium proceedings (Vol. 269, pp. 189-192). National Bureau of Standards, Miscellaneous: Publication.

Gasparyan, A. Y., Ayvazyan, L., Akazhanov, N. A., & Kitas, G. D. (2014). Self-correction in biomedical publications and the scientific impact. Croatian medical journal, 55(1), 61.

Goyal, S. (2022, January 6). Evaluation Metrics for Classification Models - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69

Greitemeyer, T. (2014). Article retracted, but the message lives on. Psychonomic bulletin & review, 21(2), 557-561.

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. PloS one, 7(10), e44118.

Heneberg, P. (2016). From excessive journal self-cites to citation stacking: Analysis of journal self-citation kinetics in search for journals, which boost their scientometric indicators. PloS one, 11(4), e0153730.

Hernandez-Alvarez, M., Soriano, J. M. G., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. Natural Language Engineering, 23(4), 561-588.

Jan, R., Bano, S., & Mehraj, M. (2018). Context Analysis of Top Seven Retracted

Articles: Should Retraction Watch Revisit the List? Context.

Jha, R., Jbara, A. A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. Natural Language Engineering, 23(1), 93-130.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

Kambria, K. (2019, August 27). Logistic Regression for Machine Learning and Classification. Kambria. https://blog.kambria.io/logistic-regression-for-machine-learning/

Koblah, D. S., Acharya, R. Y., Capecci, D., Dizon-Paradis, O. P., Tajik, S., Ganji, F., & Forte, D. (2022). A Survey and Perspective on Artificial Intelligence for Security-Aware Electronic Design Automation. ACM Transactions on Design Automation of Electronic Systems (TODAES).

Kochan, C. A., & Budd, J. M. (1992). The persistence of fraud in the literature: the Darsee case. Journal of the American Society for Information Science, 43(7), 488-493.

Kuzminykh, N. (2020, October 23). Sentiment Analysis in Python with TextBlob. Stack Abuse. https://stackabuse.com/sentiment-analysis-in-python-with-textblob/

Li, B., & Han, L. (2013, October). Distance weighted cosine similarity measure for text classification. In International conference on intelligent data engineering and automated learning (pp. 611-618). Springer, Berlin, Heidelberg.

Li, X., He, Y., Meyers, A., & Grishman, R. (2013, September). Towards fine-grained citation function classification. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 402-407).

Liu, H., Christiansen, T., Baumgartner, W. A., & Verspoor, K. (2012). BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. Journal of biomedical semantics, 3(1), 1-29.

Munawar, M. R. (2022, April 10). What is Data preprocessing, Why Do We Need That❓ - Nerd For Tech. Medium. https://medium.com/nerd-for-tech/what-is-data-preprocessing-why-we-need-that-2846b8b04bc4

Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18(60), 1-8.

Nath, S. B., Marcus, S. C., & Druss, B. G. (2006). Retractions in the research

literature: misconduct or mistakes? Medical Journal of Australia, 185(3), 152-154.
Neale, A. V., Northrup, J., Dailey, R., Marks, E., & Abrams, J. (2007). Correction and use of biomedical literature affected by scientific misconduct. Science and engineering ethics, 13(1), 5-24.
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2), 1-135.
Piller, C. (2021). Disgraced COVID-19 studies are still routinely cited.
Quick start: Build and deploy a machine learning model in a Jupyter notebook. (2022). IBM. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=gsbdtm-build-deploy-machine-learning-model-in-jupyter-notebook
Redman, B. K., Yarandi, H. N., & Merz, J. F. (2008). Empirical developments in retraction. Journal of medical ethics, 34(11), 807-809.
Resnik, D. B., Wager, E., & Kissling, G. E. (2015). Retraction policies of top scientific journals ranked by impact factor. Journal of the Medical Library Association: JMLA, 103(3), 136.
Retraction Watch. (2021a). Database search; retractions, 2000 to 2020. Retrieved December 14, 2021a from http://retractiondatabase.org/RetractionSearch.aspx
Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508-524). Springer, Berlin, Heidelberg.
Santos-d'Amorim, K., de Melo, R. R., & dos Santos, R. N. M. (2021). Retractions and post-retraction citations in the COVID-19 infodemic: is Academia spreading misinformation? Liinc em Revista, 17(1), e5593-e5593.
Schneider, J., Ye, D., Hill, A. M., & Whitehorn, A. S. (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. Scientometrics, 125(3), 2877-2913.
Sheth, B. P., & Thaker, V. S. (2014). Scientific retraction: a synonym for pseudoscience? Acta bioethica, 20(1), 93-97.
Shuai, X., Rollins, J., Moulinier, I., Custis, T., Edmunds, M., & Schilder, F. (2017). A multidimensional investigation of the effects of publication retraction on scholarly impact. Journal of the Association for Information Science and Technology, 68(9), 2225-2236.
Steen, R. G. (2011). Retractions in the medical literature: how many patients are put

at risk by flawed research?. Journal of medical ethics, 37(11), 688-692.
Steen, R. G., Casadevall, A., & Fang, F. C. (2016). Why has the number of scientific retractions increased?.
Suelzer, E. M., Deal, J., Hanus, K. L., Ruggeri, B., Sieracki, R., & Witkowski, E. (2019). Assessment of citations of the retracted article by Wakefield et al with fraudulent claims of an association between vaccination and autism. JAMA network open, 2(11), e1915552-e1915552.
Suelzer, E. M., Deal, J., Hanus, K. L., Ruggeri, B., Sieracki, R., & Witkowski, E. (2019). Assessment of citations of the retracted article by Wakefield et al with fraudulent claims of an association between vaccination and autism. JAMA network open, 2(11), e1915552-e1915552.
Teixeira da Silva, J. A., & Bornemann-Cimenti, H. (2017). Why do some retracted papers continue to be cited?. Scientometrics, 110(1), 365-370.
Teixeira da Silva, J. A., & Dobránszki, J. (2017). Highly cited retracted papers. Scientometrics, 110(3), 1653-1661.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 103-110).
Theis-Mahon, N. R., & Bakker, C. J. (2020). The continued citation of retracted publications in dentistry. Journal of the Medical Library Association: JMLA, 108(3), 389.
Trikalinos, N. A., Evangelou, E., & Ioannidis, J. P. (2008). Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. Journal of clinical epidemiology, 61(5), 464-470.
Usman, M., Mustafa, G., & Afzal, M. T. (2021). Ranking of author assessment parameters using Logistic Regression. Scientometrics, 126(1), 335-353.
Usman, M., Shafique, Z., Ayub, S., & Malik, K. (2016). Urdu text classification using majority voting. International Journal of Advanced Computer Science and Applications, 7(8).
Van Der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Research integrity and peer review, 1(1), 1-10.

Van Der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Research integrity and peer review, 1(1), 1-10.

Van Der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Research integrity and peer review, 1(1), 1-10.

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988–2008. Journal of medical ethics, 37(9), 567-570.

Wallstreetmojo, W. (2022, August 30). Correlation Matrix. WallStreetMojo. https://www.wallstreetmojo.com/correlation-matrix/

Wang, L. (Ed.). (2005). Support vector machines: theory and applications (Vol. 177). Springer Science & Business Media.

Wikipedia contributors. (2022, October 23). Retraction Watch. In Wikipedia, The Free Encyclopedia. Retrieved 09:52, January 3, 2023, from https://en.wikipedia.org/w/index.php?title=Retraction_Watch&oldid=1117803784

Yang, S., Qi, F., & Diao, H. (2020, August). Exploring the influence of publication retraction on citations in psychology science. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 501-502).

APPENDIX A.

XXXXXXXXXXXXXXXXXX

BIOGRAPHY

Ehtisham Kiyani

Ehtisham Kiyani have completed his bachelor's degree from Capital University of Science and technology Islamabad, Pakistan. He worked with different international clients as a freelancer. Moreover currently he is in the final semester of the Master's degree of computer engineering at Beykoz University, Turkey. His main goal is to be a motivational speaker and help the people to live their lives with joy and happiness.