Analyzing the Influence of Social Media Activity on the academic community

CS43-1

Final Report

5703 Group Based Capstone Project

Group Members

1. Member 1 Qirui Chen (480004321)
2. Member 2 Yuzhe Zhou (53069186)
3. Member 3 Linfeng Yu (530366706)
4. Member 4 Qian Yu (530458816)
5. Member 5 Lin Zhang (530207959)
6. Member 6 Yutong Wu (530717300)

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

19 April 2025

CONTRIBUTION STATEMENT

Our group, taking project CS43-1, with group members Qirui Chen, Yuzhe Zhou, Linfeng Yu, Qian Yu, Yutong Wu, Lin Zhang, would like to state the contributions each group member has made for this project during this semester:

- group member 1 name: Qirui Chen (Provide team members with appropriate data sets, group sampling of a data set to reduce the data set to a readable size and participate in predictive model design and data visualization)
- group member 2 name: Yuzhe Zhou (EDA, research about the relationship between exposure and the number of citation and output the visualization about the result, create the flowchart of the project.)
- group member 3 name: Linfeng Yu (Perform exploratory data analysis on data, clean and balance it, contribute to creating predictive models and finding the best hyper-parameter manually.)
- group member 4 name: Qian Yu (detailed contributions during whole semester)
- group member 5 name: Yutong Wu (Responsible for data cleaning, modeling, and analysis in the prediction model. Designed the project framework, assigned tasks, communicated with the tutor and client for requirement clarity, coordinated resources, and led team collaboration.)
- group member 6 name: Lin Zhang (Led the initial feature engineering for network dataset, conducted comprehensive model comparisons and performance evaluations, performed hyperparameter tuning for optimization, and handled part of video production for final presentation.)

All group members agreed on the contributions listed on this statement by each group member.

Signatures:

Abstract

This project investigates the influence of social media engagement on academic visibility, focusing on platforms like Twitter and LinkedIn and their impact on scholarly citation rates. With social media reshaping traditional citation metrics, understanding its role in scholarly impact is increasingly important. This study aims to validate the link between social media exposure and citation growth, identify influential social media factors affecting citations, and develop predictive models to guide researchers in optimizing their work's online visibility. Key methodologies include citation data collection, exploratory data analysis, feature selection, and machine learning modeling. The findings will offer actionable insights, helping researchers effectively leverage social media to expand their audience reach and enhance their academic impact.

Contents

CHAPTER 1

Introduction

Social media is becoming more and more significant in the academic community in today's digital environment. Platforms like LinkedIn, X (previously Twitter), and TikTok have altered the way people communicate information, impacting both the academic community and public conversations. The frequency of citations in other academic works has historically been used to gauge the significance of academic research (Donelan 2016). One important metric for assessing a paper's significance and impact is its citation count. However, as social media plays a bigger role in disseminating research, it's critical to investigate how these platforms are altering academic practices and impacting study exposure and reach.

Motivation

Citation counts are only one aspect of the connection between social media and the academic community. Citations are still crucial, but social media also aids in the promotion of collaboration across disciplines, the development of academic networks, and the visibility of new research (Sugimoto et al. 2017). Knowing how to use social media effectively is essential for researchers, especially those who are just starting out, to ensure that the correct people see their work and that it receives the credit it merits (Priem and Costello 2010).

Benefits

Observing that many researchers struggle to use social media to make their work more visible gave rise to the concept for this project. This is particularly true for up-and-coming writers who might not have a large following or network. Even excellent research can be overlooked if it is not sufficiently publicized, which delays its impact and recognition (Wouters et al. 2019). This study intends to provide scholars with useful advice on how to utilize social media to promote their articles by examining how social media activities affect academic visibility.

Problem Description

Monitoring citation counts as a gauge of performance isn't the only goal of this endeavor. Our goal is to identify the most effective methods for boosting a paper's social media presence, including when to publish, which platforms to utilize, how to interact with various audiences, and how to work with subject-matter experts or influencers. In doing so, we intend to offer practical guidance that will help researchers—particularly those who are publishing new papers—reach a larger audience and make a greater impact.

Proposed Solution

Solving this problem involves more than just assisting individual scholars; it also entails developing a more vibrant and welcoming academic community. Effective use of social media by researchers allows them to connect with people outside of their immediate academic circles (Veletsianos and Kimmons 2012). This may result in more varied conversations, interdisciplinary idea exchanges, and increased chances for cooperation. Ultimately, knowing how to use social media effectively can help open up the academic process and guarantee that significant research is seen by those who stand to gain the most from it.

In overall, by giving researchers the resources and techniques they require to successfully market their work online, this project seeks to close the gap between social media and academic practices. By emphasizing the ways in which social media exposure can increase academic visibility, we intend to enable researchers to take full advantage of the digital resources at their disposal and guarantee that their knowledge-contributions are acknowledged and appreciated both inside and outside of the academic community.

CHAPTER 2

Related Literature

2.1 Literature Review

Title	**Dataset**	**Approach**	**Venue**	**Results**	**Year**	**Other Relevant Details**
If I tweet will you cite later? Follow-up on the effect of social media exposure on article downloads and citations(Tonia et al. 2020)	130 articles from the International Journal of Public Health (2012-2014)	A controlled study examining how exposure to social media (Twitter, Facebook, blog posts) affects the number of downloads and citations of articles	International Journal of Public Health	Simple social media promotion cannot significantly increase the number of downloads and citations of papers, and traditional impact indicators may not fully reflect the value of social media promotion. Future research should focus on the different functions of social media and their long-term effects on papers	2020	Implies that the impact of social media may not be fully captured by traditional citation metrics
Longitudinal relationship between social media activity and article citations in the journal Gastrointestinal Endoscopy(Smith et al. 2019)	Data regarding journal articles published in GIE from 2000 to 2016 publication status, number of citations per article, and social media exposure per article using Altmetric data were collected from the publisher.	They analyzed the exposure on Twitter of all articles published in GIE magazine in 2012, and explored the correlation between the number of Facebook posts and Mendeley readers and the citation rate of the articles.	The American Society for Gastrointestinal Endoscopy	Articles that are retweeted on Twitter are 14 times more likely to be cited than articles that are not retweeted. In addition, the number of posts on Facebook and the number of readers on Mendeley are also related to the increase in article citation rate, but the correlation is weaker than that on Twitter.	2019	The authors suggest that further randomized controlled trials (RCTs) can be conducted to evaluate the impact of different levels of social media exposure on the citation rate of a single article to better assess the causal relationship.

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
The Use of Social Media to Increase the Impact of Health Research: Systematic Review(Bardus et al. 2020)	They searched the Medical Literature Analysis and Retrieval System Online (MEDLINE), Excerpta Medica dataBASE (EMBASE), and Cumulative Index to Nursing and Allied Health Literature (CINAHL) databases using a predefined search strategy (International Prospective Register of Systematic Reviews: CRD42017057709).	Automatically or manually post on the target journal's Twitter or Facebook account, taking advantage of organic (free) distribution on social networks. Use ads on Facebook to increase the visibility of your posts (i.e. "boost content")	The Journal of Medical Internet Research	Most correlation studies show a positive correlation between traditional metrics (such as citations) and social media metrics (such as mentions)	2019	Further and better designed studies are needed to establish causal relationships between social media effects and research effects
The Patterns and Impact of Social Media Exposure of Journal Publications in Gastroenterology(Chiang et al. 2021): Retrospective Cohort Study	The 3-year citations of all full-length articles published in five major gastroenterology journals from January 1, 2012, to December 31, 2012, tweeted by official journal accounts with those that were not.	Determine the engagement patterns of publications in gastroenterology journals on Twitter and evaluate the impact of tweets on citations	The Journal of Medical Internet Research	There was significant association between article type and number of retweets on analysis of variance (ANOVA) (P<.001), with guidelines/technical reviews (mean difference 1.04, 95% CI 0.22-1.87; P<.001) and meta-analysis/systematic reviews (mean difference 1.03, 95% CI 0.35-1.70; P<.001) being retweeted more than basic science articles.	2021	Wider adoption of social media to increase reach and measure uptake of published research should be considered.
Universality of Citation Distributions: Toward an Objective Measure of Scientific Impact(Radicchi et al. 2008)	Citation data from multiple scientific disciplines.	Statistical analysis of citation distributions across various disciplines.	*Proceedings of the National Academy of Sciences (PNAS)*	Found that citation distributions are universal across disciplines, suggesting a standardized measure of scientific impact.	2008	Proposes that normalized citation distributions can be used to objectively compare scientific impact across different fields.

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
Prediction Methods and Applications in the Science of Science: A Survey(Hou et al. 2019)	Scholarly data from various sources including AMiner, APS, DBLP, and MAG datasets.	Proposes that normalized citation distributions can be used to objectively compare scientific impact across different fields.	*Computer Science Review*	No results in this article, the author discusses the traditional prediction for Science of Science and at the last part points out some issues (e.g., differences in research fields and the impact of self-citation)	2019	Highlights the challenges and open issues in the field of prediction in the science of science, including the use of deep learning for paper impact prediction.
A Supervised Learning Method for Prediction Citation Count of Scientists in Citation Networks(Bütün et al. 2017)	12 sampled datasets extracted from two citation networks (Aminer and HEP-Th).	The problem of predicting future citation counts of scientists is formulated as a link prediction problem in directed, weighted, and dynamic citation networks. The method introduces a dynamic similarity metric and uses topological features in classifiers.	2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	The proposed method performs well in predicting future citation counts, particularly when considering dynamic networks.	2017	The study emphasizes the importance of using graph structures in citation networks to enhance the accuracy of citation prediction.
Which can better predict the future success of articles? Bibliometric indices or alternative metrics(Wang et al. 2019)	617 scientific articles published in seven journals from Public Library of Science (PLOS).	A machine learning framework was established to predict the future success of articles using 23 bibliometric and alternative indices. Feature selection techniques such as Relief-F, PCA, and EWM were used, and classifiers like Naïve Bayes, KNN, and Random Forest were employed.	*Journal of Informetrics*	Both bibliometric indices and alternative metrics were found to be beneficial in predicting the future success of articles. Early citation features, early web usage statistics, and the reputation of the first author were the most valuable indicators.	2019	The study concluded that combining traditional bibliometric indices with alternative metrics provides a more comprehensive profile for predicting the future success of articles.

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
Tweets to Citations: Unveiling the Impact of Social Media Influencers on AI Research Visibility(Weissburg et al. 2024)	Over 8,000 papers shared by two AI/ML Twitter influencers (AK and Aran Komatsuzaki) from December 2018 to October 2023	The study used a matched-pair design, comparing papers shared by influencers to control papers matched on publication year, venue, and topic similarity. The researchers conducted citation analysis, geographic distribution analysis, and gender distribution analysis.	arXiv	Papers shared by influencers had significantly higher median citation counts (2-3 times higher) than control papers. No significant difference in review scores between shared and control papers, suggesting effective quality control in the matching process.	2024	Discusses implications of influencers acting as curators/gatekeepers for AI research visibility. Recommends maintaining diverse voices and perspectives in research dissemination.
Research output and visibility of librarians: Are social media influencers or distractors?(Adetayo 2023)	312 librarians from universities in southwestern Nigeria. Data collected through questionnaires	1. Descriptive survey research approach 2. Using questionnaires to collect data.	Journal of Librarianship and Information Science	1. Librarians have high research output but low research visibility. 2. Journal articles were the most commonly published type of research output.	2023	The study examined social media use among southwestern Nigerian university librarians for research. It found high research output but low visibility, with popular platforms like WhatsApp and Facebook used more than academic-specific ones.
Using social media to promote academic research: Identifying the benefits of Twitter for sharing academic work(Klar et al. 2020)	308 articles published in 2016 from 6 academic journals (3 political science, 3 communication). Gender, rank, department ranking, Twitter followers for 576 authors	1. Collected data on articles, authors, tweets, and citations 2. Used negative binomial regression models to analyse factors predicting number of tweets about articles	PLOS ONE	1. Articles tweeted about received more citations overall 2. No evidence of gender bias in likelihood of articles being tweeted 3. Solo-authored articles by women more likely to be tweeted than those by men	2020	Focused on political science and communication fields. Examined both original tweets and retweets. Considered interactions between author gender and number of authors

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
Does Tweeting Improve Citations? One-Year Results From the TSSMN Prospective Randomised Trial(Luc et al. 2021)	112 representative original scientific articles published from 2017-2018 in The Annals of Thoracic Surgery and The Journal of Thoracic and Cardiovascular Surgery	1. Prospective randomised trial 2. Articles randomised 1:1 to be tweeted via Thoracic Surgery Social Media Network (TSSMN) or a control (non-tweeted) group 3. Measured citations, Altmetric scores, and Twitter analytics at 1 year compared to baseline	The Annals of Thoracic Surgery	Tweeted articles showed significantly higher increases in Altmetric scores (9.4 vs 1.0, p<0.001), Altmetric percentiles (76.0 vs 13.8, p<0.001), and citations at 1 year (3.1 vs 0.7, p<0.001) compared to non-tweeted articles.	2021	The study was conducted by the Thoracic Surgery Social Media Network (TSSMN), a collaborative effort between leading cardiothoracic surgery journals. TSSMN delegates had a combined Twitter followership of 52,983 at the time of the study.
Early indicators of scientific impact: Predicting citations with altmetrics(Akella et al. 2021)	Altmetric data and citation counts from multiple sources	Built and tested various machine learning models (e.g., neural networks, ensemble models) to predict short-term and long-term citation counts using altmetric data.	Journal of Informetrics	Neural networks and ensemble models performed best for prediction.	2021	Found that Mendeley readership was the most critical factor in predicting early citations
Citation count prediction as a link prediction problem(Pobiedina and Ichise 2016)	Citation data from academic publications, modeled as a citation network for analysis.	The study employs graph pattern mining techniques to predict citation counts, treating the task as a link prediction problem within the citation network.	Applied Intelligence in 2016	The introduction of a new feature based on frequent graph patterns significantly improved citation prediction accuracy compared to traditional methods.	2016	This paper is particularly relevant for researchers looking to enhance citation prediction models, offering a method that aligns with altmetric studies but focuses on citation networks.
Do altmetrics work? Twitter and ten other social web services(Thelwall et al. 2013)	Altmetric data from Twitter and ten other social web services (e.g., Mendeley, Facebook)	Quantitative analysis of correlations between social media mentions and traditional citations	PLOS ONE	Found moderate to strong correlations between certain altmetrics and citation counts	2013	Highlighted limitations in data coverage and discipline differences

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
ComLittee: Literature Discovery with Personal Elected Author Committees(Kang et al. 2023)	Academic literature databases such as Semantic Scholar, combined with user feedback data and citation networks	The system uses an author-centric recommendation algorithm, combining co-authorship and citation data to optimize literature discovery dynamically.	2023 CHI Conference on Human Factors in Computing Systems (CHI '23)	The system demonstrated enhanced efficiency in discovering new authors and papers, with improved user satisfaction compared to traditional paper-centric systems.	2023	The ComLittee system is particularly effective for researchers in rapidly developing fields, aiding in the discovery and tracking of emerging research trends with a personalized approach.
Predicting citations from mainstream news, weblogs, and discussion forums (Timilsina et al. 2017)	Mentions of academic articles in mainstream news, weblogs, and discussion forums, analyzed in relation to subsequent citation counts.	The study introduces graph-based influence metrics and the "EgoMet score" to measure the impact of social media mentions on academic citations.	2017 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2017) held in Leipzig, Germany from August 23 to 26, 2017.	The study found that discussions on social media platforms can significantly influence the visibility and citation outcomes of scholarly work.	2017	This research highlights the importance of considering social media discussions in citation prediction models, offering a complementary perspective to altmetric studies.
Social media usage to share information in communication journals: An analysis of Twitter mentions and academic citations(Özkent 2022)	Articles published in 2018 from the top ten communication-based journals	The study conducted a retrospective cross-sectional analysis. Various statistical analyses, including correlation and Mann-Whitney U tests, were performed to compare the citation rates of tweeted versus non-tweeted articles.	PLOS ONE	Articles that are exposed on social media (especially Twitter) have significantly higher citation rates than those that are not exposed.	2022	The study was designed to focus on communication science journals with a Q1 quartile index and an impact factor greater than 2. Twitter mentions are positively correlated with academic citations.
To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers(Ortega 2016)	4,166 articles from 76 Twitter users and 124 articles from non-Twitter users.	Regression analysis was used to compare tweet and citation patterns of papers authored by Twitter users and non-Twitter users.	Scientometrics	While being active on Twitter can increase the dissemination of research papers, it does not necessarily mean higher citation counts.	2016	The study questions the effectiveness of social media in enhancing traditional academic impact, such as citation rates.

Title	Dataset	Approach	Venue	Results	Year	Other Relevant Details
How do scientific papers from different journal tiers gain attention on social media?(Cao et al. 2023)	The dataset consists of 170,862 scientific papers from various journals, including 35,195 papers from elite journals and 47,992 papers from non-elite journals.	Using complex network analysis and time series analysis, the study explores the dynamic diffusion patterns of papers across these tiers.	Information Processing and Management	Elite journal papers typically spread faster, deeper, and more widely than non-elite journal papers. However, non-elite journal papers can achieve considerable impact through high-impact users.	2023	The study also discusses implications for science communication and the potential for increasing the visibility of scientific research through social media.
The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations)(Ortega 2017)	The study used data from 4,176 articles in 350 journals, with information from Plum Analytics.	Student independent sample t-tests and regression analyses assessed the relationship between Twitter activity and the number of tweets and citations received by the research paper.	Aslib Journal of Information Management	Journals with their own Twitter accounts received 46% more tweets and 34% more citations than those without Twitter accounts. However, Twitter did not have as much of an impact on citations as it did on tweets. The study found that the number of followers on Twitter was the most important factor in increasing tweets and citations, although the overall impact was small	2017	The study concluded that having a dedicated Twitter account is the best strategy for journals to increase visibility of their articles.

TABLE 2.1. Literature Review on Social Media Influence in Academia

CHAPTER 3

PROJECT PROBLEMS

3.1 Project Aims & Objectives

Our project's success will be determined by achieving the following objectives, each directly tied to the project scope:

- **Delivering a comprehensive research report**: Nowadays, the Internet and social media have become the most significant parts of communication research, which are on par with traditional media such as television or newspapers (Günther and Domahidi 2017). As Yasemin Özkent mentioned "Social media research is encouraged in the field of communication because people nowadays present themselves through digital network platforms." (Özkent 2022). Therefore, this report will be based on processed data analysis and validate the relationship between social media exposure and citation counts of articles.
- **Developing a recommendation model using social network datasets**: We will use social network data to build and optimize a recommendation model that identifies key patterns and factors impacting citation counts. Different empirical studies have shown that it is possible to predict new relationships between elements attending to the topology of the network and the properties of its elements. The problem of predicting new relationships in networks is called link prediction. (Martinez2016) Link prediction finds missing links (in static networks) or predicts the likelihood of future links (in dynamic networks). Link prediction is a fast-growing research area in both physics and computer science domain. (Kumar2020)It can be observed that more and more papers pay attention to link prediction in social networks, especially in the last five years, there are thousands of papers related to this problem every year. Another interesting phenomenon is that the problem of link prediction also attracts attention from different disciplines.(Wang2014)

- **Developing and applying a machine learning model**: Link prediction in sparse networks presents a significant challenge due to the inherent disproportion of links that can form to links that do form. Previous research has typically approached this as an unsupervised problem. (Lichtenwalter2010) We will train at least five different machine learning models, select the one with the highest F1-score, and use it to recommend a social media strategy for a recently published paper, thereby demonstrating the model's practical application.

By aligning the scope with the success criteria, we ensure that each aspect of the project directly contributes to our overall goals.

3.2 Project Questions

Our task is to clearly define the core problem faced by the client: understanding how social media can be leveraged to enhance the citation count of their research papers and devising a strategic plan to achieve this goal. We will focus on identifying the key factors on social media that influence citation counts and use this knowledge to help the client increase the citation count of their papers effectively.

3.3 Project Scope

Our project will focus on three key areas:

- **Verifying the relationship between social media exposure and citation counts**: We semantically match the keywords of the article with the hot words in the industry every year to observe how the articles with hot words of different degrees are cited.
- **Developing a recommendation model using social network datasets**: By analyzing social network data, we will create a recommendation model that identifies key patterns and attributes influencing citation counts, helping researchers optimize their work for better visibility.
- **Building a predictive model**: We will develop a model using various social media and paper-related attributes to predict future citation trends. This model will help us determine which key variables are most effective in increasing citations, providing researchers with actionable insights for enhancing their paper's visibility.

CHAPTER 4

METHODOLOGIES

4.1 Methods

We intend to use a combination of data collection, exploratory data analysis, causation analysis, correlation analysis, traditional statistical prediction models, machine learning prediction models, and hypothesis testing in order to address the issue described in section 3.1 (Project Aims & Objectives). The process described in section 4.3 (Data Analysis) will specify the logical order in which these techniques will be used. Sections 5.1 (Hardware & Software) and 5.2 (Materials) go into detail on the particular tools and libraries chosen to enable this methodology.

4.2 Data Collection

The original data is from https://www.aminer.org/citation. Due to the large size of the dataset and the limitations of our computational power, we were unable to use this dataset for model prediction and further work. Therefore, we decided to perform sampling on the data (Lohr 2019). We chose stratified sampling to reduce the data size. This method divides the entire dataset into several different strata or categories, and by drawing samples from each stratum, the sample can more accurately reflect the overall structure. Hence, we first examined the distribution of citation counts to group the data accordingly.

As shown in the figure, we decided to divide the citation counts into five groups: "1-15", "16-25", "26-50", "51-200", and "200+". We set the sampling rate to 0.1, meaning we will extract 10% of the data from each stratum. To further reduce the data size, we also performed dimensionality reduction. We retained only six features: year, keywords, title, abstract, n_citation, and references (Jolliffe and Cadima 2016). Then, we performed 10% sampling based on these five groups, resulting in a new, manageable JSON dataset.

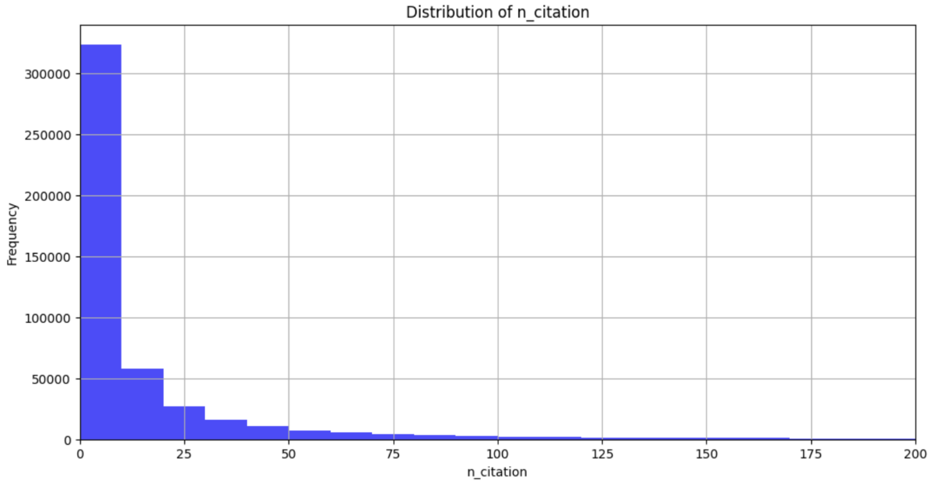


FIGURE 4.1. Distribution of Citation

4.3 Data Analysis

We employ a range of methods in our data analysis to look closely at the dataset. This flow chart's subsequent procedure. To gain a better understanding of the features of citation data, we carried out a pilot research on paper citations prior to starting the official analysis of our target data. Similar techniques will be used in the following steps to conduct feature engineering and exploratory data analysis (EDA) depending on the properties of the data.

Pilot study

In order to ensure consistency, we will first gather data from citation count systems, making sure the data is cleaned and preprocessed. For exploratory data analysis, we select 1500 data points at random from the entire dataset. We mostly concentrate on the desired variable "number of citations" during the EDA demo. To summarize important characteristics like mean, median, standard deviation, and skewness—which show the concentration trends, variability, and distribution of the data—we first created descriptive statistics (Figure 4.2). We discovered from the statistics section that the distribution of our target variable is extremely skewed. In this situation, using log transformation is a good technique to bring the distribution closer to normal because many publications get few or no citations, while only a small number of papers receive many (Lee 2020). To evaluate the quality of the data, we also looked at outliers, zeros, and missing numbers.

Stats | Histogram | KDE Plot | Normal Q-Q Plot | Box Plot | Value Table

Overview

Approximate Distinct Count	139
Approximate Unique (%)	13.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	16000
Mean	2.2142
Minimum	0
Maximum	9.6601
Zeros	146
Zeros (%)	14.6%
Negatives	0
Negatives (%)	0.0%

Descriptive Statistics

Mean	2.2142
Standard Deviation	1.5117
Variance	2.2853
Sum	2214.2236
Skewness	0.3637
Kurtosis	-0.09153
Coefficient of Variation	0.6827

Quantile Statistics

Minimum	0
5-th Percentile	0
Q1	1.0986
Median	2.1972
Q3	3.2581
95-th Percentile	4.7875
Maximum	9.6601
Range	9.6601
IQR	2.1595

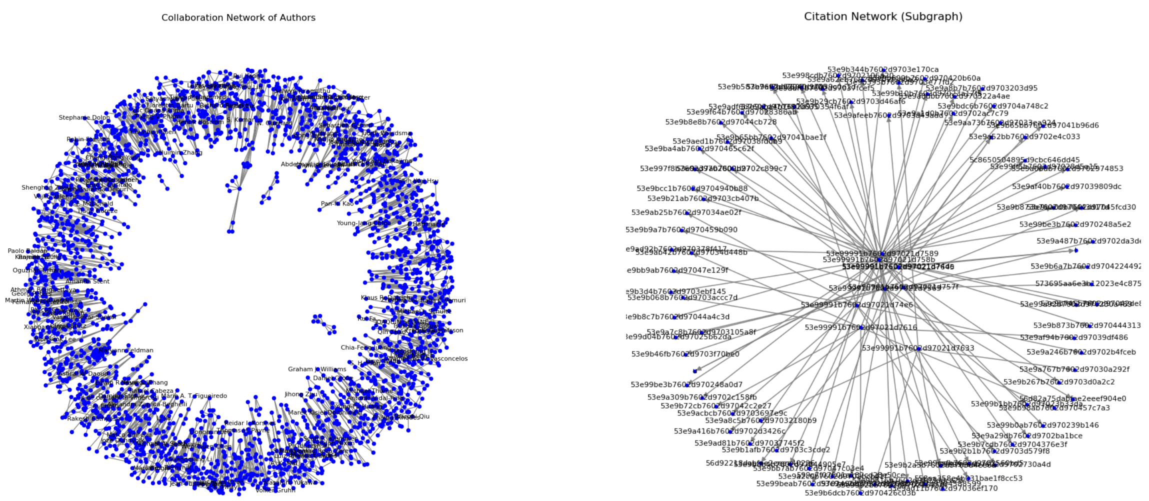
FIGURE 4.2. Statistics Summary

We used bar charts and line graphs to examine the growth of published papers and the distribution of citations in order to comprehend citation trends over time. We were able to identify some unique patterns that we can utilize in our future research. To see the connections between publications and find important academic works, we also carried out citation network analysis. Co-authorship patterns were also uncovered by collaborative network analysis, emphasizing the impact of these connections on citation counts.

The Chisquare test and MultiLabelBinarizer are used to separate the strings in the list and determine the correlation of each one with the number of citations. Lastly, we attempted to use one hot to convert categorical data to numerical data, which can facilitate the use of correlation heatmap and PCA. However, since the type of several attributes is a string list, a standard one hot will not be helpful. Together, these methods offer a thorough comprehension of the information and direct our evaluation of social media's influence.

An author cooperation network is depicted in Figure 4.3 A. An author is represented by each blue dot, and author collaborations, such as co-authorship of articles or cooperative research initiatives, are represented by the lines joining the dots. A decentralized collaboration pattern is indicated by the circular form, which implies a dense and extensive network of collaboration without a clear central hub. While some clusters show regular cooperation among particular groups, the majority of authors are linked to one another, creating a vast, interconnected network. This arrangement makes it easier to see how researchers from different academic disciplines or in a particular topic collaborate overall.

A citation network subgraph is shown in Figure 4.3 B. A publication or paper is represented by each node, and a citation relationship—where one document cites another—is indicated by each line. Highly cited publications that are regarded as influential in their field—possibly fundamental studies or thorough reviews—are suggested by central nodes with numerous connections. Papers with fewer citations or those mentioning only a few others may be represented by peripheral nodes with fewer connections. The structure of the graph identifies clusters—subfields or specialized study groups—where papers regularly cite one another. Understanding the flow of knowledge and the most influential articles within a field of study is made easier with the aid of this citation network visualization.



(A) Collaboration Network (B) Citation Network

FIGURE 4.3. Comparison of Collaboration and Citation Networks

Feature Engineering(pilot study)

We use heat maps(Figure 4.4) to observe the correlation and importance of each feature. Then select the features we need to conduct experiments.

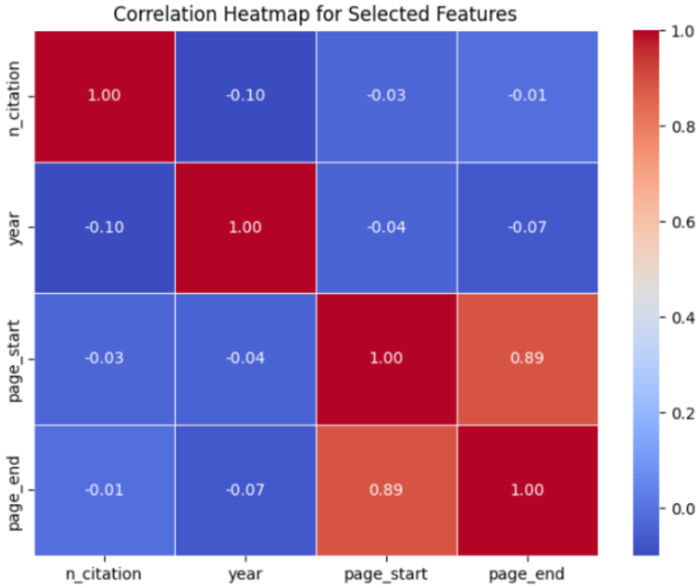


FIGURE 4.4. heat map

And we will also use neural network models to handle complex nonlinear relationships. Help us build more complex prediction models.

Exploratory Data Analysis and Feature Engineering (The full data after data collection)

Although the data processing will differ slightly based on each research question, we applied the same initial data processing steps. Using the same dataset ensures that we can analyze the data from different dimensions consistently.

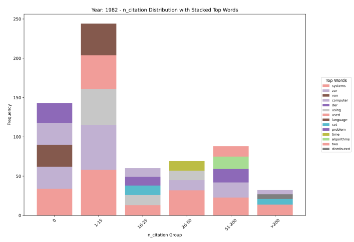
Missing Value

- For some basic numerical missing values, we filled them with 0 as a standard approach. Additionally, since the goal of our research is to help improve citation counts, we decided to remove data where the citation count is 0.
- After performing Exploratory Data Analysis from pilot study, we found that the variables containing critical information: keyword, title, and abstract. They still

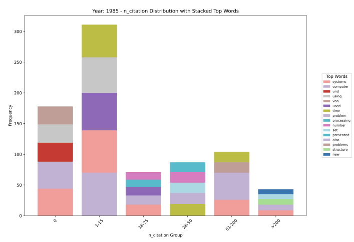
had a large amount of missing data. According to several research papers, there is a strong correlation between these three variables (Garcia et al. 2019). Based on this finding, we decided to combine keyword, title, and abstract into one variable, called combined text. During the merging process, we aimed to clean the combined text by removing words that are not useful for predicting the response. To achieve this, we used nltk (Wang and Hu 2021) and a list of common stopwords provided by our professor to eliminate unnecessary conjunctions and words. Additionally, we removed duplicate words that appeared in keyword, title, and abstract to ensure the efficiency of the combined text.

Visualization

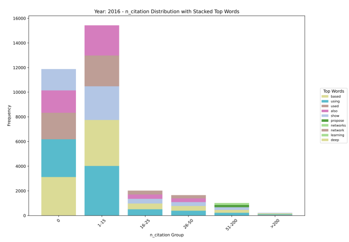
- To identify the trends in keywords for each year, we visualized the citations in the sampled data (He 1999). First, we combined the keywords, abstracts, and titles, and removed any duplicate words. Then, using the nltk package, we removed stop words and empty strings. After that, we manually filtered out additional non-essential words for further refinement. Next, we grouped the data based on the number of citations: "1-15", "16-25", "26-50", "51-200", and "200+". Once the grouping was complete, we plotted the top five most frequent words for each citation group by year.



(A) Example 1 of key words Distribution



(B) Example 2 of key words Distribution



(C) Example 3 of key words Distribution

FIGURE 4.5. Examples of key words Distribution

The following are three examples of visualizations; all visualizations will be presented at the end of the report.

Balance Data

- Based on the results of our data collection, we classified the data into different ranges and visualized them to check for balance. As shown in the image on the left, after categorizing the data based on different citation counts, we observed that there was a disproportionately large amount of data in category 0 (very low citations), while category 4 (very high citations) had significantly fewer data points. Therefore, we decided to downsample the data in category 0 and upsamplethe data in category 4. As shown in the image, we made sure to preserve the original distribution as much as possible, so we did not downsample category 0 too aggressively.

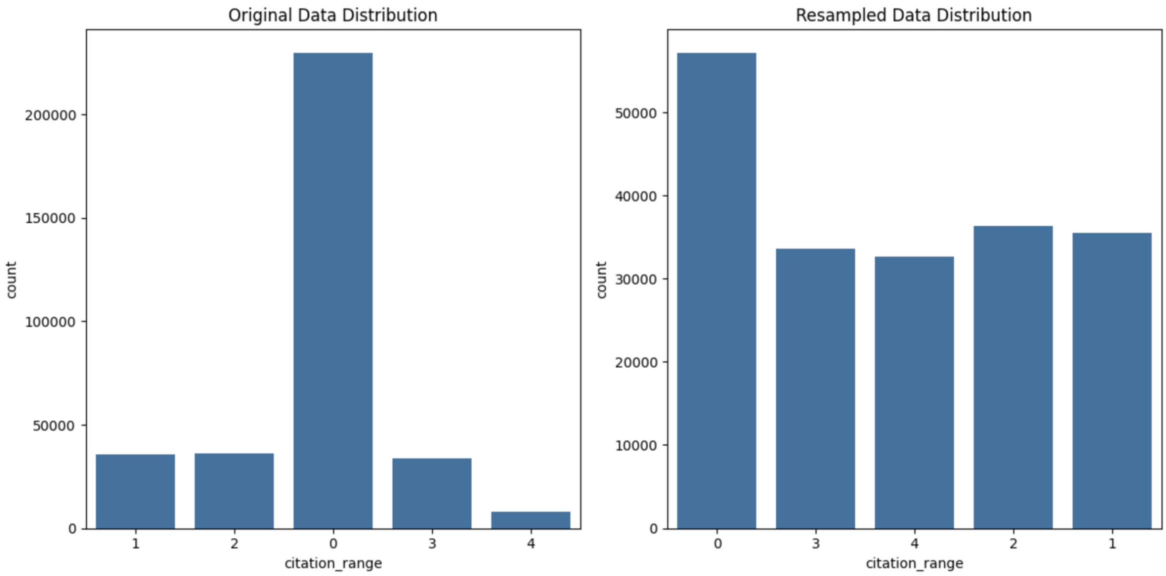


FIGURE 4.6. Banlanced Data

Model Selection

Based on the scope outlined in section 3.3, our model selection will be tailored to address three main areas.We also provided a flow chart for each area to better understand our model:

(1) **Relationship between exposure and citation number analysis**: Since we need to find the relationship between exposure and citation, we need to match the keywords of each paper in the dataset with the current hot words of different levels in the generated industry library. As a variant model of BERT, `paraphrase-MiniLM-L6-v2` can solve the problem of semantic matching efficiently and lightly, and calculate the **cosine similarity** of the matching to filter out data with higher matching degree.

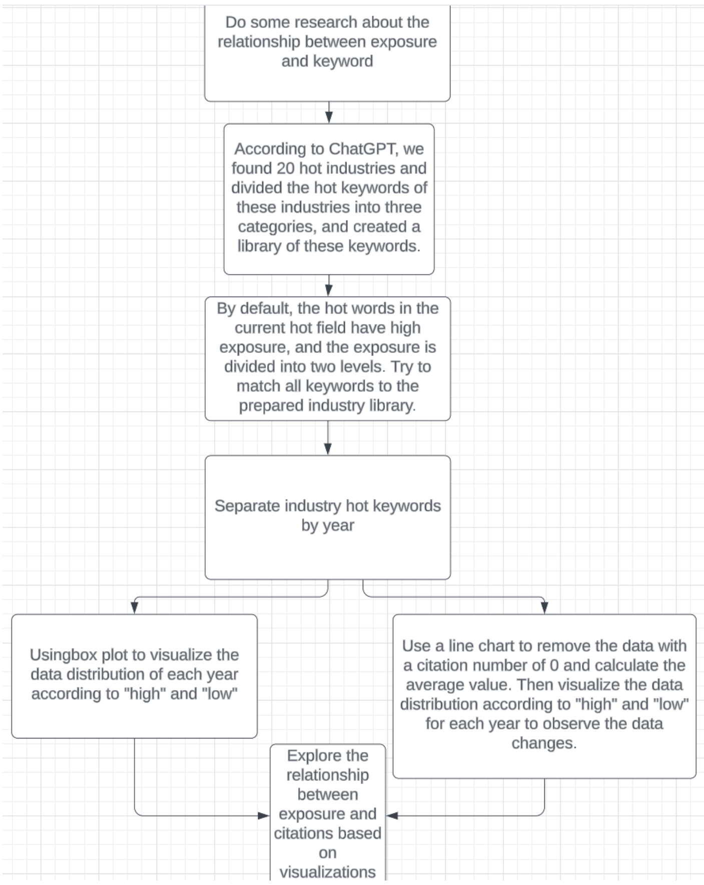


FIGURE 4.7. Relashipship Between Exposure And Citation Model Introduction

(2) **Developing a recommendation model using social network datasets**: To build a recommendation model, we will utilize `Random Forest` and `XGBoost` models, which are well-suited for handling large-scale social network data. These models will help us predict and recommend optimal strategies for improving citation counts based on social media exposure and other relevant attributes.

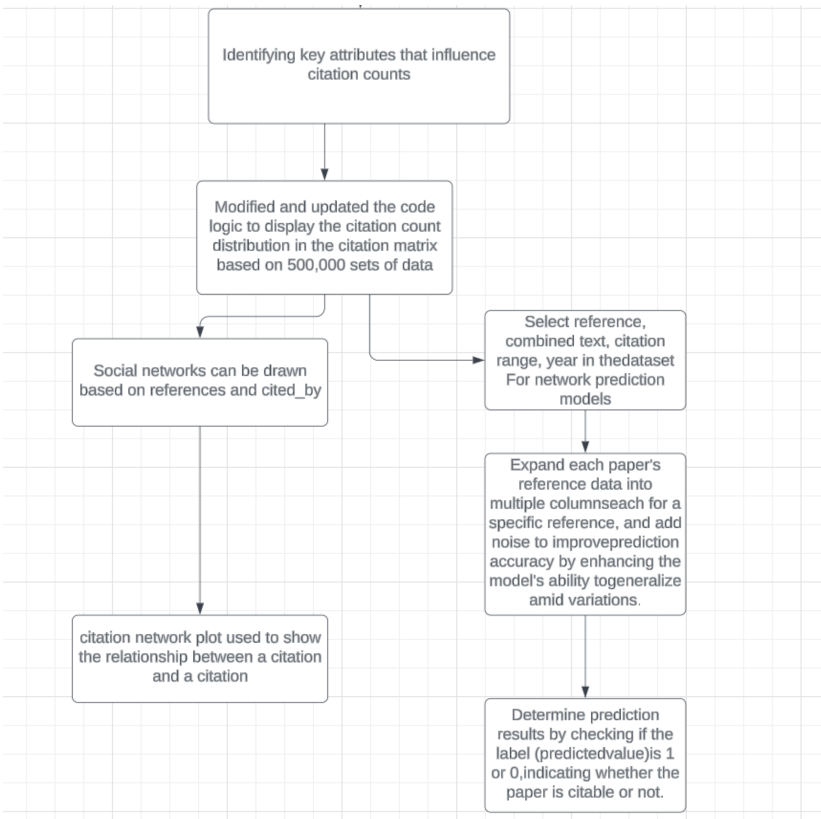


FIGURE 4.8. recommendation model using social network

(3) **Regression and classification model for Prediction citation counts**: Since our key explanatory variable is `combined text`, which is unstructured data, we applied Natural Language Processing (NLP) techniques using the pre-trained `RoBERTa model` for feature extraction and transformation. After extracting the features from the text using RoBERTa, we developed both traditional machine learning models and deep learning models to handle prediction tasks.

For the **regression tasks**, we used traditional machine learning models, including `Ridge Regression`, `Random Forest`, and `XGBoost`, to predict citation counts.

For the **classification tasks**, we built traditional machine learning models such as `XGBoost` and `LightGBM`, as well as a deep learning model, `LSTM`, to classify the citation ranges.

This model helps us identify the key variables that are most effective in increasing citation counts, providing researchers with practical insights to enhance the visibility of their work.

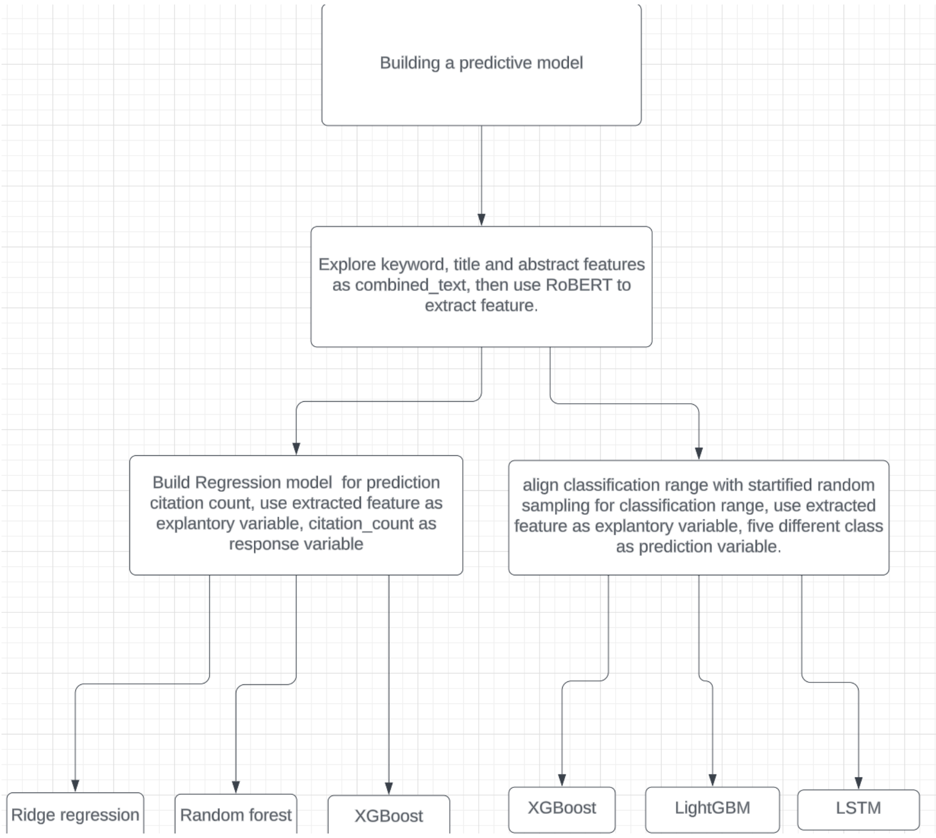


FIGURE 4.9. Prediction Model Introduction

Model Training and Validation

To ensure our models generalize well to unseen data and avoid overfitting, we will use the dataset selected in section 4.2 and split it into training, validation, and test sets. This will allow us to fine-tune the models and assess their performance accurately. Additionally, given the potential limitation in the size of our dataset, we will employ techniques such as cross-validation or bootstrapping to maximize the use of available data and prevent underfitting. These methods will help us ensure that our models are robust and perform well across different data subsets.

Evaluation Metrics

To assess the accuracy and effectiveness of our models, we will use a variety of metrics that are appropriate for the specific data distribution and model characteristics. These may include accuracy, precision, recall, F1-score, or other metrics that best reflect the model's performance on the given task(Chicco and Jurman 2020). By employing multiple evaluation metrics, we can ensure a comprehensive understanding of how well the model meets the project objectives.

4.4 Deployment

As our project is focused on studying the impact of social media on the academic community, it doesn't involve developing or delivering a software system that would require deployment. The primary output of our work is data analysis and a research report, which will be shared with stakeholders in the form of presentations and written documentation. Therefore, a detailed deployment plan is not necessary for this project.

CHAPTER 5

RESOURCES

5.1 Hardware & Software

- **NumPy**: Working with matrices and carrying out intricate mathematical operations required for our data analysis is made simpler using NumPy, which is used for numerical computations and managing big datasets.
- **Scikit-Learn**: For creating machine learning models such as classification, and regression
- **TensorFlow/PyTorch**: When we needed to analyze data with more intricate patterns for deep learning jobs, we utilized TensorFlow and PyTorch.
- **NLTK/Spacy**: For tasks involving natural language processing, such as text analysis on social media. These libraries aid in deconstructing textual material and identifying practical linguistic patterns.
- **Matplotlib/Seaborn**:For data visualization, including plots of correlation patterns and network structures.
- **ChatGPT**: To acquire assistance with a range of queries, we utilized ChatGPT as a consulting tool. It gave us concepts and advice that improved our comprehension and methodology.
- **Google Colab (A100 GPU)**: To expedite code execution, particularly when executing complex models, we utilized Google Colab's A100 GPU.
- **Slack**: Used to monitor progress and facilitate team collaboration. It enables us to share updates, interact effectively, and maintain team cohesion.

5.2 Materials

- **Social Media Data Access**: Access would depend on the platforms used such as Linkedin and Twitter for API access. Depending on the depth of access, special permissions or developer accounts are required.
- **High-Performance Computing (HPC)**: Access to HPC clusters or cloud-based computing platforms such as AWS, Google Cloud, or Microsoft Azure may be required if complex machine learning models or large-scale data are being used.
- **GPU Access**: Available through on-premise GPUs or cloud providers, GPU access can greatly accelerate training times for deep learning models.
- **API Access**: Some of them may incur some usage restrictions or even costs for extensive use, so one has to budget for using those APIs.

5.3 Roles & Responsibilities

Our jobs are rotated every week, and everyone's position will be selected from these 6 positions:

- **Project Manager (Qian Yu)**: supervises the project schedule and makes sure that deadlines are fulfilled. They serve as the primary point of contact for the customer, setting up meetings, facilitating communication, and responding to inquiries and issues. They are responsible for keeping an eye on developments and adjusting as necessary to keep the project moving forward.
- **Lead Developer (Lin Zhang, Yutong Wu)**: creates the database and other essential features, as well as the system architecture. In order to diagnose and optimize the system, they collaborate closely with other developers and guarantee the quality of the code.
- **Database Administrator (Yuzhe Zhou)**: focuses on data performance and integrity when designing and managing the database. They work with the development team to guarantee smooth integration and manage data storage, retrieval, and changes.
- **Quality Assurance (QA) Specialist (Qirui Chen)**: develops and carries out test plans to make sure the system satisfies all specifications. They find flaws, carry out testing during development, and collaborate with developers to fix problems prior to deployment.

- **Documentation and Presentation Specialist (Linfeng Yu)**: prepares all project documentation, ensuring it is clear and accurate. They also create presentation materials for client meetings and the final presentation, working closely with the Project Manager.

CHAPTER 6

MILESTONES SCHEDULE

Milestone	Tasks	Reporting	Date
Week-1	Analysis and design stage, gather data and create system mockup	Client meeting to review the project	04-08-2024
Week-2	Architecture design	Client meeting to review the work plan	11-08-2024
Week-3	Design work plan	None	18-08-2024
Week-4	Create database	None	25-08-2024
Week-5	Proposal Report Due		01-09-2024
Week-6	Get the final dataset, finished EDA and feature engineering		08-09-2024
Week-7	Finish object 1 & 2	None	15-09-2024
Week-8	Finish object 3 models created but without tuning parameter	None	22-09-2024
Week-9	Progress Report Due		29-09-2024
Week-10	Finish all the models tuning and start writing the Final report	Client meeting to deploy the system	06-10-2024
Week-11	Preparing final presentation		13-10-2024
Week-12	Final Presentation		20-10-2024
Week-13	Final Report (thesis)		27-10-2024

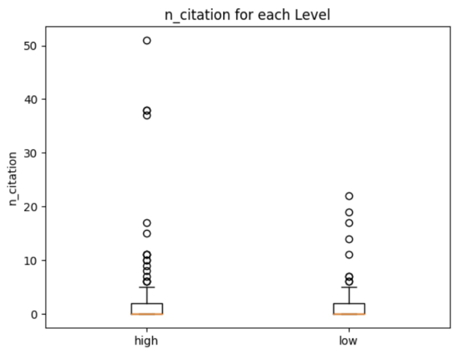
TABLE 6.1. Project Milestones and Tasks

CHAPTER 7

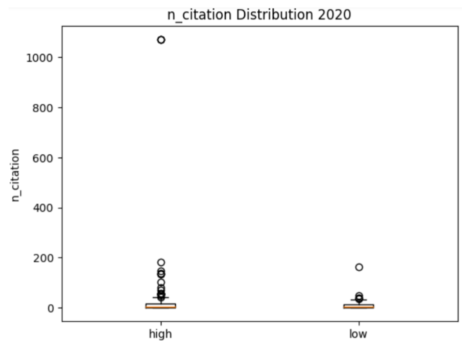
RESULTS

7.1 Relationship between social media exposure and citation counts

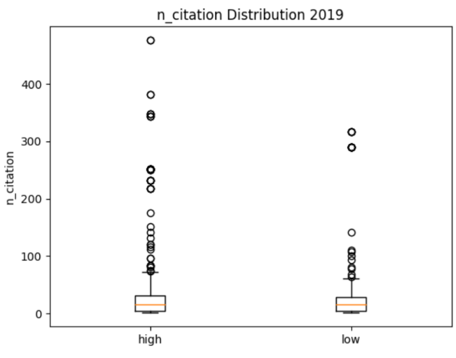
Due to the size of the data and the different hot words every year, we have performed feature selection on the original data and cut it by year and I also only select the cos similarity bigger than 0.6 data which means the papers' keyword can match with the industry keywords well. Through the box plot results of the annual citation volume and high-heat and low-heat keywords, we can see that no matter which level of keywords, there will always be some citations near 0, but there are indeed some data with higher citations.



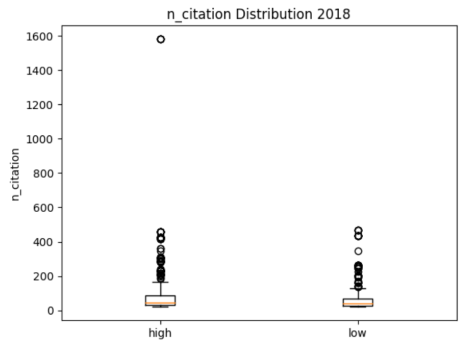
(A) Distribution of Citation 1



(B) Distribution of Citation 2



(C) Distribution of Citation 3



(D) Distribution of Citation 4

FIGURE 7.1. 2x2 Layout of Citation Distributions

If we remove the articles with 0 citations and calculate the average citation volume of each level, we will get a line chart like this, that is, the average citation volume of high-heat keywords will be greater than that of low-heat keywords in some cases.

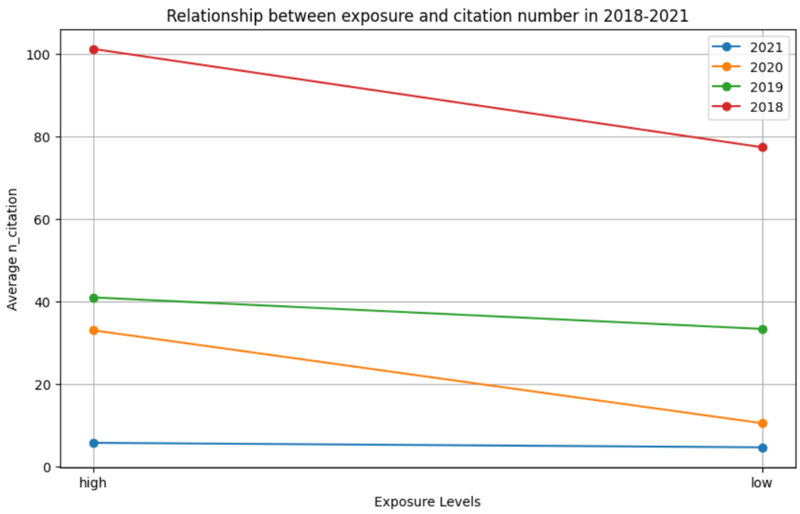


FIGURE 7.2. Distribution of Citation

Because we assume that hot keywords will have higher exposure on major social media platforms, we can conclude that although high exposure has a positive impact on the increase in citations, the two are not a simple linear relationship. The level of citations is also affected by other factors such as traffic purchases, whether the field is vertical, and the long-tail effect of the paper. Social media exposure can promote the growth of citations to a certain extent, but it is only one of the many factors that affect citations.

7.2 Network

In the network section we mainly focus on predictive and recommendatory modeling of network connections. The so-called citation network refers to the relationship between cited and cited articles. First of all, we choose a small data sample for the first model, through the article_id organization to extract the cited_by column of data, thus constructing a dataset containing both the cited and cited attributes, which is convenient for the model to learn the citation network structure of the article more directly. Based on this, we expand to large data samples, and instead of extracting the cited relationship, we train the model to capture the cited information in the cited relationship by itself. The specific steps are divided into, in the first step, starting from the original data, we first labeled each cited article of each article as a positive class (label 1), after which the citation lists of other articles were randomly selected 10 as interference and labeled as a negative class (label 0). After that, machine learning models such as random forest and XGBoost were used to predict the labeled classes (label: 0, 1). From the results, the models are more capable of predicting the positive class and can effectively restore the citation network relationship between articles. But for the negative class, although the accuracy is still good, it still incorrectly misclassified most of the interferences into the positive class. The test results on big data are relatively better, although the columns of the cited relations are not extracted in advance, the model captures the information of the citation network in the data very well. Some improvement was achieved in the effectiveness of both positive and negative class prediction.

Class	Precision	Recall	F1-Score	Support
0	0.62	0.30	0.41	166
1	0.57	0.84	0.68	186

TABLE 7.1. Random Forest on small sample dataset with attribute of cited_by

Class	Precision	Recall	F1-Score	Support
0	0.5583	0.5468	0.5525	68113
1	0.6874	0.6973	0.6923	97341

TABLE 7.2. XGBoost on big dataset without attribute of cited_by

7.3 Prediction model

7.3.1 Regression Model

We used three models: Ridge Regression, Random Forest, and XGBoost to predict `citation_count`. The results are shown in the table below:

Model	Mean Squared Error (MSE)	R^2 Score
Ridge Regression	1.9623	0.0767
Random Forest Regressor	1.8680	0.1211
XGBoost Regressor	1.6523	0.2226

Although the MSE values are relatively low, it's important to note that we scaled the data during preprocessing. Therefore, the R^2 score serves as a more accurate measure of model performance in this context. Among the models, XGBoost achieved the highest R^2 score, which was only around 22.26%. Based on this, we conclude that regression models do not perform well for predicting `citation_count` with this dataset, as the models failed to explain a significant portion of the variance in the data.

7.3.2 Classification Model

Due to the suboptimal performance of the regression models, we decided to group the data into five categories (0, 1, 2, 3, 4) based on the values from each stratum that were previously generated using stratified random sampling. Since our focus is on understanding how to achieve higher citation counts, we removed records where `citation_count` was zero, as these cases are not relevant to our analysis. Using RoBERTa, we extracted features from the `citation_count` data to predict the range of citation counts. For the classification task, we built models using XGBoost, LightGBM, and LSTM, leveraging both traditional machine learning and deep learning methods to classify the data into these five categories. The results are shown in the tables below.

TABLE 7.3. XGBoost Model Performance

Class	Precision	Recall	F1-Score	Support
0	0.39	0.80	0.53	8565
1	0.22	0.06	0.10	5328
2	0.25	0.13	0.17	5447
3	0.31	0.18	0.23	5031
4	1.00	0.94	0.97	4894
Accuracy				0.4573

The overall accuracy is 45.73%. For Class 4, which represents high-citation papers, XGBoost achieved a precision of 1.00, a recall of 0.94, and an F1-score of 0.97, indicating that the model was highly effective at identifying papers in this category. However, for lower-citation categories (Classes 0, 1, 2, 3), the performance was less optimal. For example, Class 0 achieved an F1-score of 0.53, while Class 1 only had an F1-score of 0.10, indicating the model's difficulty in distinguishing papers with fewer citations.

TABLE 7.4. LightGBM Model Performance

Class	Precision	Recall	F1-Score	Support
0	0.37	0.83	0.51	8565
1	0.22	0.04	0.06	5328
2	0.25	0.08	0.12	5447
3	0.28	0.13	0.18	5031
4	0.73	0.74	0.74	4894
Accuracy				0.4102

The overall accuracy is 41.02%. Similar to XGBoost, LightGBM performed well in predicting Class 4, with an F1-score of 0.74. However, its performance dropped significantly for the other classes. For example, Class 0 achieved an F1-score of 0.51, but Class 1 performed poorly, with an F1-score of only 0.06.

TABLE 7.5. LSTM Model Performance

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.39	0.75	0.51	8565
1	0.20	0.00	0.01	5328
2	0.25	0.09	0.13	5447
3	0.27	0.27	0.27	5031
4	0.82	0.91	0.86	4894
Accuracy				0.4568

The overall accuracy is 44.68%. LSTM also demonstrated strong performance in predicting Class 4, with an F1-score of 0.86. However, for lower-citation classes, its performance was similar to that of LightGBM and XGBoost, with low F1-scores for Class 1 (0.01) and Class 2 (0.13). LSTM had difficulty distinguishing between papers with similar citation counts, especially in the lower categories.

CHAPTER 8

DISCUSSION

8.1 Relationship between social media exposure and citation counts

As the influence of social media in the scientific community grows (Regenberg 2019), more and more academic journals are using social media platforms such as Twitter and Facebook to share and promote research projects and results, and increase visibility and engagement within and outside the academic community (Cylkowski 2020). And Olena Zimba and Armen Yuri Gasparyan said, “If editors manage their journals’ Twitter, Facebook, and other popular social media accounts in an ethical manner, they can involve influential authors in post-publication communications and expand the social impact of the journal.” (Zimba and Gasparyan 2021). Thus, this study found that higher social media exposure can indeed increase the visibility of articles, help spread articles in a wider academic and public community, and give more scholars the opportunity to access the article and cite its content. But there also situations where no matter how high the exposure is, there will always be data near 0 citations and high exposure also have low citation number, low exposure also have high citation number. We discuss and analyze this phenomenon:

1. As we observed in the line chart, the overall citations of articles with older years are generally higher, which shows that there may be a time lag between exposure and citation, which may be because of the monotonic growth of the citation network, where older papers tend to have more citations than more recent papers(Tóth et al. 2020). The article may have just been published, and although it has received high exposure on social media, academic citations usually take time to accumulate. In addition, since it will requires researchers to spend time reading, understanding the papers, high social media exposure does not guarantee immediate citations(Hare et al. 2023). Therefore, high social media exposure has not yet been converted into academic citations in the short term.

2. Exposure on social media does not necessarily reflect the academic value of the article. Some articles may receive wide exposure on social media because of attractive topics, novel titles, or easy-to-spread discussion content, but they do not have sufficient academic depth or innovation, resulting in low citations. This situation is especially common in articles with public appeal but limited academic value. Thus, the popularity of certain topics on social media does not necessarily reflect the academic impact of the article(Hassan et al. 2023).

3. Social media has different audiences from academia. Users on social media are broader and may include non-academics or scholars outside the field. These users may share or discuss articles but do not directly cite them. Therefore while the exposure of the articles might be high because the topic is interesting, this may not increase the citation rate of a scientist's papers.(Branch et al. 2023).

8.2 Network

citation network refers to the network structure compiled by the citation relationship between articles. It has articles as nodes, citation relationships as links, and citation active-passive relationships as the direction of the links. (Daud2020) This model turns out to perform quite admirably with great adaptability and practicality in binary classification problems for citation forecast, especially doing an excellent job in classifying Class 1 papers, which are cited. The recall score of 0.79 obtained on the model denotes that it captures most of the cited papers, something important to the researchers in order for them to understand the potential impact of their work. With a precision score of 0.67, this model is highly reliable in predicting when a paper is going to be cited-that is, it is often right when it predicts a citation. In terms of early identification of high-impact papers, such reliability is of value for both researchers and academic institutions in resource allocation and promoting research output.

From a design viewpoint, though the two classes are imbalanced, with cited papers fewer, it did very well, since it keeps its predictions balanced across classes. The model makes use of features like titles, keywords, and abstracts that have interpretative value in citation prediction. Generally speaking, highly cited papers reflect some distinguishing characteristics, and the model learns a pattern to predict effectively for citation potential.

From an optimization point of view, it could be better, especially tuning the balance between precision and recall. That could have been further improved by hyperparameter tuning or by resorting to more informed feature extraction methods, such as TF-IDF or word embeddings. Also, domain knowledge may be applied in feature engineering to capture such subtle differences that exist in the contents of academic papers and hence improve prediction accuracy.(Zhao2015)

In summary, its high recall and reliable precision for the model's prediction of cited papers are adaptable to key academic text features, turning this into a very valuable tool in supporting researchers by analyzing and predicting academic impact for their works. This is not only essential for assessing the prospective effect of research, but such a metric provides a quantitative basis for handling academic publishing and research management. It will help in proper resource allocation and overall quality improvement in research.

8.3 Prediction Model

Although the model didn't perform well in predicting the lower citation categories (Classes 0, 1, 2, 3), the results are still meaningful for our research goal. Our research question is focused on predicting highly cited papers to help researchers understand how their publications might perform in the future. If the model doesn't predict a high citation count, it could mean there's room for improvement in the paper's features.

In our experiments, the XGBoost model did especially well in predicting the highest citation category (Class 4), achieving an F1-score of 0.97. We believe this is related to how we initially defined the category ranges. Class 4 represents papers with more than 200 citations, so there is likely greater variation in this category, making it easier for the model to capture distinctive features. In contrast, for Classes 0, 1, 2, and 3, the citation ranges are smaller, and the data differences are not as pronounced, making it difficult for the model to distinguish between them and learn their characteristics.This difficulty in predicting lower citation categories may also reflect the "Matthew effect" in scientific citation patterns, where highly cited papers continue to accumulate citations(Merton 2016).

Even though the model struggled with the lower citation categories, these findings still provide valuable insights. We can use this information to examine which aspects of keywords, titles, or abstracts might be impacting citation counts. For instance, if papers with lower citations share certain patterns in their titles or keywords, we could focus on improving those specific areas. Making these features more detailed or clearer might help increase citation numbers.Studies have shown that specific elements, such as the use of trending keywords or title length, can influence a paper's citation impact (Aksnes et al. 2019). This suggests that optimizing these features could be an effective strategy to increase visibility and citations. This has important implications for journals, research institutions, and even the academic publishing industry, as it could help boost the visibility of papers.

Overall, while the model's accuracy for the lower citation categories wasn't perfect, it still gives us a good starting point for understanding how to improve the visibility and citation rates of lower-cited papers.

8.4 Conclusion

Through the identification and discussion of the findings' implications and relevance, the analysis offers a perceptive comprehension of project outcomes. It critically looks at how these outcomes help to achieve important project goals, fill in gaps, and fix problems that have been found. The analysis identifies a number of significant variables, including time lag, academic merit, audience type, and self-promotion, that affect the correlation between social media exposure and citation counts.

With a recall of 0.79 and a precision of 0.67, the citation prediction model that was created demonstrates a noteworthy level of accuracy in identifying publications that are likely to be referenced. This provides researchers and institutions looking to forecast the impact of their study with useful information. While there are still issues with lower-citation categories, the model's ability to identify highly cited articles (Class 4) shows promise. With wider ramifications for journals and organizations looking to boost paper visibility and citation potential, the findings also point up opportunities for improvement, particularly in improving aspects like keywords, titles, and abstracts. All things considered, this paper makes a substantial contribution to our understanding of how social media might affect academic effect and reach while outlining doable strategies for improving academic engagement and citation results.

CHAPTER 9

LIMITATIONS AND FUTURE WORKS

This project faced several key limitations during its execution. The first major limitation was computational power cost. Given the large volume of data, particularly for high-dimensional feature analysis like keywords and abstracts, the lack of sufficient computational resources impacted both the efficiency and accuracy of model training. We were unable to utilize more complex models or perform extensive hyperparameter tuning, which limited the overall performance of the models. Future work could address this by leveraging higher performance computing to improve model predictions.

Another significant limitation was our inability to access large social media APIs. Although social media platforms potentially have a remarkable influence on citation counts, due to restrictions in accessing these data, we were unable to incorporate social media metrics (e.g., tweets) into our model. This restricted our ability to analyze the exposure of papers across online platforms. Future studies could focus on collaborating with social media platforms to gain access to their APIs, allowing a more in-depth exploration of social media's role in academic influence.

Furthermore, since there is no feature about exposure in our dataset, we decided to use the keywords of each article to correspond to the current hot words in different industries every year, and match the hot words of each industry with the keywords of the article to get the exposure. This leads to a problem that since we can only divide the exposure into levels, we can't use the model to fit the data to get a line chart that expresses the specific relationship, but can only show whether there is a relationship between them, and how exposure affects the citation volume. Moreover, our industry hot words are generated by chatgpt, which includes 20 industries, and each level has 10 hot keywords, which may cause incomplete industry coverage and incomplete keyword coverage, resulting in a slightly insufficient sample data size.

In exploring “The relationship between social media exposure and citation counts is positively correlated", we discovered some limitations.Despite high social media exposure, some articles still receive few or no citations, while some low-exposure articles achieve high citation counts. This may result from a time lag between exposure and citation accumulation. Additionally, social media exposure does not necessarily indicate academic value, as the audience on social media differs from that in academia. In some cases, high exposure is driven by authors or institutions, which does not always translate into citations from other scholars, leading to a disconnect between high exposure and low citations.(Weng2013)

And our model also has different limitations.The network model is highly effective in binary classification for citation prediction, particularly for identifying cited papers (Class 1). However, accuracy could still be improved through hyperparameter tuning and advanced feature extraction methods, such as TF-IDF or word embeddings. Since one paper often contains only a part of keywords that a user is interested in, recommender system returns a set of papers that satisfy the user’s need of keywords. However, each paper of an existing paper citation network hardly has cited relationships with others, so the correlated links among papers are very sparse. In addition, while a mass of research approaches have been put forward in terms of link prediction to address the network sparsity problems, these approaches have no relationship with the effect of self-citations and the potential correlations among papers (i.e., these correlated relationships are not included in the paper citation network as their published time is close)(Liu et al. 2019). What’s more, the predictive model struggled with lower citation categories (Classes 0-3) but performed well in predicting the highest citation category (Class 4), likely due to greater variability in Class 4. The smaller citation ranges and less pronounced data differences in lower categories made it challenging for the model to capture distinctive characteristics.

Additionally, time limited the process of our project. Given the short timeframe, we were unable to analyze a wider range of variables or test our methods on these large datasets. Such additional experiments and analyses could have provided a more comprehensive evaluation of the model’s applicability and robustness. Future research could extend the timeline to explore more different reliable datasets, enhancing the model’s generalizability and adaptability.

Future work can expand in several directions. First, we need to overcome the problem of insufficient computing power, so that more complex deep learning models can be applied, such as using Transformer to build a more powerful citation prediction system to capture the complex citation relationships between papers. Second, we could optimize the feature extraction process, especially by introducing more advanced text preprocessing techniques and Natural Language Processing models to improve the prediction accuracy for low citation papers. In addition, we also need to enhance our ability to obtain multi-source data, by connecting to the API interfaces of major social media platforms, to achieve automatic collection and continuous updating of multi-dimensional data such as user behavior, interaction, and comments, thereby enhancing the accuracy of data analysis and improving the authenticity of data.

Bibliography

Adetayo, Adebowale Jeremy (2023). 'Research output and visibility of librarians: Are social media influencers or distractors?' In: *Journal of Librarianship and Information Science* 55.3, pp. 813–827.

Akella, Akhil Pandey et al. (2021). 'Early indicators of scientific impact: Predicting citations with altmetrics'. In: *Journal of Informetrics* 15.2, p. 101128.

Aksnes, Dag W, Liv Langfeldt and Paul Wouters (2019). 'Citations, citation indicators, and research quality: An overview of basic concepts and theories'. In: *Sage Open* 9.1, p. 2158244019829575.

Bardus, Marco et al. (2020). 'The use of social media to increase the impact of health research: systematic review'. In: *Journal of medical Internet research* 22.7, e15607.

Branch, T. A. et al. (2023). 'Controlled experiment finds no detectable citation bump from Twitter promotion'. In: *bioRxiv (Cold Spring Harbor Laboratory)*. DOI: 10.1101/2023.09.17.558161. URL: https://doi.org/10.1101/2023.09.17.558161.

Bütün, Ertan, Mehmet Kaya and Reda Alhajj (2017). 'A supervised learning method for prediction citation count of scientists in citation networks'. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 952–958.

Cao, Renmeng et al. (2023). 'How do scientific papers from different journal tiers gain attention on social media?' In: *Information Processing & Management* 60.1, p. 103152.

Chiang, Austin Lee et al. (2021). 'The patterns and impact of social media exposure of journal publications in gastroenterology: Retrospective cohort study'. In: *Journal of Medical Internet Research* 23.5, e25252.

Chicco, Davide and Giuseppe Jurman (2020). 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'. In: *BMC genomics* 21, pp. 1–13.

Cylkowski, K. (2020). 'Impact of social media on academic journals'. In: *The Journal of Perinatal & Neonatal Nursing* 34.4, pp. 287–288. DOI: 10.1097/jpn.0000000000000487. URL: https://doi.org/10.1097/jpn.0000000000000487.

Donelan, Helen (2016). 'Social media for professional development and networking opportunities in academia'. In: *Journal of Further and Higher Education* 40.5, pp. 706–729. DOI: 10.1080/0309877X.2015.1014321.

Garcia, Débora Cristina Ferreira, Cristiane Chaves Gattaz and Nilce Chaves Gattaz (2019). *The relevance of title, abstract and keywords for scientific paper writing*.

Günther, E. and E. Domahidi (2017). 'What communication scholars write about: An analysis of 80 years of research in high-impact journals'. In: *International Journal of Communication* 11, p. 21. DOI: 10.15496/publikation-30502. URL: https://doi.org/10.15496/publikation-30502.

Hare, M. et al. (2023). 'Do you cite what you tweet? Investigating the relationship between tweeting and citing research articles'. In: *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2306.16554. URL: https://doi.org/10.48550/arxiv.2306.16554.

Hassan, D. G., M. E. Tantawi and M. G. Hassan (2023). 'The relation between social media mentions and academic citations in orthodontic journals: A preliminary study'. In: *Journal of the World Federation of Orthodontists* 12.3, pp. 125–130. DOI: 10.1016/j.ejwf.2023.05.003. URL: https://doi.org/10.1016/j.ejwf.2023.05.003.

He, Qin (1999). 'Knowledge Discovery through Co-Word Analysis'. In: *Library Trends* 48.1, pp. 133–159.

Hou, Jie et al. (2019). 'Prediction methods and applications in the science of science: A survey'. In: *Computer science review* 34, p. 100197.

Jolliffe, I. T. and J. Cadima (2016). 'Principal component analysis: a review and recent developments'. In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences* 374.2065, pp. 20150202–20150202. DOI: 10.1098/rsta.2015.0202.

Kang, Hyeonsu B et al. (2023). 'ComLittee: Literature Discovery with Personal Elected Author Committees'. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20.

Klar, Samara et al. (2020). 'Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work'. In: *PloS one* 15.4, e0229446.

Lee, Dong Kyu (2020). 'Data transformation: a focus on the interpretation'. In: *Korean journal of anesthesiology* 73.6, pp. 503–508.

Liu, Hui et al. (2019). 'Link prediction in paper citation network to construct paper correlation graph'. In: *EURASIP Journal on Wireless Communications and Networking* 2019.1, pp. 1–12.

Lohr, Sharon L. (2019). *Sampling: Design and Analysis*. Second edition. CRC Press.

Luc, Jessica GY et al. (2021). 'Does tweeting improve citations? One-year results from the TSSMN prospective randomized trial'. In: *The Annals of thoracic surgery* 111.1, pp. 296–300.

Merton, Robert K (2016). 'The Matthew effect in science'. In: *Science* 146, p. 496.

Ortega, José Luis (2016). 'To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers'. In: *Scientometrics* 109, pp. 1353–1364.

Ortega, Jose Luis (2017). 'The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations)'. In: *Aslib journal of information management* 69.6, pp. 674–687.

Özkent, Y. (2022). 'Social media usage to share information in communication journals: An analysis of social media activity and article citations'. In: *PLoS ONE* 17.2, e0263725. DOI: 10.1371/journal.pone.0263725. URL: https://doi.org/10.1371/journal.pone.0263725.

Pobiedina, Nataliia and Ryutaro Ichise (2016). 'Citation count prediction as a link prediction problem'. In: *Applied Intelligence* 44, pp. 252–268.

Priem, Jason and Kaitlin L. Costello (2010). 'How and why scholars cite on Twitter'. In: *Proceedings of the American Society for Information Science and Technology* 47.1, pp. 1–4. DOI: 10.1002/meet.14504701201.

Radicchi, Filippo, Santo Fortunato and Claudio Castellano (2008). 'Universality of citation distributions: Toward an objective measure of scientific impact'. In: *Proceedings of the National Academy of Sciences* 105.45, pp. 17268–17272.

Regenberg, A. (2019). 'Science and social media'. In: *Stem Cells Translational Medicine* 8.12, pp. 1226–1229. DOI: 10.1002/sctm.19-0066. URL: https://doi.org/10.1002/sctm.19-0066.

Smith, Zachary L et al. (2019). 'Longitudinal relationship between social media activity and article citations in the journal Gastrointestinal Endoscopy'. In: *Gastrointestinal endoscopy* 90.1, pp. 77–83.

Sugimoto, Cassidy R. et al. (2017). 'Scholarly use of social media and altmetrics: A review of the literature'. In: *Journal of the Association for Information Science and Technology* 68.9, pp. 2037–2062. DOI: 10.1002/asi.23833.

Thelwall, Mike et al. (2013). 'Do altmetrics work? Twitter and ten other social web services'. In: *PloS one* 8.5, e64841.

Timilsina, Mohan et al. (2017). 'Predicting citations from mainstream news, weblogs and discussion forums'. In: *Proceedings of the international conference on web intelligence*, pp. 237–244.

Tonia, Thomy et al. (2020). 'If I tweet will you cite later? Follow-up on the effect of social media exposure on article downloads and citations'. In: *International journal of public health* 65, pp. 1797–1802.

Tóth, I. et al. (2020). 'Mitigating ageing bias in article level metrics using citation network analysis'. In: *Journal of Informetrics* 15.1, p. 101105. DOI: 10.1016/j.joi.2020.101105. URL: https://doi.org/10.1016/j.joi.2020.101105.

Veletsianos, George and Royce Kimmons (2012). 'Networked Participatory Scholarship: Emergent techno-cultural pressures toward open and digital scholarship in online networks'. In: *Computers and Education* 58.2, pp. 766–774. DOI: 10.1016/j.compedu.2011.10.001.

Wang, Meng and Fanghui Hu (2021). 'The application of nltk library for python natural language processing in corpus research'. In: *Theory and Practice in Language Studies* 11.9, pp. 1041–1049.

Wang, Mingyang, Zhenyu Wang and Guangsheng Chen (2019). 'Which can better predict the future success of articles? Bibliometric indices or alternative metrics'. In: *Scientometrics* 119, pp. 1575–1595.

Weissburg, Iain Xie et al. (2024). 'Tweets to citations: Unveiling the impact of social media influencers on ai research visibility'. In: *arXiv preprint arXiv:2401.13782*.

Wouters, Paul et al. (2019). 'Social Media Metrics for New Research Evaluation'. In: *Springer Handbook of Science and Technology Indicators*. Springer International Publishing, pp. 687–713. DOI: 10.1007/978-3-030-02511-3_26.

Zimba, O. and A. Y. Gasparyan (2021). 'Social media platforms: a primer for researchers'. In: *Reumatologia/Rheumatology* 59.2, pp. 68–72. DOI: 10.5114/reum.2021.102707. URL: https://doi.org/10.5114/reum.2021.102707.

Appendix A

A1 Visualization of keyword distribution throughout the year

- The distribution of citations across all years of our dataset is shown below, with the five most frequently used words shown for each segment.