Norwegian School of Economics

Bergen, Spring 2022

Predicting the impact of academic articles on marketing research:

Using machine learning to predict highly cited marketing articles

Ingrid Skogeng Hansen and Magnus Torvund

Supervisor: Ivan Belik

Master Thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible − through the approval of this thesis − for the theories and methods used, or results and conclusions drawn in this work.

Preface

This thesis is written as a part of the MSc in Economics and Business Administration at the Norwegian School of Economics (NHH), with a major in Business Analytics. Writing this thesis has been a valuable experience, and it has strengthened our interest in what drives the impact of academic publications.

We wish to thank our supervisor, Ivan Belik, for valuable advice, feedback, and discussions throughout this process. His guidance has been of great help in improving our thesis. In addition, we would like to thank our friends and family for supporting and motivating us.

Norwegian School of Economics

Bergen, May 2022

Ingrid Skogeng Hansen

Magnus Torvund

Abstract

The citation count of an academic article is of great importance to researchers and readers. Due to the large increase in the publication of academic articles every year, it may be difficult to recognize the articles which are important to the field. This thesis collected data from Scopus with the purpose to analyze how paper, journal, and author related variables performed as drivers of article impact in the marketing field, and how well they could predict highly cited articles five years ahead in time. Social network analysis was used to find centrality metrics, and citation count one year after publication was included as the only time dependent variable. Our results found that citations after one year is a strong driver and predictor for future citations after five years. The analysis of the co-authorship network showed that closeness centrality and betweenness centrality are drivers of future citations in the marketing field, indicating that being close to the core of the network and having brokerage power is important in the field. With the use of machine learning methods, we found that a combination of paper, journal, and author related drivers perform better at predicting highly cited articles after five years, compared to using only one type of driver.

Table of Contents

List of Tables

List of Figures

1. Introduction

In the last decade, the number of marketing-related articles have had a rapid increase. With the rising amount of new scientific articles being published, there is a growing need to identify the potentially impactful articles from the less impactful ones (Yan et al., 2011). It is impossible to track all new research articles. Therefore, an important but challenging task in academia is to predict the future impact of a published paper on the research field (Ma et al., 2021; Abrishami & Aliakbary, 2019; Cao et al., 2016).

Impact can be estimated through various measures, but citation count is regarded as the most important indicator for this purpose (Didegah & Thelwall, 2013; Oppenheim, 1995). This thesis interprets the term *impact* as “the actual influence on surrounding research activities at a given time”, as defined by Aksnes et al. (2019). Hence, impact gives an estimation of how significant an article is to its respective field, which reflects the utility of research. Since citations is recognized as a proxy of impact, this thesis will use *citations* as the measure of article impact.

Identifying and utilizing the drivers of future article impact, is not only of importance to authors of articles, but also to researchers, journals, and other entities who need to stay updated on a field. For the readers of articles, citations work as a tool to find relevant and high-quality articles. For researchers, citations can affect their salary, funding, and their position in the field. Citations can also impact the position of a journal in a field (Stremersch et al., 2007). The greater the impact, the more established the reputation of the scholar (Li et al., 2013). This gives incentives to the producers and publishers of research articles to maximize the number of citations they can get. Furthermore, understanding the drivers behind future article citations can help researchers understand how research drives the field.

There are many different factors that can affect the number of citations. It is possible to split the variables into categories. The categories typically group variables related to the paper itself, the journal it is published in, or characteristics of the author of the article. Furthermore, social networks, such as co-authorship networks, have recently been seen as useful for predicting article impact (Colladon, 2020). The network created of relationships between authors give valuable information and knowledge of social interactions. Factors like having a strong position and being important in the author network, has previously shown to have a positive effect on citations. In addition, social relations of authors have an important role in the recognition of their research. Actors in the network can benefit by gaining more knowledge and a wider horizon, and possibly receive better outcomes (Li et al., 2013), such as enhanced quality of the research (Katz and Martin, 1997).

There has been substantial research done on the prediction of future citation counts and highly cited articles (Ma et al., 2021). However, there has not been a large focus within the field of marketing. Hence, we will in this thesis first study what drives the citations of marketing articles, before using this information to predict highly cited academic marketing articles after five years. The thesis aims to answer the following research question:

What are relevant drivers of marketing article impact between 1992 and 2022, and how can these drivers be utilized to predict highly cited marketing articles five years after publication?

With this research question, we aim to identify drivers of article citations through a literature review, test relevant drivers with the use of data from the period 1992-2022, and investigate how well these drivers predict whether an article will be a highly cited article five years after publication. We will test the drivers by classifying marketing articles from 2015 and 2016 as either highly cited or non-highly cited five years after publication. To test this, we will be using the most promising predictive methods today. We have chosen to predict citations five years after publication as this is a sufficient time interval to identify article impact due to most articles having a determining citation path by then (Abramo et al., 2011). It is also a standard timeframe for prediction of citation counts, which means that we can compare our results with previous work in citation count predictions.

To identify potential drivers, we have done an extensive literature review which includes important research that has been done on the subject. In addition, we have collected articles from 75 English-written marketing journals published between 1992 and 2022 for analysis purposes. Marketing articles from this time-period are covered to a great extent and generally have good metadata quality compared to articles prior to the 1990s. By looking at a broad timeframe, we have the possibility to find patterns of how marketing articles gain citations as well as having better ability to increase the robustness of the testing. It is also important that we have enough data from previous years to test the quality of our predictions. As we need citation data from five years after publication, the latest article we can test predictions on are from 2016.

In the following section, we will present a literature review to provide knowledge about existing research on drivers of article impact within marketing research and the citation count prediction problem. Based on found gaps in the literature from the literature review, we will form specific hypotheses to test. Further, a methods section presents the methods used to achieve our results, followed by a data section that explains how we have collected and prepared our data for analysis. After the data collection section, we present our results, before discussing how well the results managed to answer our hypotheses. Within the discussion we also include the implications for authors and readers of marketing articles. Finally, we conclude with how our results contribute to the field, and we discuss the limitations and possible future research possibilities.

2. Literature Review

This section contains past work done on citations and predictions of future citations, which we find necessary as a foundation to answer our research question. We will in this part present the important research done within 1) what citations represents, 2) drivers of citations, and 3) previous attempts to predict future article citations. The first part discusses importance of citations to researchers and readers, as well as what it represents. The second part examines the different types of drivers, and how they affect citation counts. In the last part, previous solution proposals to the citation count prediction problem are discussed, and we look at how social network metrics have been used in previous research to measure article impact.

2.1 Article Citations

Citations was primarily intended for helping researchers more effectively search through the literature (Mingers & Leydesdorff, 2015). The number of citations is the most frequently used measure to evaluate the quality of articles (Tahamtan et al., 2016). However, it has been primarily shown to reflect aspects of impact (Aksnes et al., 2019). Although research has found that the citation count is not necessarily correlated with article quality, many citations can indicate the utility an article has for others, and followingly the academic impact of the article (Nightingale & Marshall, 2012).

There are many reasons why a researcher cites the work of other researchers in their own articles (Tahamtan, 2016). It could be to support their own claims, methodology or findings, or to present a different point of view of other researchers. Some papers are cited to be criticized, while others are cited as positive or negative examples.

Article citations can affect the career and salary of an academic researcher (Stremersch et al., 2007). However, in Norway, citations do not affect funding opportunities or research policy decisions (Aksnes & Rip, 2009). Instead, having a highly cited article can give important collaboration opportunities and scientific positions. Bhandari et al. (2007) found that researchers tend to seek being published by high-impact journals to become more frequently cited. Research has found that a weak article that does not contribute with anything new may receive many citations if published in a journal with many readers (Callaham, 2002). This suggests that academic network matters for promotion of articles.

2.2 Drivers of Article Citations

The question of what influences citations have been studied in numerous aspects, such as the importance of time, authors, and journals (Ma et al., 2021). In this thesis, a driver of article impact is defined as a factor that is influencing citations. According to Tahamtan et al. (2016), the drivers of article citations can be divided into three categories: paper related, journal related, and author related drivers. As Tahamtan et al. (2016) does not define these categories, we provide definitions based on the types of content found in the different categories in this paper. We define paper related drivers as drivers based on either the content of an article or related to the content of it. We define journal related drivers as metrics related to the journal the article is published in. Lastly, we define author related drivers as characteristics and metrics related to the authors of an article.

In this section, we will elaborate on the research done within these three categories. A tabular overview of the drivers tested and on which fields can be seen in Appendix A3.

2.2.1 Paper Related Drivers

Increasing quality of a paper is found to increase the number of citations of an article (Tahamtan et al., 2016). Peer reviewed papers, and longer review times is found to result in increased article quality and higher number of citations (Tahamtan et al., 2016). Stremersch et al. (2007) found article order in journal, editorial journal awards, and article length to have a positive effect on article citations. These are factors which can reflect scientific value, as journal editors may order and allocate the placement of an article after their perceived scientific value. Furthermore, papers which introduce novel connections between clusters of co-cited references tend to generate more citations (Chen et al., 2010). Innovative and novel subjects in articles is often seen as a quality measure.

The question of whether the complexity, length, or sentiment of a paper affect citation count has been investigated in multiple studies. Text length has been studied both in terms of title, abstract, and paper length. While title length has been found to affect citations in some fields, it has not been found to affect citations in marketing (Stremersch et al., 2007). Bornmann et al. (2014) found that the number of pages impacted citations during the first few years after the publication of the paper. Other research finds no effect of number of pages (Royle et al., 2013). Furthermore, longer abstracts can lead to more citations than shorter abstracts (van

Wesel et al., 2014). In terms of complexity, Colladon et al. (2020) found that articles with high lexical diversity tend to attract more citations. In addition, they found that sentiment of abstract can impact citations. Warren et al. (2021) investigated the abstraction, technical language, and passive writing in marketing articles and found that articles with high degree of such characteristics made them less likely to be cited. In other words, when articles are written clearly, they will be more easily understood, which increases the probability of making an impact.

Several research papers found accessibility and visibility to be related to citations (Henneken et al., 2006; Yu and Wilson, 2004). However, there are a few papers that have a contradicting result. Moreover, articles which are published in open-access journals are found to be cited more compared to non-open-access journals, given that they are published in a top-50 journal (Tahamtan et al., 2016). The opposite effect takes place for bottom-ranked journals. Furthermore, diversity and number of keywords in a paper are found to increase number of citations (Chakraborty et al. 2014; Rostami et al., 2013; So et al. 2014). It has also been found that the number of references, the variety of references, as well as their prestige, can increase citations (Tahamtan et al., 2016).

Time-based paper related drivers

Early-stage citation count has been shown as an effective predictor of future citation counts in several research publications (Ma et al., 2021; Ruan et al., 2019; Abrishami & Aliakbary, 2019; Abramo et al., 2019, Stegehuis et al., 2015). Early-stage citation counts can tell something about the future pattern of citations of an article as they reveal the articles initial reception from the research community. While there exists a lot of patterns, they are generally based on when and for how long the knowledge become useful for researchers. There are two major types of successful article patterns based on when a paper become useful: “Classics” (van Raan, 2004) and “Shooting star” (Ye & Bornmann, 2017). The classics gets a substantial number of citations quickly and obtain a considerable number of citations even several years after publication, indicating that their content has relevance for a long period (Baumgartner & Leydesdorff, 2013). The shooting stars peak in yearly citations quickly (3-4 years), but fade quickly as well, indicating a transitional relevance of the knowledge it contains.

Early-stage citation counts should be used with caution as papers published early in a year has a significant additional timeframe compared to those published late in a year (Levitt & Thelwall, 2011). Increasing the time-window for collecting early-stage citation counts improves long-term predictions (Aksnes et al., 2019). However, this comes at a cost of having to wait longer before articles can be predicted.

Although variables based on number of citations in early years can be considered as a paper related variable, it is important to be aware of the time aspect. Being able to collect information after the date of publication can be a huge advantage, as the variable will have some information about the initial *true* impact of the article. In addition, it can absorb the effect of other types of variables that has an effect on early-stage citations. This makes these variables different from other paper related drivers.

2.2.2 Journal Related Drivers

In the quest for more citations, researchers try to get their articles published in journals with high impact (Tahamtan, 2016). Bornmann et al. (2013) and Bornmann et al. (2014) found that journal variables are significant characteristics to include when predicting article impact. The impact factor of the journal in which the article is published could be considered as an indicator for the quality of the article. If the quality of the paper is high, the higher are the chances of it being published by credible journals, leading to a higher probability of it getting attention. In addition, the prestige of a journal can be a measure of the article quality (Dervos and Kalkanis, 2005). Garner et al. (2014) found that articles that are published in high-impact journals tend to have a closer and faster citation rate than articles published in journals with lower impact. A weak article may get a relatively higher number of citations when published by a high-impact journal (Callaham et al. 2002). However, Bornmann and Williams (2013) found that an article with a broader topic can receive more citations even when they are published in low-impact journals.

Although there are several studies that support that journal impact factor increase citations, there are studies that have contradicting results. Various studies have not found a positive correlation between journal impact factor and number of citations in articles (Willis et al., 2011; Leimu and Koricheva, 2005). Whether the journal impact factor is influencing citations or not might therefore depend on other factors and vary within fields of research.

Only a few studies have been investigating SCImago Journal Rank (SJR) or other journal rankings as drivers of citations. Sohrabi & Iraj (2016) found that the SCI quartile had a statistically significant effect. Using quartile 1 as the base level, all the other quartiles, in a descending order, had an increasingly negative coefficient. Hu et al. (2020) also used SJR as one of their journal related drivers that they compared with keyword popularity features.

2.2.3 Author Related Drivers

Many studies have shown that there is a positive correlation between the number of authors of a paper and the number of citations a paper gets. A broader social network to spread the knowledge within gives a wider audience to promote the paper (Tahamtan et al., 2016). This is contrary to what is shown in the marketing domain where Stremersch et al. (2007) argue that increasing amounts of authors has a negative effect on citations. Further, they argue that increasing amounts of authors can reduce the intellectual ownership, and hence decrease the willingness to promote the article. Since natural science articles can have significantly more authors than marketing papers due to technological complexity, the negative effect of diminishing ownership to promote research may for marketing papers therefore be greater than the increased potential knowledge diffusion effect of having many authors.

Social Network Drivers

Social influence aspects, such as the social network of an author, may impact citations. An example is that articles from widely connected authors have been found to have a higher probability of being cited (Chakraborty et al. 2014). A proxy to authors social network can be author productivity. High productivity researchers are more likely to have a greater social network, which in turn leads to more citations (Tahamtan et al., 2016). Similarly, if a paper has authors from several organizations, then it is found to be cited significantly more than those from one organization (Puuska et al., 2013). International collaboration has also been found to positively impact citations. However, impact can vary depending on domain, and what type of network generating the highest citations differs between fields of research (Tahamtan et al., 2016).

Rosenzweig et al. (2016) found that social network has an impact on marketing research articles, and that female researchers and researchers originating from less economically advanced countries are more likely to utilize their social network. Another research found that the more connected scholars in coauthor networks in the marketing discipline are more important (Goldenberg et al., 2006). Thus, their work can transfer more easily in the network, and they can get more citations.

2.3 Prediction of Article Impact

The prediction of article impact can be divided into two tasks; 1) predicting the future citation counts of each specific article and 2) identifying future highly cited papers (Wang et al., 2019). The primary differences between these tasks are the methods applied, where task 1) is a regression problem and task 2) is a binary classification task. However, variables which are important in one task are likely important in the other as both problems are based around predicting future citation counts, where the binary problem is predicting whether the article will have citation counts above or below a given threshold. The prediction window varies but are typically five or six years after publication (Ma et al. 2021; Hu et al. 2020; Sarigöl et al., 2014).

According to Ma et al. (2021) the citation count prediction task was first brought to light by Yan et al. (2011) and has since then got attention from numerous researchers. A majority of research papers are hardly cited at all, while a few are cited considerably (Tahamtan et al., 2016). Knowing what drives article citations and being able to identify the highly impactful articles early is therefore of high value.

2.3.1 Prediction of Highly Cited Articles

The task of identifying highly cited papers is defined as a classification problem. Many previous efforts have been done, and it has become clear that there are numerous drivers which contribute to article impact, as presented in 2.2.

There have been tested numerous drivers of article citations on prediction. In terms of paper related drivers, semantic and sentiment text analysis of abstract has been tested with promising results (Colladon, 2020). Using early-stage citation counts, both alone and combined with other variables, has improved predictions (Steighuis, 2015; Ma et al., 2021). Hu et al. (2020) extracted keywords from marketing articles, and retrieved keyword popularity metrics from Google Trends, Google Scholar, and ResearchGate. They found that using journal and author variables were better than including keywords-based popularity features. The three journal variables that were tested were Journal Impact Factor (JIF), 5-year journal impact factor (5-JIF), and SCImago Journal Rank (SJR), while the author related variable included was h-index. Another research found that social network variables created with centrality metrics from author collaboration networks have been found to improve predictions (Colladon, 2020).

A crucial question in the prediction of highly cited articles is the definition of what a highly cited article is. This definition determines the difficulty of achieving precise predictions. A higher threshold results in fewer highly cited articles to be identified, which leads to a more imbalanced and ultimately a harder classification problem due to less examples of the highly cited articles to learn from (Haixiang et al., 2017). The definition a highly cited article has varied greatly. Some have defined a percentage threshold such as the 10% most cited articles (Sarigol et al., 2014) or 25% most cited articles (Hu et al., 2020). Others have used average citations in a field as the threshold (Newman, 2014), the percentage of total citation counts (Wang et al., 2019), or a given amount of the most cited articles in a field, such as the 100 most cited articles (Abrishami & Aliakbary, 2019).

2.3.2 Machine Learning for Prediction of Future Citations

Various machine learning methods have been used for predicting article citations. The increase in computing power and advancements in machine learning methods have contributed to improvements in predictions of article citations. In recent years, neural network approaches have achieved promising results. Hu et al. (2020) tested various machine learning methods to predict whether marketing papers in three top marketing journals, Journal of Marketing, Journal of Marketing Research, and Marketing Science, were among the top 25% cited papers or not. Their results found Artificial Neural Nets (ANN) to perform best among Logistic Regression, C4.5, and Support Vector Machine (SVM). Furthermore, using the 5-year citation count as dependent variable, and not as a *result* of a predicted sequence, has yielded better results when using neural networks on citation count prediction (Ruan et al., 2020).

Tree-based methods have generally performed well. Sarigöl et al. (2014) found Random Forest to perform better than Naïve Bayes on a dataset of 36 000 articles. Similarly, Wang et al. (2019) found Random Forest to perform better than Naïve Bayes and KNN. Support Vector Machine (SVM) has also shown promising performances (Chakraborty et al., 2014; Xu et al. 2019). Finally, Logistic Regression has performed well, placing second among ANN, C4.5, and SVM (Hu et al., 2020).

2.3.3 Social Network Analysis for Predicting Future Citations

Research on networks can be tracked back to a well-known publication by Granovetter (1973), which is cited over 64,500 times. Through the years, social network analysis has become popular, and it has been used for studies of everything from organizations, countries (Quan-

Haase & Wellman, 2006), and journal articles (White et al., 2003), to twitter content (Yao et al., 2021), and friendship (Mutoh et al., 2016). The interest of using network analysis spans across all social sciences and is increasingly used in physics, biology, and other fields (Borgatti & Halgin, 2011). It has also been a popular tool in the marketing field (Webster & Morrison, 2004).

Different types of social network analysis can be applied to create variables for predicting future article impact. Co-authorship networks has been used to create centrality metrics which were able to predict with high precision the articles that were highly cited five year after publication (Sarigöl et al., 2014). Colladon et al. (2020) used social network analysis in combination with natural language processing of the abstract to understand which variables that were drivers of impact and could predict future success of chemical engineering papers. Furthermore, social network analysis has been used to identify which young researchers who are most likely to become successful measured by their h-index (Billah & Gauch, 2015).

The centrality measures in co-authorship networks describe the position of an author relative to others in a network (Costenbader & Valente, 2003). The centrality of an author can say something about how influential they are, or how important their field of research is. Among the most used centrality metrics are degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Centrality can be used as an important measure to find powerful and impactful authors in a network (Das et al., 2018). Only a few studies have focused on the way centrality of authors affect the citations of articles, and whether it is a driver of impact. Matveeva and Poldin (2016) found a positive relationship of the citation counts of scholars and the centrality of an author. Yan and Ding (2009) found that closeness, betweenness, and degree centrality correlated with the citation counts. Other research found that none of the centrality measures they investigated could accurately predict the future citation success of an author alone (Sarigöl et al., 2014). Biscaro and Giupponi (2014) found that centrality measures influence the early-stage citation counts. Work on citation prediction based on centrality measures for term-document networks found that document centrality measures offer a fairly high performance in identifying articles that contain a large number of impactful keywords (Klimek et al., 2016). Furthermore, Li et al. (2013) found that betweenness centrality had the most important role in using non-redundant resources in co-authorship networks, and therefore had a significant effect on citations.

3. Hypotheses

The literature review presented the extensive research that is done on drivers of citations for academic articles and the prediction of citations. However, there are some areas within the marketing field that still have not been investigated, and thus are valuable to explore. Followingly, we introduce three hypotheses that we will test. We introduce the gap in the literature, the motivation behind studying the hypotheses, and how we will test them.

Hypothesis 1

As presented in our literature review, previous studies have found strong evidence that early citation counts are a strong driver and predictor of accumulated citation counts in the following years. However, there is a gap in the previous research. As far as we have found, early citation counts, as a driver for future citations, have not been investigated in the field of marketing research.

The more time after publication we allow for, the more information we have about the article and its future impact. However, there is a trade-off with having to wait longer for this information. Since citation counts usually are monitored yearly, we therefore focus on the earliest full year of new citations after publication, and form the following hypothesis:

H1) Citations received the following year after publication highly influence the citation count of marketing articles five years after publication.

The motivation for testing this hypothesis is based on the availability and ease to check yearly citation counts for the readers of an article. If this hypothesis is confirmed true, then 1) the readers can use this as a tool to find articles with high future impact, and 2) authors of marketing articles can better understand the importance of early citations.

To test *H1*, we will use Spearman's rank correlation and multiple linear regression. The correlation will show us the strength of correlation between new citations after one year and citations after five years. The multiple regression will tell us whether the variable has a statistically significant effect on citations after five years. We will also compare the multiple linear regression with a linear regression using only citations after one year as the independent variable. By comparing the regression models, we can compare the effect of this variable compared to the other variables.

Hypothesis 2

From previous literature on the use of social network analysis for predicting citation counts, we found that centrality measures have an impact on future citation counts in other research fields. In addition, previous research has shown that author visibility and promotion is important to receive citations in the marketing domain (Stremersch et al., 2007). In several fields, centrality metrics has been shown to have a positive effect on the citations of publications. Hence, mapping the co-author network of authors in marketing looks promising.

The motivation for testing whether co-author network matters for five-year citation counts is that it can benefit authors of marketing articles to know how co-author collaboration affect article impact. In addition, there is little research on the effect of centrality measures on future citation count of marketing articles. Social network analysis can give valuable information about the structure in a collaboration network. As research has shown the benefits of using social network analysis in several fields, we are finding it advantageous to utilize it to look at the effects a collaboration network has on article impact on the marketing field.

Previous research by Sarigöl et al. (2014) have pointed out that a single network metric is not enough to capture the effect of social influence on article impact. Therefore, we want to look at the four most commonly used centrality measures: degree-, closeness-, betweenness-, and eigenvector centrality. The first three centrality measures have been included in several research articles with varying results. However, eigenvector centrality has rarely been investigated as a driver for article citation. We therefore formulate the following hypothesis:

H2) The value of centrality measures from co-authorship networks at the time of publication are drivers for marketing article citation count five years after publication

We use the same techniques as for *H1* to test *H2.* With Spearman's rank correlation we can see if there are any monotonic correlation between the centrality measures and citations after five years. In addition, a multiple linear regression will be used to investigate if there is a significant effect between any of the centrality measures and citations after five years.

Hypothesis 3

After studying relevant drivers of article impact, we are interested in how well these perform as variables for predicting future highly cited articles. There has been widespread research on which variables that are useful when predicting article impact. The variables can normally be split into three categories; paper, journal, and author related variables. Paper related drivers have been tested in several fields, often with results showing that variables like abstract length, sentiment, or lexical diversity are predictive for future citations. Both author- and journal related drivers have also been successfully used in combination to predict highly cited articles from three top marketing journals (Hu et al., 2020).

We argue that it is more relevant for the readers of marketing articles to know which articles will be highly impactful or not, than to know a given number of citations. By narrowing the number of articles that a reader needs to consider reading, less time is required to find new, impactful articles. As variables from all three categories of drivers individually has proved to be important to predict highly cited articles, it is interesting to investigate whether they work better in a combination. Hence, we present the following hypothesis:

H3) Using a combination of paper, journal, and author related drivers better predict the 5% most cited marketing articles five years after publication, compared to using drivers from only one of the categories

With a combination, we mean either using a combination of two or three driver categories. We chose to set the threshold for highly cited articles at top 5% for it to be a manageable number of articles to be suggested. We assume that it is limited how many articles one person is interested in considering reading. With over 2000 new articles yearly the last few years, and over 4000 in 2021, we argue that predicting the top 5% of those articles gives a manageable selection of articles to consider, and therefore is a relevant task to improve.

To test *H3*, we will create predictive models using four different methods. By using four different methods, we get a more accurate overview of the true predictive ability of the variable categories. We will create seven models per method; three for each of the separate categories, three where we combine two of the categories in all possible combinations, and one where we combine all the categories. This will make it possible to compare the variable categories, and test whether our hypothesis is true or not. The methods we use will be presented in the following section.

4. Methodology

This section includes an elaboration on the methods we have used to answer the hypotheses. It includes the regression model used to answer *H1* and *H2*, and the classification models used to answer *H3*, with its corresponding metrics to assess the classification models and the validity of *H3*. Lastly, we describe the network metrics present in *H2* and *H3*. We also explain the application of social network metrics for predictive purposes.

4.1 Machine Learning Models

A machine learning (ML) model is an expression of an algorithmic method applied to a specified type of observations (Parsons, 2021). There are a wide variety of models for creating a representation of a defined task, ranging from simple, two-dimensional, linear functions created by linear regression to complex, non-linear, high dimensional functions created by neural networks. What differentiates the ML models is how they estimate the expression, and how interpretable they are. Interpretability usually comes with the cost of having a more restrictive model in terms of how well it can shape to the patterns in the data. In our analysis, we will use a highly interpretable regression model to answer *H1* and *H2*. We will then use multiple different classification models to account for the possibly different patterns and relationships our variables can have to empirically test *H3*. We will in this section go through our models of choice for these tasks. The terms variable and feature will be used interchangeably depending on what is used in the theory.

4.1.1 Multiple Linear Regression

To help us answer *H1* and *H2*, this thesis will use multiple linear regression. The goal is to find the linear relationship between different variables (explanatory variables) and the citation count after five years (response variable). The multiple linear regression is a parametric model which has a high interpretability, and it can be defined as:

$$Y = \beta_0 + \beta_1 X + \ \beta_2 X_2 + \cdots + \ \beta_p X_p + \varepsilon,$$

where β_p is a coefficient which signal the effect each increase in value of variable X has on the outcome Y (James et al., 2013).

Multiple linear regression has the benefit of being able to look at several potentially important variables in one model. This may lead to a more accurate interpretation of the relationship of each individual variable with the response variable. In this way, it is possible to look at relationship of both paper, journal, and author related drivers with number of citations after five years.

As the regression coefficients are unknown, they must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, we can use the following formula to make predictions:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

The parameters are estimated by using a least squares approach where the coefficients are chosen to minimize the sum of squared residuals:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 . \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_p x_p\right)^2 \end{aligned}$$

The fit of a linear regression can be explained by R-squared (James et al., 2013). R-squared explains the proportion of variance explained, and can have a value between 0 and 1, where 1 means the fit is perfect. R-squared can be defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where TSS = $(y_i - \bar{y})^2$ and explains the total variance in the data, while RSS explains the explained variance of the regression line.

Multiple linear regression has multiple assumptions (James et al., 2013). First, it assumes that the relationship between the variables and response are linear. Second, it assumes that the error terms are uncorrelated. If this assumption is violated, then the estimated true standard errors tend to be underestimated, meaning that p-values will be lower than they should be. Third, it assumes non-constant variance of error terms, often referred to as heteroscedasticity. Violating this assumption, means that the estimated true standard errors can be too wide or narrow. Additionally, there are several considerations such as outliers, high leverage points, and collinearity. Both high leverage points and outliers can skew the fit to become less representative for most observations. Finally, presence of collinearity can pose difficulties in terms of separating the effects of individual colinear variables towards the dependent variable. This can result in a growth in standard deviation, and the p-value becoming higher than it should be.

4.1.2 Logistic Regression

To test *H3*, we will predict the articles that are among the top 5% cited articles after five years. While linear regression models are well suited when having a continuous dependent variable, it is not recommended for a binary classification problem. The logistic regression was therefore selected as one of the four models to test *H3*. Logistic regression is computationally efficient to fit to data and the results are easy to interpret. In previous research, logistic regression is used as a supervised classification method with good results for average success rate for prediction of citations (Ibanez et al., 2009). Hu et al., (2020) used logistic regression to investigate whether journal, author, or keyword-related features could better predict highly cited papers in highly rated marketing journals, and logistic regression performed well with all variable categories included.

Logistic regression calculates the probability that Y belongs to one of two categories (James et al., 2013). The difference from a linear model is that the response variable in a logistic regression is binary (Hosmer & Lemeshow, 2013). For example, the probability of a marketing article being one in the 5% most cited marketing articles after five years given its journal ranking can be written as:

$$\Pr(5\%\ most\ cited = True|\ Journal\ ranking)$$

This gives a probability between 0 and 1, where we can set a threshold in which the classifier will classify into a certain category beyond the threshold probability. The logistic regression uses the logistic function, which can be written as:

$$p(X) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\ \beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\ \beta_pX_p}}$$

where β_p is the coefficient value for predictor p with value X. The benefit of the logistic function is that we always get a value between 0 and 1, meaning that it works well for estimating probabilities. The coefficients, β_p are found by maximizing the likelihood function, where the likelihood function can be written as:

$$\ell(\beta_0, \beta_1, .., \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_i)),$$

where $p(x_i)$ is the observed probability for x_i to happen. By maximizing this function, we are maximizing the likelihood of observing the observations (X) in our data. Hence, if our training data reflects our test data, the logistic regression will generally work well.

4.1.3 Support Vector Machine (SVM)

The second model we will use to test *H3* is Support Vector Machine (SVM). SVM is a method for drawing a hyperplane to separate two classes by the largest margin. A hyperplane is essentially a boundary separating n-dimensions into two parts (James et al., 2013). This method has performed well in various settings and is good at handling large number of features on less amounts of data. Therefore, we want to test it to classify which articles are among the top 5% most cited. SVM models accurately predicted whether a biomedical article got a given number of citations within a threshold (Fu & Aliferis, 2010). Furthermore, Yan et al. (2012) found that support vector regression, a regression version of SVM, was one of the best methods for predicting citations.

SVM is a computationally effective way of enlarging the feature space for the support vector classifier to perform classification of two classes (James et al., 2013). The idea behind SVM is to choose a hyperplane that give the best generalization capacity, before finding the maximum margin between the two categories (Adankon & Cheriet, 2015). The solution to the support vector classifier is to calculate the following:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle,$$

where *S* contains the collection of indices of the non-zero vectors of the inner products of pairs of training observations. Note that α_i is only nonzero for the support vectors, which is why SVM is so computationally effective – it does not have to calculate all points.

SVM draws the hyperplane based on its kernel, which can either be linear or non-linear. In this thesis we will be using the linear kernel as this is the kernel which has performed by far the best in early testing of our data. A kernel is made up of a generalization of the inner product ($K\langle x, x_{i'} \rangle$) in the form of a function which quantifies the similarity of two observations. The idea of kernels is the possibility to map training data in a higher-dimensional feature space through a mapping function (Adankon & Cheriet, 2015). The linear kernel of two observations, $x_i, x_{i'}$ can be defined as:

$$K(x_i, x_{i'}) = \sum_{j=1}^{n} x_{ij}\, x_{i'j}$$

where n is the number of variables (dimensions). The benefit of using a kernel instead of an enlarged feature space with functions, is the computational benefit of avoiding working explicitly in the enlarged feature space. However, a weakness of SVM is that it does not perform very well with large data sets, due to rapidly increasing computation requirements (Nalepa & Kawulok, 2018).

SVM usually performs better than logistic regression when the classes are well separated, while logistic regression is often preferred in more overlapping scenarios. Hence, we include SVM to account for the possibility of the classes (top 5% vs non-top 5%) being well separated.

4.1.4 LightGBM

The third model we will use to test *H3* is a Gradient Boosting Model (GBM), called LightGBM. Gradient boosting combines the use of gradient descent with the boosting tree algorithm. GBM are decision-tree based models. This means that the feature space is divided into sections by splits (James et al., 2013). GBM are shown to be good at predicting article impact, and we therefore expect it to provide a representative result. Akella et al. (2021) used gradient boosting with good results to investigate how well altmetric features could be used to predict whether an article would receive more than a median number of citations. It was also used by Galli & Guizzardi (2020), where gradient boosting was the best of three models to find the top drivers for prediction of citations.

One of our main reasons for testing *H3* with this method is that it is non-parametric, meaning that it does not take assumptions about the distributions of our data, and nor does it classify the data into a theoretical distribution such as Logistic Regression. Instead, it learns the distribution by trial and error. This gives us a novel use of our variables. In GBM, trees are grown sequentially to iteratively improve the fit based on gradient information regarding the error from previously grown trees. This way, the model learns the importance of each feature. The sequential trees grown are weak learners. Weak learners are producing hypotheses about the data which are slightly better than random guessing and can with enough trials learn almost the entire distribution of the training data (Schapire, 1990). The way GBM learns is by creating new weak learners to be maximally correlated with the negative gradient of the loss function (Natekin & Knoll, 2013). A loss function is a function of the error of the predicted versus true outcomes of training observations. The gradient is the steepness of the slope of the loss function, and therefore tells the model whether it is improving from previous trees, and how steep the change of the loss is. If the gradient is negative, then the model is improving and descending to either a saddle point, a local optimum, or a global optimum depending on the specified hyperparameters, training data, and variables (Lee et al., 2016). The model change is determined by multiplying the gradient value with the learning rate specified. If the rate of change is small enough, then the model will end the search and the best model will be returned.

In this thesis we will use a GBM called LightGBM as it has better categorical variable support, while in general being faster than the other gradient boosting models (Bentéjac et al., 2020). LightGBM is created by Ke et al. (2017) and is a gradient boosting model developed by Microsoft. Usage areas include ranking, classification, and other machine learning projects. It has not previously been used to classify highly cited articles. However, it has been used for prediction within several different fields like chemical-toxicity, blood glucose, wind power, and cryptocurrency price trend (Zhang et al., 2019; Wang & Wang, 2020; Ju et al., 2019; Sun et al., 2020).

The primary benefit of LightGBM is the computational efficiency. Gradient Boosting can be very time consuming to use when handling big data, due to the high number of splits, data instances, and dimensions to consider. LightGBM reduces the computational complexity of all three tasks. The time spent on finding the optimal splits was reduced by implementing a histogram-based decision tree learning algorithm (Ke et al., 2017). The issue of data instances was reduced through Gradient-Based One-Sided Sampling (GOSS). GOSS excludes a significant part of data instances with small gradients and use the rest to estimate information gain. It can obtain quite accurate estimations of the information gain with a smaller data size because the larger gradients contain most of the information. The issue of high dimensionality was reduced with Exclusive Feature Bundling (EFB). EFB bundle mutually exclusive features to reduce the number of dimensions in the calculations. In sum, these improvements have created a significantly faster gradient boosting model.

A potential shortcoming with using LightGBM is that there is a risk of overfitting (TechLeer, 2018). This is because LightGBM split the tree leaf-wise, which can produce much more complex trees which are too specialized to the training data. However, overfitting is mostly a problem if the data it is trained on is very small. Our dataset should be large enough to avoid this shortcoming. More technical details about LightGBM can be found in Appendix A4.

4.1.5 TabNet

The fourth and final model we will use to test *H3* is a deep neural network model (DNN) called TabNet (Arik & Pfister, 2020). Gradient boosting models have been considered to be the best practice for tabular data (Shafi, 2021). However, in 2019, Google came with TabNet. This complex model is using neural network methods and was shown to outperform the tree-based models across several benchmarks. TabNet has not been used to predict article impact previously. However, it has recently been tested on forecasting of electric load, prediction of hospital no-show, and long-term rainfall with good results (Borghini & Giannetti, 2021; Boughorbel & Kadri, 2021; Xu et al., 2020; Yan et al., 2021).

TabNet seeks to improve the data representation capacity and the feature importance abilities of GBM (Arik & Pfister, 2020). To do this, they combine the advantageous feature space splitting from sequential decision trees with the data representation capacity of a DNN attention model. TabNet is inspired by the decision tree functionality of splitting the feature space into regions through decision boundaries, formerly presented in 4.1.4. TabNet creates hyperplane-like decision boundaries through creating linear combinations of features, where the coefficients of the hyperplanes are determined by sequences of attention. Attention, often called transformer, is a deep learning architecture which was first introduced by Vaswani et al. (2017). The primary strength of attention is its ability to compute data representations and identify the important parts of a sequence (Doshi, 2021). In TabNet, the sequence is a set of variables, which in our case is our paper, journal, and author variables which will be presented in 5.4. TabNet uses sequential attention on each observation to determine the variables which should contribute to each decision step. The feature selection is done for each observation and can be different for each output. The attention is done in sequence to make the model learn and reason why the variables with contributions from the previous sequence were contributing. The decision boundaries are then drawn based on the aggregated contribution of the variables from the sequences of attention, where the vote from each sequence is equally important. The aggregated contribution explains, like GBM, which features are more and less important. In sum, TabNet creates output predictions which are based on learning from the most contributing variables of each observation in the training set.

In sum, this creates a novel platform for testing our variables in *H3* as it with its sequential attention structure may utilize variables and connections between variables different compared to Logistic Regression, SVM, and GBM.

4.1.6 Predictive Evaluation Metrics

To evaluate the predictive performance of our models to determine the validity of *H3*, we will be using several well-known classification metrics, namely ROC AUC, F1, Precision, Recall, and Accuracy.

ROC AUC

ROC AUC is combining the true positive rate with the false positive rate (Burkov, 2019). One of the primary benefits of using it is that it is easy to evaluate if the model is better than a random classifier. An AUC higher than 0.5 means that the classifier is better than a random classifier, and the higher the AUC value is, the better the model generally performs. An additional benefit of ROC AUC is that it is differentiable, meaning that a model can use it as a loss function.

F1

The F1 score is the harmonious mean of precision and recall (Taha & Hanbury, 2015). It can be defined as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 is useful when recall and precision is equally important, as their importance is balanced. The F1 score is therefore highly useful to test *H3* as we would like a model which both can find as many of the top 5% articles as possible, while avoiding predicting many articles which will not become one of the top 5% cited articles.

Precision

Precision is the proportion of the predicted true cases which is correct (Burkov, 2019). It can also be interpreted as the probability that a predicted true positive in fact is a true positive. Precision can be defined as:

$$Precision \stackrel{\text{def}}{=} \frac{TP}{TP + FP},$$

where TP is true positives, and FP is false positives.

Precision is useful to measure in the cases where there is a high value in having correctly predicted true cases. In our case, having a high precision would mean having suggested less articles which will *not* become one of the top 5% most cited articles after five years. A higher precision is therefore highly useful in our case as it means that we have a more precise pool of true top 5% article predictions to consider reading through. Typically, a higher precision comes at the expense of lower recall and vice versa. Whether precision or recall is more valuable depends on whether the cost of missing a potential highly cited article is higher than the time usage cost of having to go through more suggestions.

Recall

Recall specifies how many of the true cases the model managed to predict (Burkov, 2019), and can be defined as:

$$Recall \stackrel{\text{def}}{=} \frac{TP}{TP + FN},$$

where TP is true positives and FN is false negatives.

In our case, a higher recall means that the model identifies more of the most impactful articles five years ahead. Having a higher recall typically comes at the expense of a lower precision score. This highlights the importance in our case of calculating a score which balances Recall and Precision, such as F1.

Accuracy

Accuracy is the proportion of correct predictions, where correct predictions are true positive and true negative cases (Burkov, 2019). Accuracy can be defined as:

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

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

The benefit of Accuracy is that it explains how well the model performs on predicting both true and false cases correctly. However, accuracy explain less about the performance of the model in cases where a model is set to predict an unbalanced dataset. If a dataset has two classes, where one class is 90% of the dataset and the other class is 10% of the dataset, the model would have got an accuracy of 90% by only predicting the largest class. This would not be better than simply having a random person guess the largest class each time. As a result, using Accuracy on an imbalanced dataset should be a supporting measure, and not the main measure of focus.

4.2 Social Network Metrics

In this section we will elaborate on social network metrics. Previous research has shown that the patterns in the network of authors can influence the number of citations an article gets. The centrality an author has in the network can therefore be an interesting variable to include in this thesis. In this section, we will describe what a social network is and why co-authorship networks can be a helpful tool, elaborate on how the measures are created, and lastly, we will describe the applicability of the centrality measures in our thesis.

Brass (2002) claims that network theory is about the consequences of variables in the network. Social network analysis aims to predict and test theories about the structure of relationships among social entities (Butts, 2008; Wasserman & Faust, 1994). A social network consists of a set of nodes that are tied by one or more relations (Wasserman & Faust, 1994). Nodes are also called network actors and can be defined as units that are connected by the relations of the pattern that are being studied. Any unit that can be connected to other units can be studied as nodes, but the most common units are either persons or organizations (Marin & Wellman, 2010). The relational ties are connecting the nodes, and can be the transfer of resources, friendship, web links, or any possible connection.

In a co-authorship network, the relational ties reflect the collaboration between two authors. The ties form paths that indirectly link nodes by interconnecting through shared endpoints. It is the patterns created in a network that may be used for analysis and prediction (Borgatti & Halgin, 2011). The network can be used to create centrality measures. Where citation counts can be a measure of impact of articles, centrality can express the impact of an author on the field (Yan & Ding, 2009). The centrality of an author says something about how central, or powerful, the author is in the network of authors. Furthermore, having a strong centrality have in some cases been shown to have an impact on article citations.

To perform a co-authorship network analysis the co-authorship must be formatted into specific adjacency matrices, edge lists, or adjacency lists, to map the relationships between the nodes (Fonseca et al., 2016). When a pair of nodes share authorship of the same article, the intersection gets the number 1. Otherwise, it gets the number 0. When authors collaborate more than once, the number is equal to the total articles co-authored. Furthermore, the data can be visualized as a network or statistical analysis, and metrics can be calculated and interpreted.

4.2.1 Centrality in Networks

The position of individual actors in a network can provide important insights. The position of a node can reflect power or prestige, depending on the characteristics of the linkages (Giuliani & Pietrobelli, 2011). Nodes that are central may be in advantageous positions, relative to less central nodes (Freeman, 1979). Centrality can be a useful measure of the impact of a node. The higher the centrality, the stronger the influence. The centrality measures have shown to work as drivers of impact in some scientific fields. We use centrality measures to test *H2*. Thus, we look closer at four centrality measures: degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. These centrality metrics are commonly used to find central actors in networks.

Degree centrality

One of the most central methods of measuring centrality is by counting the number of direct ties a node has to other nodes in the network, called *degree centrality*. It represents the total strength of the direct connections of a node. The equation for degree centrality is as follows:

$$C_D(n_i) = d(n_i),$$

where $d(n_i)$ is the degree of author n_i.

Degree centrality can be used to find nodes who quickly connect with the wider network, and that have easier access to resources and information (Giuliani & Pietrobelli, 2011). However, in some cases having a central position can be disadvantageous. For example, Paruchuri (2010) found that the innovative performance follows an inverted U-shape as information above a certain level can overwhelm the node. In previous work on scientific collaborations, the degree centrality was identified to have a positive correlation with the citation count of an academic article (Uddin et al., 2013). This indicates that authors with more connectivity are contributing to knowledge creation in a higher scale than other authors. Thus, degree centrality can help identifying the authors in the marketing field that are contributing with knowledge and innovation.

Betweenness centrality

In certain situations, the perks of being central is due to the control a node have over the flow of information. *Betweenness centrality* measures how many times a node lies in-between the shortest network paths that connect the other nodes (Wasserman & Faust 1994). The power of the centrality is related to being essential to the network in the terms of flow of information. A high betweenness centrality serve as an indirect connection between pairs, leading to the high brokage power (Borgatti et al., 2013). Betweenness centrality formula is as follows:

$$Betweenness\ centrality = \sum_{\mathrm{j<k}} \frac{g_{jk}(i)}{g_{jk}},$$

where $g_{jk}(i)$ is the number of the shortest paths that includes node *i.* To normalize the formula, it can be divided by its maximum $(n-1)(n-2)/2$.

The betweenness centrality of an author can be seen as a driver of impact, as it has been shown in research that it correlates with the citation count (Uddin et al., 2013; Yan & Ding, 2009). Authors with high degree centrality are likely to have access to rich knowledge, which in turn can increase the quality of their publications. Betweenness centrality can therefore show us which authors within the field of marketing research that have control of information flow, and we will test whether it is an important driver of the citations of a marketing article.

Closeness centrality

To measure the embeddedness of a node in a network, one can use *closeness centrality*. Short distances means that the closeness is higher (Freeman, 1978). The closeness is measured as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. The formula for closeness centrality is as following:

$$Closeness\ centrality = \frac{1}{\sum_{j=1}^{n} d_{ij}},$$

where d_{ij} is the length of the shortest path that are connecting the nodes i and j. To normalize the closeness centrality, one can multiply its value by the maximum $(n-1)$ (Colladon et al., 2020).

Closeness centrality is measuring network property rather than academic impact, and several studies has shown that closeness centrality has low or no significant influence on citation counts (Yan & Ding, 2009; Uddin et al., 2013). This thesis will test if this is the case for marketing research, or if the positioning of an author has a significant effect on the future citation count of a marketing article.

Eigenvector centrality

Eigenvector centrality is a metric used to measure the influence of a node within a network (Golbeck, 2013). A node in the network is more central when the nodes connected to themselves are well-connected. Connections to more influential nodes are more important than the connections to less influential nodes (Zhang et al., 2021). To find the eigenvector score one has to solve the following equation:

$$A \times c = \lambda \times c,$$

where A is the adjacency matrix for a graph, c is a vector of the degree centralities of each node (eigenvector) , and λ is a scalar (eigenvalue) (Cimenler et al., 2014). The eigenvector centrality has been found to have a positive effect on productivity of authors (Ariel Xu et al., 2020). In addition, it computes the importance of an author by how frequently they collaborate with other important authors (Diallo et al., 2016). However, we were not able to find previous research on the effect it has on article impact. Thus, we want to test whether it influences the number of citations.

4.2.2 Application of Centrality Measures

Given the promising usage of co-author collaboration network by Colladon (2020) on Chemical Engineering paper citation prediction, we find it promising to represent the social network of an author in a co-author network to investigate *H2* and *H3*. In this section we will therefore explain how we apply the theory about centrality measures to our thesis.

We created yearly co-author collaboration networks from 1992 to 2016 of all authors of our 33 754 marketing papers published between 1992 and 2016. The networks were created yearly to avoid information leakage about future social network of an author This data is further explained in section 5. This allowed us to calculate each authors centrality scores for each year. Thus, we created in total 25 co-author networks. The scores were then linked up with the authors of articles based on the publication year of the articles.

In the cases where there was more than one author on an article, we considered only the best centrality scores across the centrality scores from all authors of the article. Hence, the best degree, betweenness, closeness, or eigenvector centrality could come from different authors. The intuition behind this is that we assume that the author group utilize the network strengths of the author(s) within the author group. In addition, to make it easier to compare the values we have chosen to normalize all the measures between 0 and 1. This is further explained in section 5.3. Moreover, each of the different centrality measures will be introduced again when we present our author related variables in section *5.4.3 Author related variables*.

5. Data

5.1 Software

To carry out our analysis, we have used the programming languages R (4.0.4) and Python (3.7.1). R have been used through RStudio, and Python have been used through Visual Studio Code (VScode). All parts of our data preparation and cleaning have been done with R libraries, which can be found in section A1 in the appendix. All predictions have been done in Python through libraries which can be viewed in appendix A2.

5.2 Data Collection

In this part we will present the data we used and where we retrieved it from. Since our field of interest is marketing, we first had to find which articles we would consider as marketing articles. There are numerous overviews of journals such as Financial Times, Journal Citation Reports (JCR), SCImago Journal Rank (SJR), the Source Normalized Impact per Paper (SNIP), Citescore, and The Chartered Association of Business Schools (ABS). However, Financial Times provides only a small selection of top journals. JCR, SJR, SNIP, and Citescore on the other hand, provide a large selection of journals, but they are only quantitatively based. ABS, however, has a large selection of 76 reviewed and ranked marketing journals. Their primary differentiation is that they combine metrics from JCR, SJR, SNIP, and Citescore with peer-review, editorial, and expert judgements of each journal's research standard (Chartered Association of Business Schools, 2021). We therefore selected journals based on the ABS list of marketing journals as it provided both a wide selection of journals and a thorough quantitative and qualitative selection process.

Today there are three primary sources for collecting Scientometric data about scholarly articles: Google Scholar (GS), Web of Science (WoS), and Scopus (Mingers & Leydesdorff, 2015). Each source comes with its own benefits and disadvantages. The main benefits of GS are the citation coverage, and the larger amounts of total articles available (Martín-Martín et al., 2018). However, GS suffers from a lower quality data for scientometric research (Mongeon & Paul-Hus, 2015; Mingers & Leydesdorff, 2015). WoS has the primary benefit of having article coverage dating back to 1900 (*Web of Science Group*, 2019), but has a significant underrepresentation of journals within marketing according to the 2021 ABS list of journals.

In addition, all the marketing journals which can be found in WoS, are also present in the Scopus database. While Scopus only has article coverage until the 1970s, it has similar data quality to WoS (Mongeon & Paul-Hus, 2015), but a significantly larger journal coverage for marketing journals in the ABS list. In fact, we found Scopus to have all 76 marketing journals in the ABS list, while WoS only was listed to have 42 (*Academic Journal Guide 2021*, n.d). In addition, Scopus is one of the most popular databases, created by Elsevier which is a highly acknowledged publisher of quality journals. It has a large coverage, is up to date, and offers rich metadata. As a result, we chose to use Scopus as our data source.

The data used for this thesis was retrieved from Scopus and ABS, with the final dataset being a merge of *four* datasets: two article datasets, an ABS-list ranking dataset, and a yearly citation count dataset. The data from Scopus was retrieved using a query, specified in appendix A5, where we specifically requested articles written in English from the 76 marketing journals in the ABS list. The main article data was imported to a data frame using the *convert2df* function in the Bibliometrix R library as this gave us the correct data format to later create co-author networks in this package (Aria & Cuccurullo, 2022). To get more details on the publication date of an article, and whether an article is open-access or not, we retrieved a second dataset with additional data from Scopus through the Rscopus package (Muschelli, 2019). The third dataset we imported was the yearly citation count data containing data about new citations retrieved each year from 1992 to 2022. Details about the Scopus datasets can be seen in appendix A6. The fourth and final dataset was imported from ABS (*Academic Journal Guide 2021*, n.d) to create our ABS Academic Journal Guide (AJG) variable. The contents of this dataset can be seen in appendix A7.

5.3 Merging and Pre-processing the Datasets

Given that we had in total four datasets merged to create all variables which we wanted to analyze, pre-processing steps and decisions were required. Since some articles had missing data, several pre-processing steps were necessary to get an accurate data set for further analysis.

Merging

First, the four datasets were merged to one dataset. The yearly citation dataset and article datasets were merged by DOI, as this is a unique identifier for each article. The ABS dataset was merged by ISSN, as this is a unique identifier for journal publications.

Missing Values

Missing values (NA) can signal at least four things: 1) that information is missing but existing elsewhere, i.e., on WoS but not Scopus, 2) that information has been lost in merging or pre-processing issues, 3) that the information simply does not exist, or 4) that the missing values reflect a zero value. There are numerous removal or imputation options for observations with missing values, depending on whether the missing value is missing at random (MAR), missing completely at random (MCAR), or missing from the subpopulation (Acock, 2005).

We *removed* articles which were missing one or more of the following variables: 1) DOI, 2) citations used in article, 3) page number, 4) open-access info, 5) publication month, and 6) author. These variables are seen as vital to the quality of the data, and imputing values for variables such as author or page-number would not be feasible. Furthermore, we could not find the missing values to have a systematic pattern and therefore assumed that these were MCAR and that removal of these would therefore not bias our dataset in a major way. The removal of missing values reduced our dataset from 35 409 articles to 33 754 articles and it went from having articles from 76 to 75 journals. The journal that was removed contained articles without DOI and were only a small subset with translated articles, as its original language was not English.

Furthermore, we *imputed* NA values under author keywords to 0 as we found articles with NA values to not have published author keywords. In addition, we *kept* articles with missing abstract as we found most of the articles with NA value in abstract to not have an abstract in their article. However, there were articles with missing abstracts on Scopus which had abstracts in their pdf text. Our solution to this was to create a dummy variable *contains_scopus_abstract* to account for this.

Tokenization

To create natural language processing (NLP) variables from the abstract, we had to create a separate tokenized data frame of each articles abstract. Tokens are words, groups of characters, numbers, or special characters. They are found by splitting a text into countable, unique pieces of text elements. We used the function *unnest_tokens* from tidytext (De Queiroz & Fay, 2021) to create a tokenized data frame.

Tokenizing a text can result in non-meaningful tokens, which neither represents sentiment, complexity, or diversity of a text, such as a group of numbers, or special characters. We therefore chose to remove digits by filtering out any digits through a regular expression.

Standardization

To help with interpreting the coefficients of the average sentiment of the article abstract, we used standardization. This was possible since the variable were approximately normally distributed, but not around zero, while we wanted to center the distribution around zero. This allowed us to interpret whether having a more positive or negative abstract than the majority of marketing papers are positively or negatively correlated with citation counts. To do this, we used the Z-score normalization formula as mentioned by Patro & Sahu (2015) on all abstract sentiment scores *i*:

$$x_{scaled_i} = \frac{x_i - \mu}{\sigma},$$

where μ = mean abstract sentiment, and σ = standard deviation. Variables which have been standardized contain the word standardized in the variable name.

Normalization

Since the co-author centrality variables had different absolute value ranges, we used normalization to be able to compare the coefficients of our co-authorship centrality variables. Normalization scales all values to become between 0 and 1, where 0 is the lowest value and 1 is the highest value of a variable. Normalizing observation i of a variable x can be written as:

$$x_{scaled_i} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

5.4 Variables

Based on promising drivers from the literature review, we created variables based on our pre-processed and merged dataset. From our background research it became evident that article citations can be explained from three major categories of drivers: paper, journal, and author related drivers. We wanted to have variables which approximate the most important aspects discussed in 2.2 Drivers of Article Citations. Hence, the following tables present the selected variables within each driver category.

5.4.1 Paper Related Variables

Variable name	Description	Creation
First_page	Represents paper quality. The placement of an article in a journal has been proposed to signal its perceived quality and importance in the journal.	Created by splitting PP (Pages from-to in journal) into two columns and used the from-page column.
PAGES	Represents paper importance. From the literature review, we found that number of pages can indicate the quality and importance of an article, where a longer article is perceived to be of better quality and importance as journals can often be restricted by length.	$first\ page - last\ page$
Keyword_number	Represents paper content. Increasing number of keywords has previously been shown to impact article citations.	Created by counting the number of words separated by a semicolon.
AB_words	Represents abstract content. In our literature review we found research from a more generalized dataset that suggested that the length of abstract may impact citation count, and we therefore wanted to test this on a marketing dataset.	Created by counting the number of words in abstract separated by space.
Lexical_diversity_normalized	Represents the complexity of reading the paper content. While increasing lexical diversity in abstract has been found to positively impact citations in a technical field, we are curious as to whether the same can be said for a non-technical field such as marketing.	$\frac{unique\ abstract\ words}{abstract\ words}$ This result in a normalized score between 0 and 1, where 1 means that all words are different, and a score of 0 means that all words are the same.
citations_in_article	Represents the solidity and plausibility of the paper. As presented in the literature review, more citations in an article can mean that more thorough research is done and therefore be a proxy of solidity and plausibility.	Created by counting the number of references separated by semicolon.
ave_sentiment_standardized	Represents the paper content. As we presented in the literature review, sentiment of abstract has	The sentiment was calculated by using the *sentiment_by*

	been implied as a variable which impacts future citation counts.	function in the package sentiment (Rinker, 2021). We also standardized it as mentioned in 4.3.4 standardization.
pub_month	Represents the publication month of the paper. To compensate for the lack of monthly citation data, we included month of publication as a variable. This variable is created to correct for the head start article gets when they are published early in a year in terms of having more time to accumulate citations.	Created by extracting the month from the publication date variable.
openaccess	Represents the availability of the paper as our literature review found documents which were more easily available were more likely to be cited.	Retrieved from Scopus
contains_scopus_abstract	Represents the availability of the paper content. Our data showed that several articles had missing abstract on Scopus. To account for this, we created a dummy variable.	Created by setting a value 1 if the article has NA value on abstract and 0 if not.

Table 1: Paper Related Variables

Time-Based Paper Related Variables

Variable name	**Description**	**Creation**
new_citations_after_1y	The number of new citations received the year after publication.	Retrieving the new citations received the year after the publication year

Table 2: Time-Based Paper Related Variables

5.4.2 Journal Related Variables

Variable name	Description	Creation
AJG2021	Represents journal ranking, impact, and quality. From our literature review, we found that journal ranking, and impact were found to be predictors of article citations on a wide range of academic fields, including marketing. While the metrics purely based on weighted average of journal metrics has been well tested, using rankings based on peer review, editorial and expert judgements assessments has yet to be done. Hence, we test ABS AJG as a measure for journal quality and impact.	Data retrieved from: https://charteredabs.org We then merged the score by ISSN number.

Table 3: Journal Related Variables

5.4.3 Author Related Variables

Variable name	Description	Creation
Author_count	Accounts for the increased probability of having a better author network due to being more authors on a paper.	Counted the author names separated by semicolon.
normalized_AU_degree	Represents the general academic network of each author up to a given year. Higher degree centrality means that an author has written articles with more people in the network.	See 4.2.2
normalized_AU_betweenness	Represents knowledge and brokerage power of the academic network of each author up to a given year. A higher betweenness centrality means that an author has a higher brokerage power in terms of information flow, as more people must go through this author to get in contact with other people.	See 4.2.2

normalized_AU_closeness	Represents the proximity to other in the academic network of each author up to a given year. Lower closeness centrality means that an author has a shorter path to all other marketing academics in the network.	See 4.2.2
normalized_AU_eigenvector	Represents the connectivity an author has to other authors with great academic network of each author up to a given year. A higher eigenvector centrality means that an author has written with many well-connected people in the marketing academia.	See 4.2.2

Table 4: Author Related Variables

5.5 Data Validation

Since we want to predict future values, it is essential to avoid information leakage from years beyond the years which the models are trained on. This is to ensure that a model is not getting help from future events, leading to an artificially high performance. Therefore, for the variables which are time sensitive, we try to ensure that they are created based on the data which was available *up to* the publication year of the different articles. For instance, the co-author network variables are created for each year, with data only about their co-author collaborations up to the publication year, to avoid their future collaborations to impact their scores. We have, however, included one variable from the future on all observations from 1992 to 2022 and that is the ABS journal ranking. The reason for using the 2021 ranking on all years is first and foremost that there is no AJG ranking prior to 2010, meaning that we would have uncertain AJG values for the articles in the 1990s even with using AJG2010. Secondly, the rankings have stayed almost identical from 2010 to 2022 and we therefore consider the journal rankings to be lowly biased from future events. An additional reason is that the AJG2021 ranking has rankings for more journals compared to AJG2018, AJG2015, and AJG2010.

5.5.1 The Validation Set Approach

To ensure validity when testing *H3,* we will be using the validation set approach. The validation set approach means that we will train the model on a training set and test it on a separate test set (James et al., 2013). This is done to estimate the predictive ability of the model on data which it has not been fitted on.

5.5.2 The Training and Test Set

To test the third hypothesis, we tested the predictive models on articles published in 2015 and 2016. This allowed us to test the models on yearly variations, while avoiding sacrificing training data size too much. To prevent training a model with citation data from a time-period which we are predicting the test error on, we avoided training the models on data five years prior to the test set articles. This meant that we could only use data from 1992 to 2009. The reason for not using articles after 2009 is that articles from 2010 received their fifth-year citation count in 2015, which is included in the test set. Including this citation count in the training set would therefore reveal information about the citation behavior present in the test set.

Our training set contains 17802 articles, and the test set contains 4993 articles, meaning that we had a 77/23 split. The test set contains marketing articles published on Scopus in 2015 and 2016 in 75 marketing journals. 525 of the test set articles have above 44 citations and were categorized as top 5% cited articles after five years, while 4468 were not. This is the data set used to test *H3*.

5.5.3 The Full Dataset

The full dataset consisted of 33 754 marketing articles published between 1992 and 2016 from 75 of the 76 ABS-ranked marketing journals available on Scopus. It also contained five-year citation count for all articles. The full dataset contained 21 columns, and an overview of these columns can be seen in appendix A8. The descriptive statistics about each column can be seen in appendix A9. This is the dataset used to test *H1* and *H2.*

6. Results

In this part, we present the results from the descriptive and predictive part of our analysis. In the first part, we will investigate our drivers and how they affect article citations after five years. From this part, we will draw conclusions on our first and second hypothesis. In the predictive part, we will investigate our third hypothesis by predicting the 5% most cited marketing articles five years post publication. From our dataset, we found that the 95th percentile of five-year citations of articles published between 1992 and 2016 were 44 citations. Hence, for an article to be among the 5% most cited articles after five years it must have above 44 citations.

In Figure 1, we present our data on the yearly number of English-written marketing articles published on Scopus from any of the 76 ABS listed journals published between 1992 and 2022. We find that the increase in new marketing articles per year has increased with almost 70% in the last five years from 2646 new articles in 2016 to 4489 new articles in 2021. This shows the increased importance of predicting future highly impactful articles.

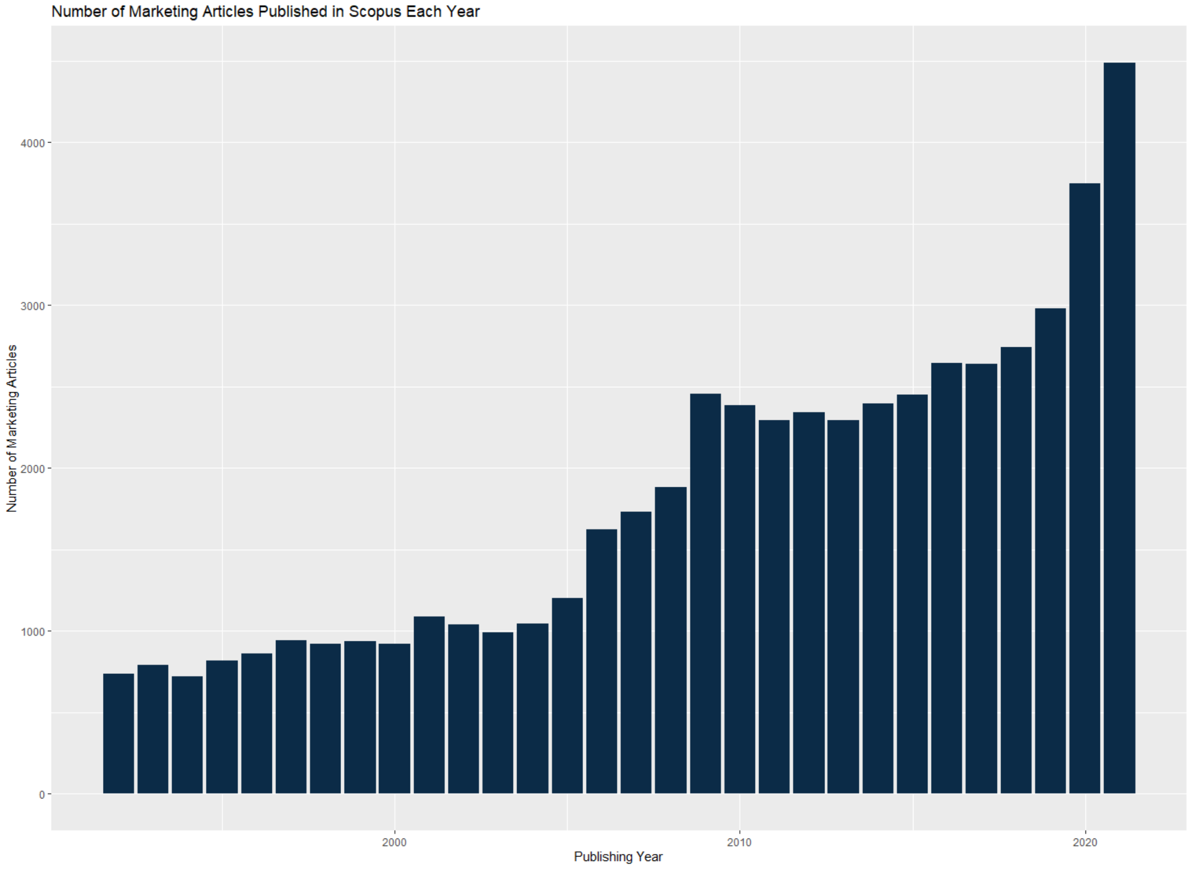


Figure 1: Number of marketing articles published in ABS ranked journals in Scopus per year according to our data.

6.1 Identifying drivers of five-year citation count

To identify the promising drivers of five-year citation count for academic marketing articles, we first looked at the correlation between our variables and the five-year citation count, as this method can reveal strong predictors. Secondly, we made multiple linear regression models to examine the variables joint linear relationships and statistical significance. In this section, we will be using the full dataset of marketing articles published between 1992 and 2016, which is the dataset presented in in 5.5.3 containing 33 754 articles from the 75 marketing journals.

6.1.1 Correlation of Variables

Since we had categorical variables and outliers, we chose to use Spearman's rank correlation. This correlation is monotonic, which makes it less sensitive to outliers and able to handle categorical variables. Monotonic correlation has previously been used to identify promising variables (Ruan et al., 2020; Chakraborty et al., 2014). The correlation plot will give us an impression of which variables have a strong relationship with the dependent variable.

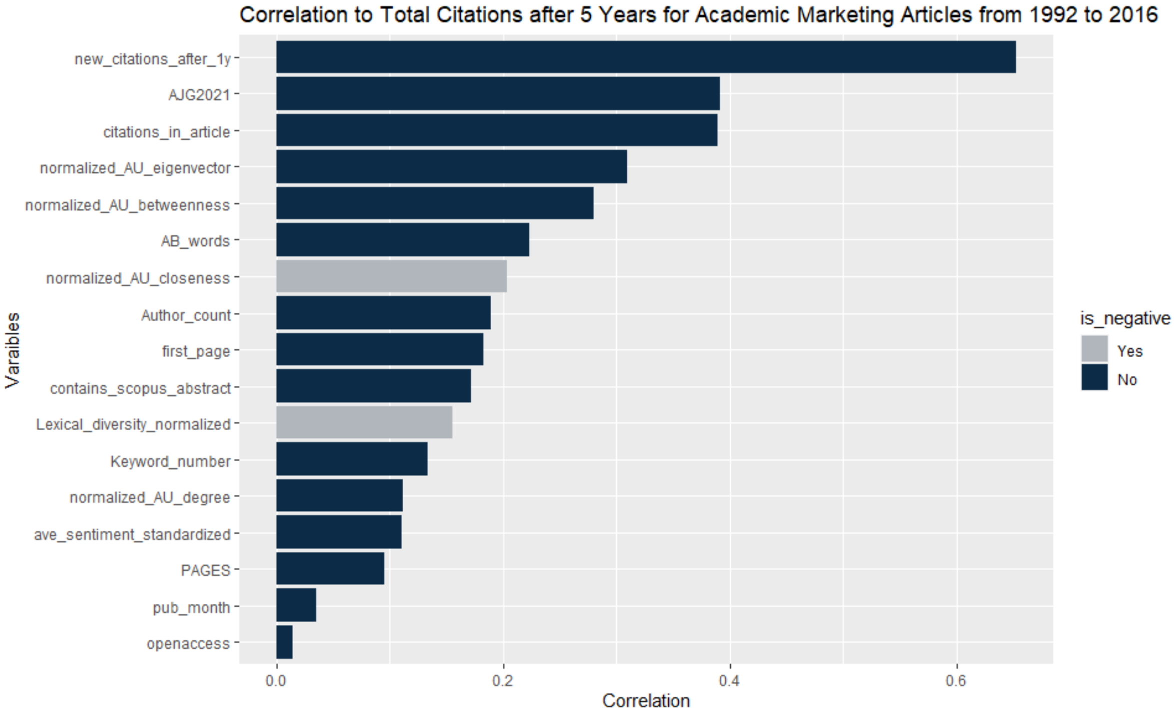


Figure 2: Spearman rank correlation of our selected variables

A correlation above 0.3 is considered moderate, while a correlation above 0.6 can be categorized as strong (Akoglu, 2018). In the correlation bar plot, the blue color represents a positive correlation, while the grey color represents a negative correlation. We see from Figure

2 that *new citations after one year* is strongly correlated to citations after five years, followed by *AJG2021*, *citations in article*, and *eigenvector centrality* at a moderate correlation. As citations after one year is a part of the total citations after five years, it is no surprise that it is highly correlated. The results therefore encourage that *H1* might be true.

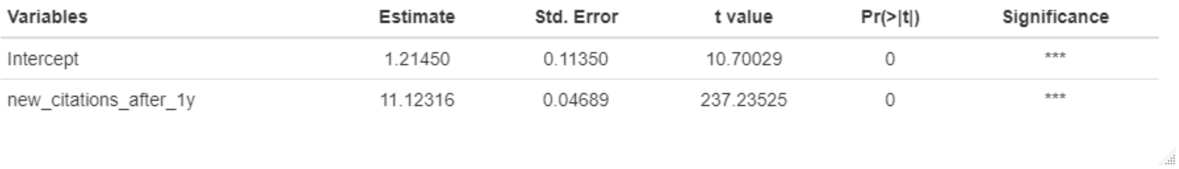
6.1.2 Regression Results

To get an understanding of our variables and help us answer the first part of our research question, we divided the multiple linear regression into six different regressions: new citations after one year, all variables, paper related, author related, journal related, and paper-, journal-, and author related combined. The regressions are created to identify their linear significance given the interactions with each other, as well as alone within their category. From the regressions, we will also determine the validity of *H1* and *H2*.

New citations after 1y regression results

To understand the individual effect of citations obtained the year after publication, we created a linear regression, which could be formulated as following:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{new\_citations\_after\_1y} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	1.21450	0.11350	10.70029	0	***
new_citations_after_1y	11.12316	0.04689	237.23525	0	***

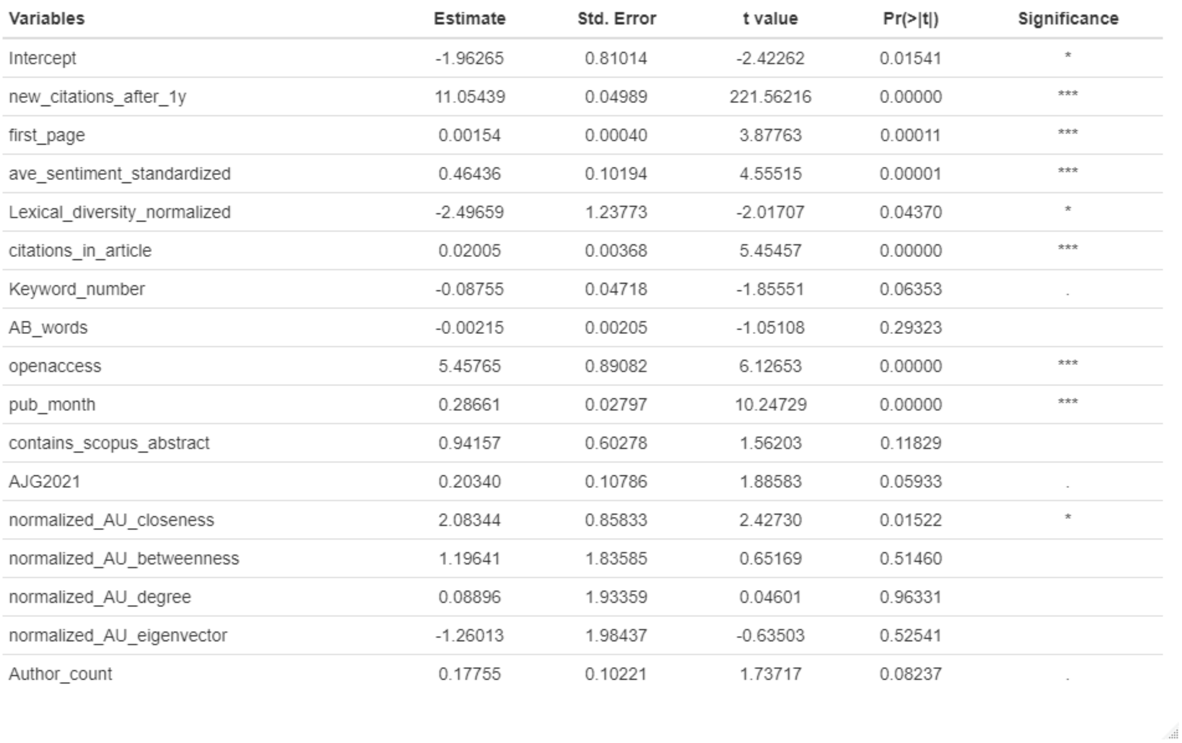
Table 5: Regression results using one year citation count

When only using the *new citations after one year* variable we get an adjusted R-squared of 0.6251. This value is quite high, meaning that the variable explains much of the variance in citations after five years. Thus, the explanatory power is relatively high, even when only having one variable. From Table 5, we see that the coefficient estimate is a 11.12 increase in total citations after five years for each additional new citation the following year after publication. This result in, combination with the Spearman’s rank correlation, indicates that this variable is a significant driver of total marketing article citation count after five years, supporting *H1*. In addition, it is a promising variable for our predictions.

Full regression results

To understand the added effect of paper, journal, and author variables, we created a full regression including both *new citations after one year* and all the variables from the other categories. The regression is expressed as following:

$$total\_citations\_after\_5y \ = \beta_0 + \beta_1 X_{new\_citations\_after\_1y} + \beta_2 X_{first\_page} + \beta_3 X_{ave\_sentiment\_standardized} + \beta_4 X_{\text{Lexical}\_diversity\_normalized} + \beta_5 X_{Keyword\_number} + \beta_6 X_{citations\_in\_article} + \beta_7 X_{AB\_words} + \beta_8 X_{openaccess} + \beta_9 X_{pub\_month} + \beta_{10} X_{contains\_scopus\_abstract} + \beta_{11} X_{AJG2021} + \beta_{12} X_{Author\_count} + \beta_{13} X_{normalized\_\text{AU\_closeness}} + \beta_{14} X_{normalized\_AU\_betweenness} + \beta_{15} X_{normalized\_AU\_degree} + \beta_{16} X_{normalized\_AU\_eigenvector} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	-1.96265	0.81014	-2.42262	0.01541	*
new_citations_after_1y	11.05439	0.04989	221.56216	0.00000	***
first_page	0.00154	0.00040	3.87763	0.00011	***
ave_sentiment_standardized	0.46436	0.10194	4.55515	0.00001	***
Lexical_diversity_normalized	-2.49659	1.23773	-2.01707	0.04370	*
citations_in_article	0.02005	0.00368	5.45457	0.00000	***
Keyword_number	-0.08755	0.04718	-1.85551	0.06353	.
AB_words	-0.00215	0.00205	-1.05108	0.29323	
openaccess	5.45765	0.89082	6.12653	0.00000	***
pub_month	0.28661	0.02797	10.24729	0.00000	***
contains_scopus_abstract	0.94157	0.60278	1.56203	0.11829	
AJG2021	0.20340	0.10786	1.88583	0.05933	.
normalized_AU_closeness	2.08344	0.85833	2.42730	0.01522	*
normalized_AU_betweenness	1.19641	1.83585	0.65169	0.51460	
normalized_AU_degree	0.08896	1.93359	0.04601	0.96331	
normalized_AU_eigenvector	-1.26013	1.98437	-0.63503	0.52541	
Author_count	0.17755	0.10221	1.73717	0.08237	.

Table 6: Full regression results from 1992-2016

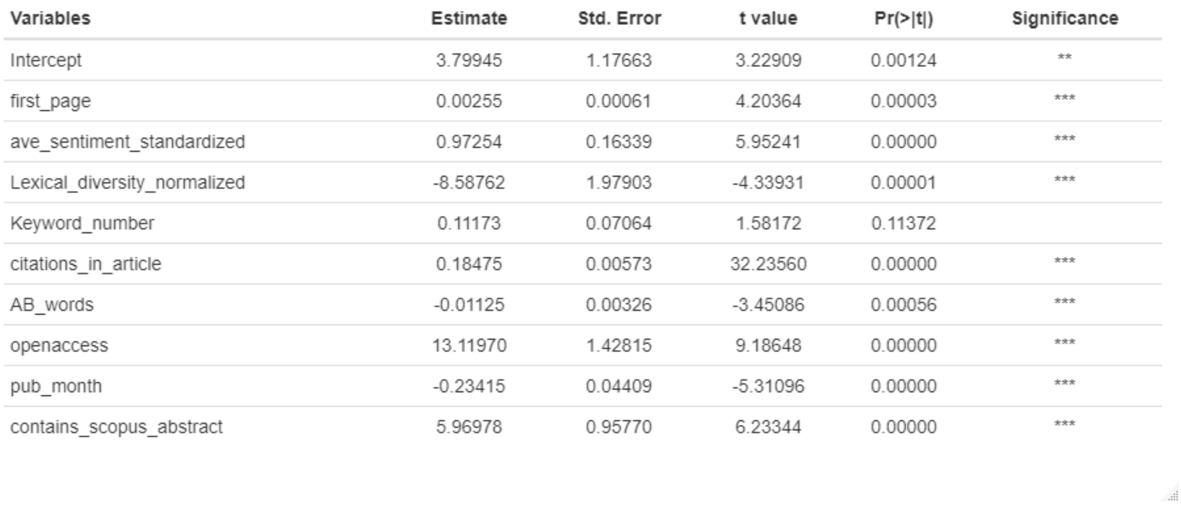
The regression had an adjusted R-squared of 0.6283, which is only slightly higher than the R-squared of 0.6251 from only including citations after one year. We notice that *new citations after one year, placement in journal, abstract sentiment, lexical diversity, number of citations in article, open-access, month of publication*, and *closeness centrality* is significant below a 5% level. *Closeness centrality* is the only centrality score variable which is significant. Surprisingly, the *AJG ranking* have a p-value slightly above 5%.

Another observation from this regression is the large coefficient value of the *new citations after one year* variable compared to the other variables. The regression containing only citations after one year as a variable had only a slightly lower R-squared value compared to the full regression. This indicates that the *new citations after one year* variable is explaining most of the variance of the variables included. Therefore, the results further support *H1* which states that citations after one year is a highly influential driver for article citations after five years.

Paper related regression results

To investigate the specific significance of our paper related variables, we constructed a multiple linear regression which could be formulated in the following:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{first\_page} + \beta_2 X_{ave\_sentiment\_standardized} + \beta_3 X_{Lexical\_diversity\_normalized} + \beta_4 X_{Keyword\_number} + \beta_5 X_{citations\_in\_article} + \beta_6 X_{AB_{words}} + \beta_7 X_{openaccess} + \beta_8 X_{pub\_month} + \beta_9 X_{contains\_scopus\_abstract} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	3.79945	1.17663	3.22909	0.00124	**
first_page	0.00255	0.00061	4.20364	0.00003	***
ave_sentiment_standardized	0.97254	0.16339	5.95241	0.00000	***
Lexical_diversity_normalized	-8.58762	1.97903	-4.33931	0.00001	***
Keyword_number	0.11173	0.07064	1.58172	0.11372	
citations_in_article	0.18475	0.00573	32.23560	0.00000	***
AB_words	-0.01125	0.00326	-3.45086	0.00056	***
openaccess	13.11970	1.42815	9.18648	0.00000	***
pub_month	-0.23415	0.04409	-5.31096	0.00000	***
contains_scopus_abstract	5.96978	0.95770	6.23344	0.00000	***

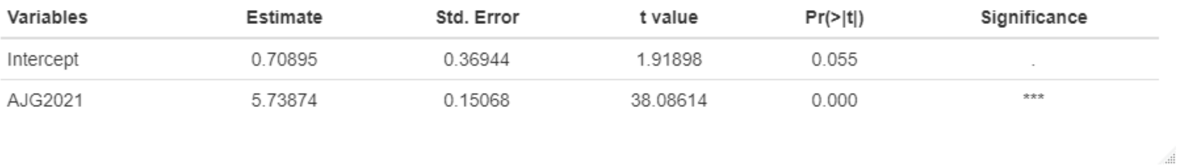
Table 7: Paper related regression results

The multiple regression of paper related variables has an adjusted R-squared of 0.04264. From Table 7, we find all variables except *number of keywords* to be significant below a 5% level. The results support some of the findings of research in other fields, and it finds that these variables can also be considered as drivers of impact in the marketing field.

Journal related regression results

The purpose of the journal related regression was to first understand the importance and significance of *AJG2021*, and secondly understand which AJG ranking increase gives the highest increase in expected citations after five years. To understand the first aspect, we included AJG2021 as a single variable:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{AJG2021} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	0.70895	0.36944	1.91898	0.055	.
AJG2021	5.73874	0.15068	38.08614	0.000	***

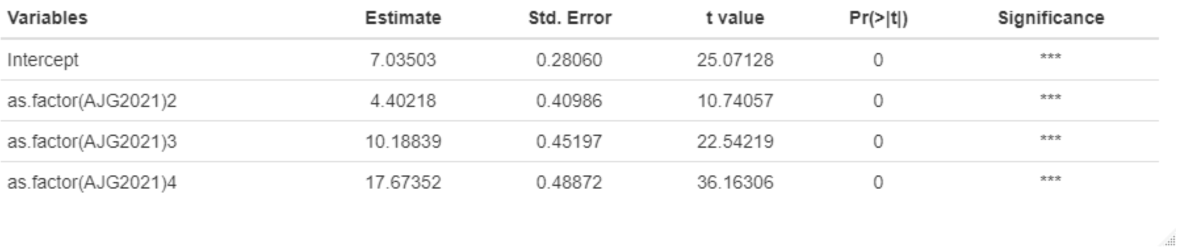
Table 8: Journal related regression results

We find *AJG2021* to be significant below 5%. Publishing in a higher ranked journal leads to a positive increase in total citations after five years with an expected increase of 5.7 more citations per ranking increase. The adjusted R-squared was 0.04118, which is almost the same as all paper related variables combined. This signals that our journal related variable is explaining close to the same amount of variance as our paper related variables.

To identify the second aspect, we regressed the *AJG2021* as a factor with dummy variables on each ranking level as follows:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{AJG2} + \beta_2 X_{AJG3} + \beta_3 X_{AJG4} + \varepsilon,$$

where β_0 is AJG1 and X_{AJGi} is binary.



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	7.03503	0.28060	25.07128	0	***
as.factor(AJG2021)2	4.40218	0.40986	10.74057	0	***
as.factor(AJG2021)3	10.18839	0.45197	22.54219	0	***
as.factor(AJG2021)4	17.67352	0.48872	36.16306	0	***

Table 9: Regression model for journal related variables as factor

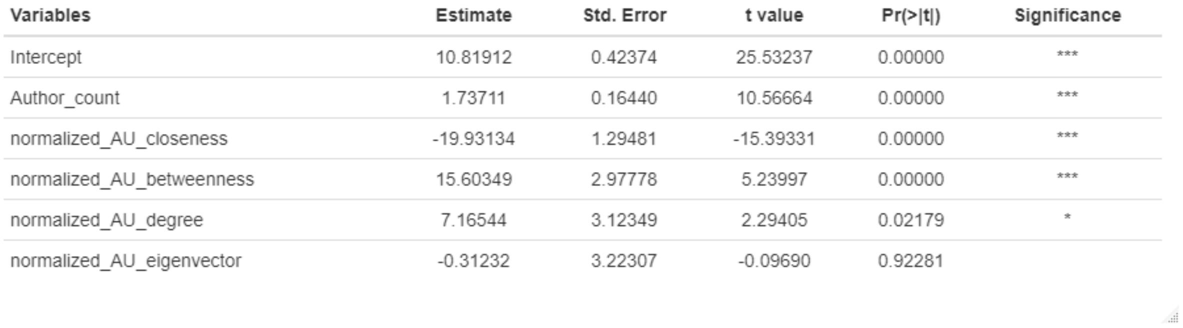
The results shows that publishing in an *AJG 2* ranked journal is expected to lead to an increase of 4.4 citations, while publishing in an *AJG 3* ranked journal gives an expected increase of 10.2 citations, compared to publishing in an *AJG 1* ranked journal. Publishing in an *AJG 4* ranked journal leads to the largest increase in expected citations, of 17.7. The adjusted R-squared in this regression increased from 0.04118 to 0.04173, showing that having *AJG2021* as a dummy variable slightly improved explanation of variance.

The regression results presented in Table 8 and Table 9 show that AJG ranking is a significant driver of article citations after five years. This also gives us the impression that it can be a qualified variable to use in the prediction of highly cited articles after five years to test *H3*.

Author related regression results

To investigate author related variables in isolation and directly test *H2*, we created a regression containing only author related variables. The number of authors works as a correction for the increased probability of having an author with higher centrality score. This regression can be formulated as the following:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{Author\_count} + \beta_2 X_{normalized\_\text{AU\_closeness}} + \beta_3 X_{normalized\_AU\_betweenness} + \beta_4 X_{normalized\_AU\_degree} + \beta_5 X_{normalized\_AU\_eigenvector} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	10.81912	0.42374	25.53237	0.00000	***
Author_count	1.73711	0.16440	10.56664	0.00000	***
normalized_AU_closeness	-19.93134	1.29481	-15.39331	0.00000	***
normalized_AU_betweenness	15.60349	2.97778	5.23997	0.00000	***
normalized_AU_degree	7.16544	3.12349	2.29405	0.02179	*
normalized_AU_eigenvector	-0.31232	3.22307	-0.09690	0.92281	

Table 10: Author related regression results

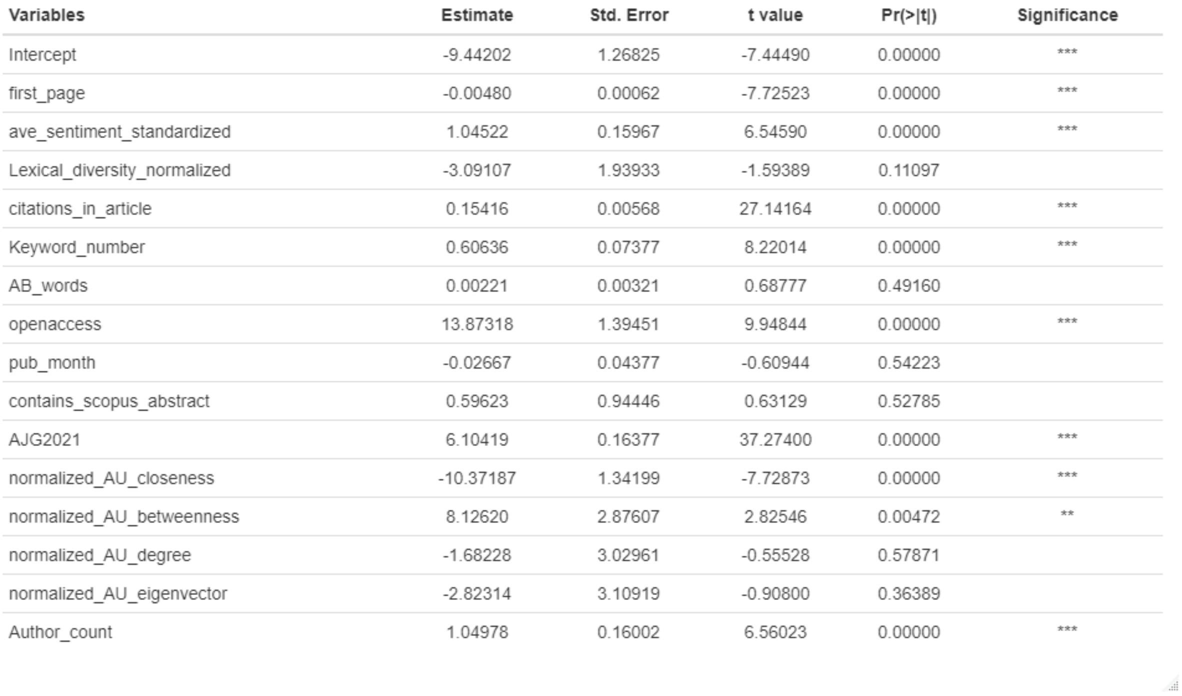
From the multiple regression of author related variables presented in Table 10, we find that all variables except *eigenvector centrality* are significant below a 5% level. The adjusted R-squared is 0.01745, meaning that these variables explain relatively less of the variation in citations after five years compared to the paper and journal variables. Higher values increase total citations after five years for *betweenness* and *degree centrality*, while lower values of *closeness centrality* increase the number of expected citations after five years. Importantly, we find the eigenvector centrality to have a p-value of 0.92, which indicates that it is highly insignificant. Since the centrality variables are normalized, we can compare their coefficient sizes. We see that *closeness* and *betweenness centrality* has the largest impact on citations, while *author count* has a very small effect, and *eigenvector centrality* has a non-significant effect on the total number of citations after five years.

Our second hypothesis (*H2*) was that the values from centrality measures are drivers of citations after five years. All the centrality measures except *eigenvector centrality* was significant in some degree. We therefore find support for the hypothesis in these results.

Paper, journal, and author related variables combined

To compare the effect of all variables without information about the impact the year after publication to the effect of all variables with this information, we created a multiple linear regression of all variables but *new citations after one year*. This regression can be expressed as:

$$total\_citations\_after\_5y = \beta_0 + \beta_1 X_{\mathrm{fi}rst\_page} + \beta_2 X_{ave\_sentiment\_standardized} + \beta_3 X_{\mathrm{L}exical\_diversity\_normalized} + \beta_4 X_{\mathrm{Keyword\_number}} + \beta_5 X_{\mathrm{cita}tions\_in\_article} + \beta_6 X_{AB\_words} + \beta_7 X_{\mathrm{openaccess}} + \beta_8 X_{pub\_month} + \beta_9 X_{contains\_scopus\_abstract} + \beta_{10} X_{AJG2021} + \beta_{11} X_{Author\_count} + \beta_{12} X_{\mathrm{normalized\_AU\_closeness}} + \beta_{13} X_{\mathrm{normalized}\_AU\_betweenness} + \beta_{14} X_{\mathrm{normalized}\_AU\_degree} + \beta_{15} X_{\mathrm{normalized}\_AU\_eigenvector} + \varepsilon$$



Variables	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	-9.44202	1.26825	-7.44490	0.00000	***
first_page	-0.00480	0.00062	-7.72523	0.00000	***
ave_sentiment_standardized	1.04522	0.15967	6.54590	0.00000	***
Lexical_diversity_normalized	-3.09107	1.93933	-1.59389	0.11097	
citations_in_article	0.15416	0.00568	27.14164	0.00000	***
Keyword_number	0.60636	0.07377	8.22014	0.00000	***
AB_words	0.00221	0.00321	0.68777	0.49160	
openaccess	13.87318	1.39451	9.94844	0.00000	***
pub_month	-0.02667	0.04377	-0.60944	0.54223	
contains_scopus_abstract	0.59623	0.94446	0.63129	0.52785	
AJG2021	6.10419	0.16377	37.27400	0.00000	***
normalized_AU_closeness	-10.37187	1.34199	-7.72873	0.00000	***
normalized_AU_betweenness	8.12620	2.87607	2.82546	0.00472	**
normalized_AU_degree	-1.68228	3.02961	-0.55528	0.57871	
normalized_AU_eigenvector	-2.82314	3.10919	-0.90800	0.36389	
Author_count	1.04978	0.16002	6.56023	0.00000	***

Table 11: Regression results for paper, journal, and author related variables combined

When regressing paper, journal, and author related variables we get an adjusted R-squared of 0.0875. Table 11 show that *lexical diversity, abstract length, publication month, Scopus abstract, degree centrality, eigenvector centrality*, and *author count* are *not* significant.

Compared to the full regression, removing the adjustment for citations one year after publication result in *betweenness centrality, AJG2021, number of keywords*, and *author count* becoming significant as well, while *lexical diversity* loses its statistical significance. We observe that both *closeness* and *betweenness centrality* is still significant, which supports *H2*.

The concluding remarks from the regression analysis is that *new citations after one year* is a dominant driver of article citations obtained after five years, supporting *H1*. We also find co-author centrality measures to be a significant driver of article citations in all regressions present. Hence, we find support for *H2.* The slightly higher R-squared value from the full regression indicates that other variables may help correct the outcome variable slightly and can therefore still be useful predictors. From the multiple linear regressions, we do not know their non-linear interactions or exactly how well the variables other than new citations after one year will translate into predictive performance. This will therefore be the topic in section 6.2.

Test of Regression Validity

To check the validity of the multiple regression assumptions, we visualized the residual plots versus the fitted regression. From Figure 3, we see no clear non-linear pattern of residual error, except when including *new citations after one year*, which approaches exponential error increase with fitted values above 500.

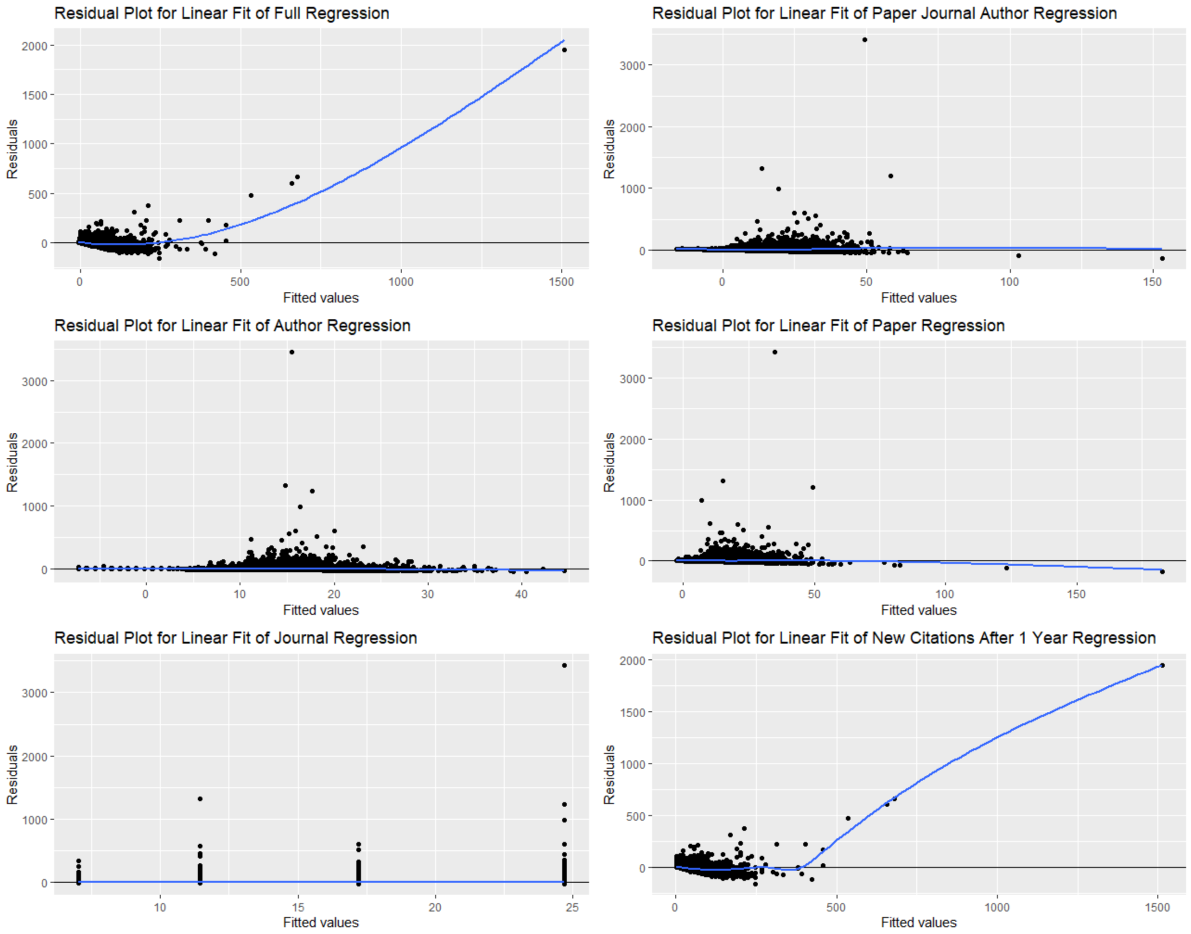


Figure 3: Linear Fit of the Residuals versus the Fitted Values of the Regression.

We do see indications of non-constant variance of error terms (Heteroscedasticity) for all but the author related regression in the form of an increase in error with the increase in fitted values. This is concerning for the interpretation of the regression results in these regressions. However, they do not change the conclusions to *H1* and *H2* as the author regression is homoscedastic, and the *new citations after one year* variable is extremely significant and highly unlikely to become much less significant even after correcting for the heteroscedasticity. However, it means that the standard-error and therefore the p-values from the paper, paper-journal-author, and journal regressions are dubious, and likely lower than true.

While we do see a few outliers and high leverage points, we do not consider these as data errors, as they represent valid examples of unusual citation counts after five years. However, it is worth noticing, as it indicates that the model does not capture all observations perfectly.

The issue of multicollinearity was checked though a correlation plot of the variables. From Figure 4 we see that *eigenvector centrality, betweenness centrality, degree centrality*, and *betweenness centrality* have a high degree of correlation. The variables with a high degree of correlation can therefore result in growth of the estimated standard deviation, causing a higher p-value for the correlated variables. However, since *H2* is considering the effect of any centrality measures rather than individual centrality measure effects, we do not consider the multicollinearity to be a critical issue.

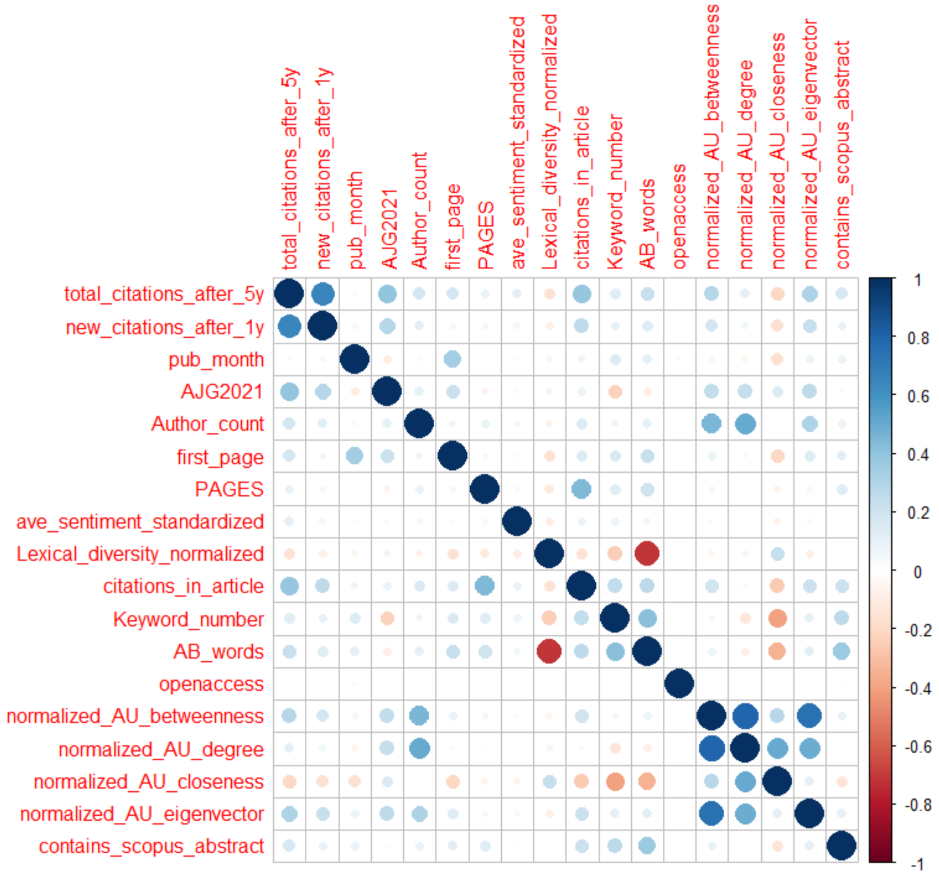


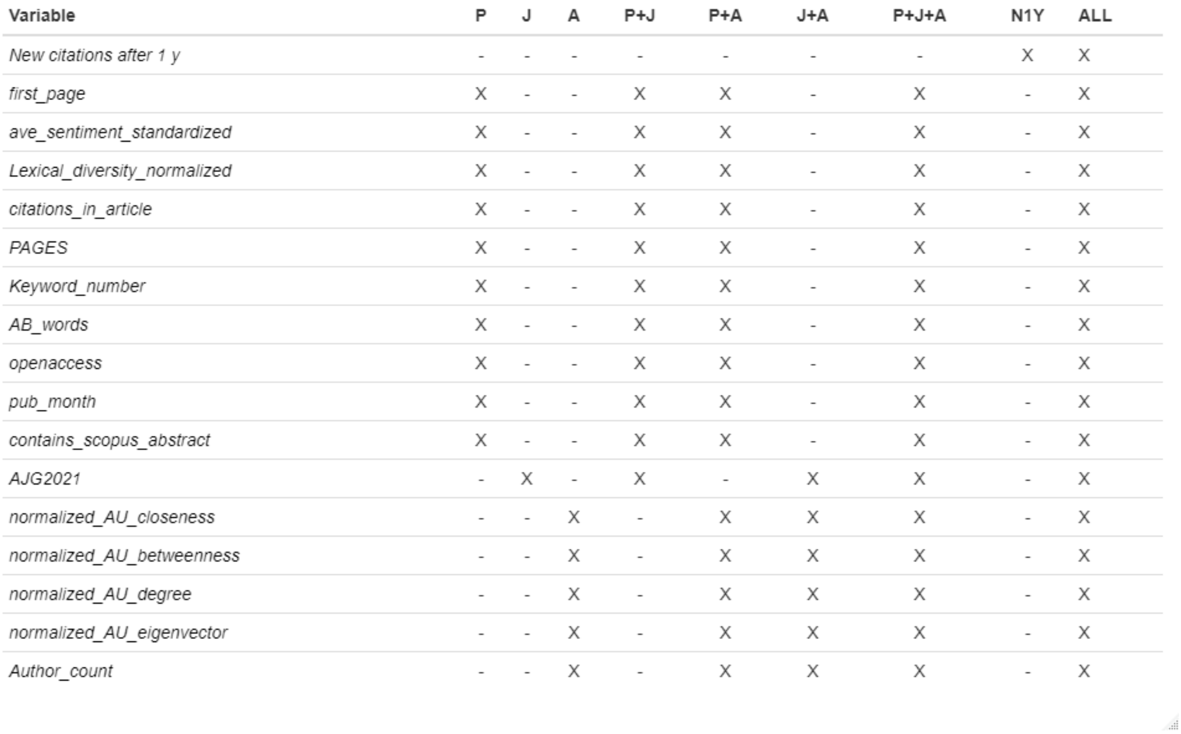
Figure 4: Correlation plot of all variables

6.2 Predictive results

To test our third hypothesis, which involves predicting whether a marketing article will be among top 5% most cited articles five years after publishing or not, we test each variable group and combination of the group on four different models. The first model is logistic regression, the second is SVM, followed by Light GBM, and finally we test TabNet. An interesting point about these models is that they use the variables in different ways and therefore likely perform differently on the same subsets. This will therefore strengthen the robustness of our results. In this section we will be using the training and test set presented in 5.5.2.

6.2.1 Variable Selection Strategy

Since some models take quite a lot of time to estimate, it is impossible to test all possible combinations of the variables as this would have resulted in 4*(2^{15}-1) = 262 143 models to train. As a result, we created a variable testing strategy to test our third hypothesis. With the variable testing strategy, we trained one model for each subset, where the variable testing subsets can be seen in Table 12. The model was then tested on an unseen dataset containing marketing articles from 2015 and 2016 as explained in 5.5.2. To ensure neutrality in our model performance, we consider precision and recall to be equally important. Hence, we calculated the probability threshold for the maximum F1 value for each trained model on the test set. This way, we could get a sense of how well the models could possibly perform in a use case scenario and ensure an equal model comparison on the metrics related to F1 (Accuracy, Recall, and Precision). The maximum F1 value was calculated by looping through all threshold values between 0.001 and 0.9 with a 0.001 step size. The F1, Accuracy, Precision, and Recall from the calculated optimal threshold value are reported in Table 13.



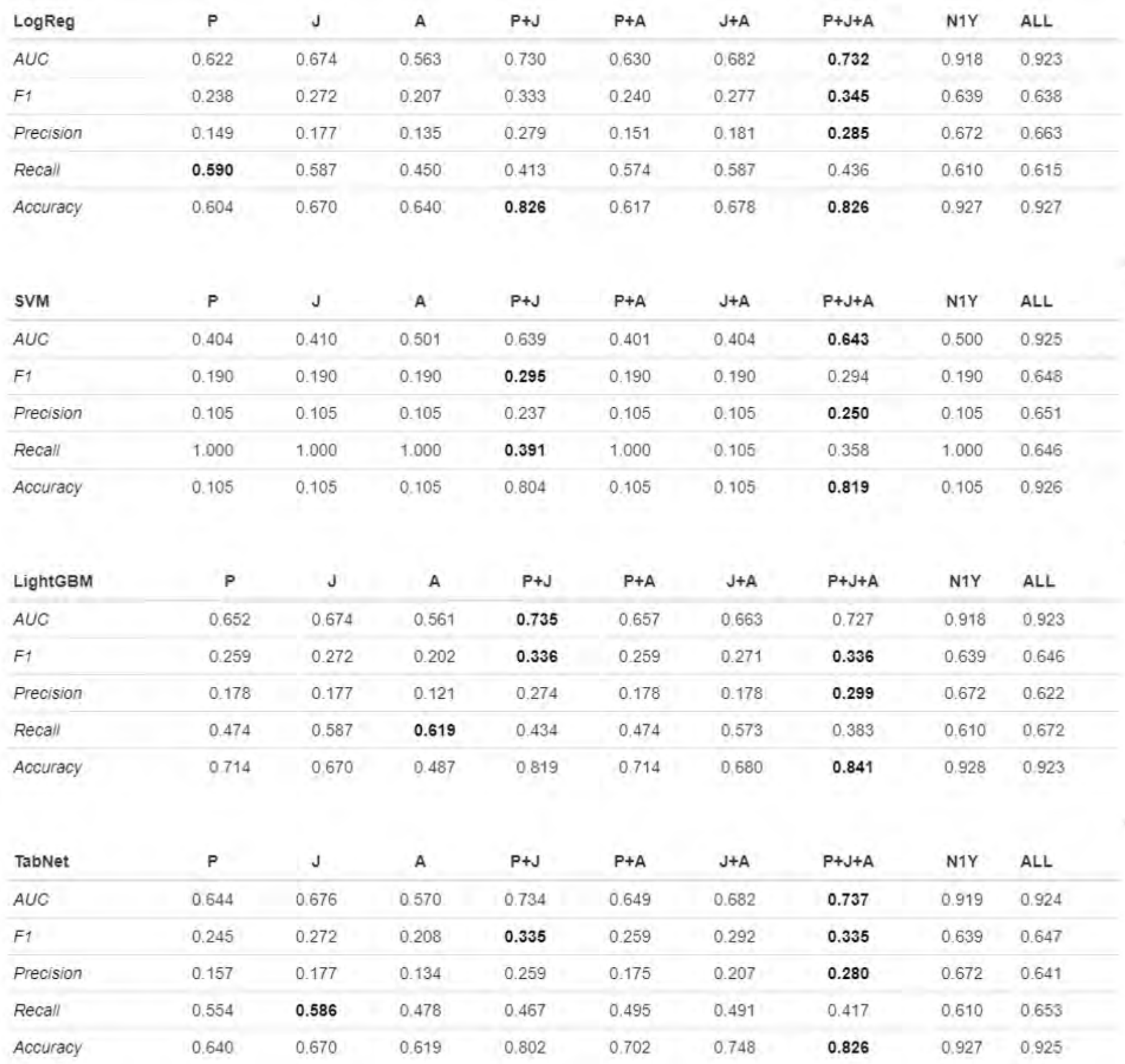
Variable	P	J	A	P+J	P+A	J+A	P+J+A	N1Y	ALL
New citations after 1 y	-	-	-	-	-	-	-	X	X
first_page	X	-	-	X	X	-	X	-	X
ave_sentiment_standardized	X	-	-	X	X	-	X	-	X
Lexical_diversity_normalized	X	-	-	X	X	-	X	-	X
citations_in_article	X	-	-	X	X	-	X	-	X
PAGES	X	-	-	X	X	-	X	-	X
Keyword_number	X	-	-	X	X	-	X	-	X
AB_words	X	-	-	X	X	-	X	-	X
openaccess	X	-	-	X	X	-	X	-	X
pub_month	X	-	-	X	X	-	X	-	X
contains_scopus_abstract	X	-	-	X	X	-	X	-	X
AJG2021	-	X	-	X	-	X	X	-	X
normalized_AU_closeness	-	-	X	-	X	X	X	-	X
normalized_AU_betweenness	-	-	X	-	X	X	X	-	X
normalized_AU_degree	-	-	X	-	X	X	X	-	X
normalized_AU_eigenvector	-	-	X	-	X	X	X	-	X
Author_count	-	-	X	-	X	X	X	-	X

Table 12: Variables contained in the different subset groups

Testing the different categories of variables is an approach which has previously been used to test how various variable categories affect predictive performance on different models (Hu et al., 2020). To be able to test *H3*, we made subsets with each category separated. We tested all possible combinations of the three categories, where they only had data available at the publication date to ensure an equal comparison. For performance comparison, we made a subset consisting of only *new citations after one year* and one subset with all variables including new citations after one year. This would allow us to see the additional effect of paper, journal, and author variables with new citations after one year.

6.2.2 Variable Testing Results

To predict the 5% most cited marketing articles, we applied our four selected methods and created nine different models for each of them. For each model we measured the ROC AUC, F1, Precision, Recall, and Accuracy metrics which we will compare and analyze. We will primarily compare ROC AUC, F1, Precision, and Accuracy, as Recall can be artificially high when the model is not working.



LogReg	P	J	A	P+J	P+A	J+A	P+J+A	N1Y	ALL
AUC	0.622	0.674	0.563	0.730	0.630	0.682	**0.732**	0.918	0.923
F1	0.238	0.272	0.207	0.333	0.240	0.277	**0.345**	0.639	0.638
Precision	0.149	0.177	0.135	0.279	0.151	0.181	**0.285**	0.672	0.663
Recall	**0.590**	0.587	0.450	0.413	0.574	0.587	0.436	0.610	0.615
Accuracy	0.604	0.670	0.640	**0.826**	0.617	0.678	**0.826**	0.927	0.927

SVM	P	J	A	P+J	P+A	J+A	P+J+A	N1Y	ALL
AUC	0.404	0.410	0.501	0.639	0.401	0.404	**0.643**	0.500	0.925
F1	0.190	0.190	0.190	**0.295**	0.190	0.190	0.294	0.190	0.648
Precision	0.105	0.105	0.105	0.237	0.105	0.105	**0.250**	0.105	0.651
Recall	1.000	1.000	1.000	**0.391**	1.000	0.105	0.358	1.000	0.646
Accuracy	0.105	0.105	0.105	0.804	0.105	0.105	**0.819**	0.105	0.926

LightGBM	P	J	A	P+J	P+A	J+A	P+J+A	N1Y	ALL
AUC	0.652	0.674	0.561	**0.735**	0.657	0.663	0.727	0.918	0.923
F1	0.259	0.272	0.202	**0.336**	0.259	0.271	**0.336**	0.639	0.646
Precision	0.178	0.177	0.121	0.274	0.178	0.178	**0.299**	0.672	0.622
Recall	0.474	0.587	**0.619**	0.434	0.474	0.573	0.383	0.610	0.672
Accuracy	0.714	0.670	0.487	0.819	0.714	0.680	**0.841**	0.928	0.923

TabNet	P	J	A	P+J	P+A	J+A	P+J+A	N1Y	ALL
AUC	0.644	0.676	0.570	0.734	0.649	0.682	**0.737**	0.919	0.924
F1	0.245	0.272	0.208	**0.335**	0.259	0.292	**0.335**	0.639	0.647
Precision	0.157	0.177	0.134	0.259	0.175	0.207	**0.280**	0.672	0.641
Recall	0.554	**0.586**	0.478	0.467	0.495	0.491	0.417	0.610	0.653
Accuracy	0.640	0.670	0.619	0.802	0.702	0.748	**0.826**	0.927	0.925

Table 13: Results from the variable group testing. Bold markings show the best performing model (of paper, journal, and author combinations) on metrics across the variable subsets.

From the results in Table 13, we can see that combining a group of drivers gives a considerably higher ROC AUC, F1, Precision, and Accuracy on all models. Combining paper, journal, and author related variables achieves a higher ROC AUC on all models except Light GBM. Furthermore, it achieves a higher F1 score on all models except SVM. The TabNet model had the highest ROC AUC value of 0.737. Light GBM had a F1 score of 0.336, a precision score of 0.299, and an accuracy of 0.841, which was the highest compared to the other paper, journal, and author related models. The result from the testing therefore supports our third hypothesis (*H3)*, which says that using a combination of either paper, author, and journal related drivers will give more accurate predictions compared with using variables from only one category.

In addition, we wanted to show that it is possible to utilize the drivers presented in the first part to predict highly cited articles. In the first part of the results, we found that *new citations after one year* is a strong driver of citations after five years. When including it in the predicting models we notice that it performs extremely well, both as the only variable, but even better in combination with all the other variables. The accuracy of the predictions further supports *H1.* Our results show that the drivers we found to be impactful in the first part are useful to predict the future highly cited articles. Again, it supports that a combination of variables is the best option for prediction of highly cited articles (*H3*).

7. Discussion

The discussion includes findings from previous literature and from our results, the interpretation made by the authors of this thesis, as well as its implications for marketing researchers.

7.1 Drivers of Citations in the Marketing Field

By investigating the drivers of citations in the marketing field, we found that the number of new citations received in the first year after the publication year was the largest driver of citations. This supports our first hypothesis and previous research findings from other fields publications (Ma et al., 2021; Ruan et al., 2019; Abrishami & Aliakbary, 2019; Abramo et al., 2019, Stegehuis et al., 2015). However, this variable differs from the others by being time dependent. One will have to wait one year after publication before being able to use the number of citations to predict citations after another four years. Hence, it will not be possible to use the variable on articles that has been published before this citation variable is available. This may possibly lead to important and current articles being ignored early after publication. Citation count after one year is therefore not a suitable variable to use when the goal is to find the most impactful recently published research in a field.

When running the full regression, we notice that the effect of new citations after one year is very high compared to the other drivers, which confirms our hypothesis (*H1)* that this variable has a high influence on article citation after five years in the marketing field. In addition, this variable has a considerably higher correlation to total citations after five years compared to the other variables. Why is it that the number of citations received the year after publication is a strong driver for total citations obtained four years later? One explanation can be found in the study field of citation trajectories of successful papers. Baumgartner & Leydesdorff (2013) found that citation trajectories of breakthrough papers as well as "excellent" papers were expected to show high levels of citations from the beginning. In other words, the papers with high impact were found to immediately be recognized as major contributions to the field. Their results were found on natural science papers, and our results now supports that this effect is true for marketing papers as well.

Looking at the author related drivers, we find the number of authors of an article, degree centrality, betweenness centrality, and closeness centrality to influence article citations. Both betweenness centrality and degree centrality have in previous research been found to be a significant driver (Colladon et al., 2020; Yan & Ding, 2009). However, when combining paper, journal, and author related drivers in a regression, degree centrality no longer has any significant effect on citations after five years. Closeness centrality, betweenness centrality, and author count are still showing a significant effect. From our results it appears that having a low closeness centrality value, and thus a short path to other authors, can be a driver of article impact. This is contradicting with previous research that found low or no significant influence on citation counts (Yan & Ding, 2009; Uddin et al., 2013). However, these studies were performed on publications from the field of steel structure, and library and information science, which may have different network effects compared to the marketing field. The research also found that betweenness centrality correlated with citation counts, which is supported in our results. The network of an author may therefore have some importance when it comes to the impact of their published articles. Both position among other authors, how much an author controls the information flow, and having more connectivity can lead to an article having more impact.

When including citations one year after publication, only closeness centrality was shown to have a significant effect on citations after five years. In a study by Biscaro and Giupponi (2014) they found that degree centrality and betweenness centrality had a positive effect on citations in the first two years after publication. This might therefore explain these centrality measures are not significant when we include a variable with citation count after one year. Since the new citation after one year variable already contain information about the initial impact, the non-significance when including the citations after one year variable suggests that degree and betweenness centrality does not have any additional significant impact on article citations beyond the year after publication.

Both the author related regression and the regression with a combination of paper, journal, and author related drivers supports the hypothesis that value of centrality measures are drivers of citations five year ahead (*H2*). Including only a few variables and comparing the different centrality measures, we see that three out of four measures have a significant effect in the author related regression. However, including other types of variables in a larger regression lessens the importance of the centrality measures. Even though only closeness centrality and betweenness centrality have a significant effect in the combination regression, we still think the results are solid enough to conclude that co-author centrality measures are drivers in the marketing field.

To answer the first two hypotheses, we looked at the drivers in separate categories and combined in a full regression. These results can be used to give an indication of whether the variables will be suitable for prediction purposes as well, and thus important before answering our third hypothesis. We found that several variables within each category were significant, and thus considered to be drivers of article citation. Paper related variables, as article placement in journal and the abstract sentiment, had a positive significant effect on article citations after five years, which is supporting the findings of previous research (Stremersch et al., 2007; Colladon et al., 2020). AJG ranking was the only journal related variable that we included. It has not been used much in research on citation counts. The variable did show to be a driver of article impact and could therefore be important when predicting highly cited articles after five years. Lastly, as author count, closeness centrality, and betweenness centrality had statistically significant effect on citations after five years, we assumed that these would also be important for prediction purposes.

7.2 Prediction of Highly Cited Marketing Articles

In the predictive part of our thesis we used Logistic Regression, SVM, LightGBM, and TabNet to predict highly cited marketing articles after five years. The threshold was set based on the 95th percentile of five-year citation counts from marketing articles published between 1992 and 2016. Hence, the target was to predict the 5% most cited articles. Our results showed that most of our models had the best performance measure when having a combination of paper, journal, and author related drivers. Some of the models had a few of the measures performing better when using a combination of paper and journal related drivers, but these were primarily on the recall measure, which is prone to increase when the model is performing poorly. Interestingly, Hu et al. (2020) investigated three top marketing journals and found that a combination of journal and author related drivers worked well. However, they did not test any combinations with paper related drivers. In our thesis, we have included substantially more journals. In addition, we included other variables which might be the reason we obtain a better performance using a different combination. Overall, as all our methods created models that performed substantially better when combining different drivers, this supports our hypothesis proposing that a combination of drivers from paper, journal, or author related drivers are better than using drivers from only one of the categories (*H3).*

We emphasize that the combination of all three categories and the paper-journal combination did have a substantially better result than the combinations journal-author and paper-author. The latter two combinations did not have a strong difference in performance compared to the models with variables from only one category. This means that not all combinations of variables from different categories can give a better result. Although this does not exclusively support *H3*, we still think that the great improvement in performance in two of the models is enough to strongly support our hypothesis. In addition, the results points to the importance of using the right combination of drivers. None of the models with only one category of variables came close to the results in the two combinations which gave the best results.

To utilize the impact of the drivers that we found in the first part of the thesis, we made two additional prediction models; one that used only citations after one year as a variable, and one that added citations after one year to the combined paper, journal, author model. The results showed that all our four methods are considerably better at predicting the top 5% most highly cited marketing articles after five years when including this variable. The results indicate that there is a clear advantage of waiting one year to include citations after one year. The most prominent practical benefit is the increase is in the precision of correct predictions of top 5% articles, which changes from less than 0.3 to over 0.6. Using only citations after one year as a variable works well for all models except SVM, but we argue that the poor performance from SVM is due to the method not being designed for only one variable (Nalepa & Kawulok, 2018). Including citations after one year in a full model give great performance using all methods. The results support *H1*, as the predictions get considerably better when including the variable. This is a clear indication that the citations after one year highly influence citations after five years.

Both the TabNet and the Logistic Regression models have best prediction results when using variables from both paper, journal, and author related drivers. Hence, we recommend using one of these models if the goal is to predict highly cited articles and using variables from all categories. Logistic regression is the only one of our four models that have been previously used for prediction of citations in the marketing field (Hu et al., 2020). Our findings indicate that it is a decent model to use to predict highly cited articles. TabNet is the newest model that we use in our tests, and it has not been tested on citation prediction before. However, it is made to perform well on tabular data, and have been presenting good results when predicting rainfall, among other things (Xu et al., 2020). We therefore conclude that TabNet is a suitable model for predicting highly cited articles in the marketing field.

7.3 Implications for Marketing Researchers

Our results can primarily impact two groups; the readers and the authors of marketing articles. For the readers of marketing papers, the predictive results provide two key takeaways. First, if the paper has a citation count from the first full year after publication, then simply looking at the number of new citations the article got the year after publication is highly effective for identifying articles which will be impactful five years after publication. We would argue that there are diminishing returns on using additional data to identify impactful articles if enough time has passed to use the new citation after one year variable. Second, if the article is too recent for this to be available, using a machine learning model can help identifying impactful articles after five years with using a combination of paper and journal related variables. Which model to use would depend on the types of variables available, and preferred area of performance. If the reader has a strong knowledge of marketing papers, then using a model with higher recall performance such as Logistic Regression or TabNet may be preferred as the reader would be more able to filter out the false positives from the suggestions. On the opposite side, if the reader is less known with marketing papers, using a model with higher precision such as LightGBM may be preferred, as the reader would be less required to filter out the false positive suggestions.

For the authors of marketing articles, we have four key takeaways with regards to impact on marketing academia. First, getting citations in the year after the publication year yields a significant return on later impact, where each additional citation is expected to result in eleven additional citations five years after publication. Second, having a short path to other marketing academics and being the connection between other authors, is beneficial for impacting the field of academic marketing research. Third, which journal the article is published in matters. Isolated from other variables, the higher AJG ranked the journal an article is published in, the higher impact is expected. We argue that this is especially important for receiving early citations. Our results shows that all increase in journal ranking is positive, however the biggest increase in expected citations is achieved by publishing in an AJG 4 ranked journal. One final takeaway is that making the article open-access is expected to lead to more impact, as the results found that it had a positive effect on citations after five years. Having open access to an article makes it easier for people to access it. Most people will not have access to every journal, meaning they will miss out on some potentially important articles. As a result, publishing an article with open access can lead to more citations.

8. Conclusion

This thesis has studied the important drivers of article impact in the marketing field and tested the ability of the drivers to predict highly cited articles after five years. To investigate what impacts article citation we have used data from 75 different marketing journals across a time-period of 30 years. Specifically, the aim of the thesis was to find relevant drivers of impact in articles published within marketing journals. Further, we wanted to use these drivers to predict highly cited articles five years after publication.

To do this, we have done an extensive literature review on the drivers of citations and prediction of article citations. Based on the literature review, we found gaps in the research which were the foundation for our hypotheses. The experiments on our three hypotheses were performed using both traditional and innovative machine learning methods to confirm or reject our hypotheses. To the best of our knowledge, this is the first research of its kind on articles from all English-written journals in the ABS journal ranking, as previous research only has focused on articles from a few top marketing journals. This research will therefore be relevant to a broader research audience within marketing academia.

Our research found several statistically significant drivers of impact on marketing articles. New citations one year after publication was the strongest driver of impact in academic articles within the marketing field. For paper related drivers, placement of article in journal, sentiment of abstract, citations in article, number of keywords, and open access had a statistically significant impact on citations. Furthermore, we found that the journal related variable that takes journal ranking based on expert assessment (AJG2021) is a significant driver of article citations. Regarding our author related variables, we find author count, closeness centrality and betweenness centrality to be significant drivers.

Our research found promising results regarding the predictive ability of our drivers. We found that combining paper, author, and journal related drivers resulted in promising predictive results. After comparing subsets with different groups of variables, we found only a marginal contribution by adding author variables to a model with paper and journal variables. However, combining variables from the different categories still performed considerably better compared to using variables from only one category. Furthermore, having one year to gather information leads to a significantly better model. In fact, using only citations after one year on any model other than SVM leads to a better model than only using paper, journal, or author related drivers combined. Combining all the variables gives an even better performance. The performance from the best model makes it possible for readers who want to stay updated on the marketing field to find highly impactful articles to read, without guessing among the thousands of articles published every year.

This thesis has used both acknowledged methods and methods that have never been tested on citation count predictions before. Although Light GBM and TabNet have not been tested on prediction of article citation previously, they show that they can perform better than other traditional machine learning methods if data about citation counts after one year is not yet available. The results are a contribution to the research of articles in the marketing field, as we have analyzed a large dataset with a broad range of marketing journals. Many of our variables has not been tested as drivers of impact in the marketing field earlier, but our results support the use of found drivers for future prediction purposes. Furthermore, this research has tested methods not previously used in prediction of highly cited articles with promising results.

8.1 Limitations and Future Research

This thesis has limitations that should be considered in future work. The primary limitation is that we have only considered documents of the type "article". Due to computational restrictions, we therefore excluded conference papers and other forms which still may contain new research. Our results will therefore not fully apply to conference papers, reviews, books and book chapters, editorials, letters, and other document types as the type of document has been shown to be affecting citations (Fu and Aliferis, 2010; Stegehuis et al., 2015). In addition, we only used one data source, Scopus, for getting article data. This means that our results primarily apply to Scopus data. If Scopus is missing data or have faulty data about an article, our data will be less accurate than if we cross checked the data from other sources. Future research can therefore cross check article data with other sources to ensure even better data accuracy and test the hypotheses on other document types.

Another limitation of our thesis is that we have only tested one journal metric as a journal related variable. This means there are journal related drivers and aspects which we have not considered, which may further improve the predictive ability of journal related variables in *H3*. Hence, in future research it would be useful to compare the predictive power of AJG2021 with other journal metrics such as Journal Citation Reports (JCR), SCImago Journal Rank

(SJR), the Source Normalized Impact per Paper (SNIP), Citescore, as well as combinations of these.

Another limitation is that the observations from early 1990s may not be as correct as the observations from after 2000. This may be due to the co-authorship network being calculated from 1992. For authors publishing in the 90s, we exclude their potential collaborations prior to 1992. The centrality measures might therefore not be representative for authors that had their publishing peak before 1992. Future research should therefore correct for this.

Lastly, given the importance of early citations in marketing articles, future research should study the drivers of early citations in marketing to improve predictions right after publication. Our results, in addition to previous research, points to journal and author related variables being more important for citations early after publication. However, other variables may be even more important, and could be an interesting subject to further explore.

9. References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics*, *5*(4), 659–667. https://doi.org/10.1016/j.joi.2011.06.004

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics*, *13*(1), 32–49. https://doi.org/10.1016/j.joi.2018.11.003

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*(2), 485–499. https://doi.org/10.1016/j.joi.2019.02.011

Academic Journal Guide 2021. (n.d.). Chartered Association of Business Schools. Retrieved February 4, 2022, from https://charteredabs.org/academic-journal-guide-2021-view/

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*, *67*(4), 1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Adankon, M. M., & Cheriet, M. (2015). Support Vector Machine. *Encyclopedia of Biometrics*, 1504–1511. https://doi.org/10.1007/978-1-4899-7488-4_299

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, *15*(2), 101128. https://doi.org/10.1016/j.joi.2020.101128

Akoglu, H. (2018). User's Guide to Correlation Coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*, *9*(1), 215824401982957. https://doi.org/10.1177/2158244019829575

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, *38*(6), 895–905. https://doi.org/10.1016/j.respol.2009.02.001

Andriy Burkov. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Aria, M., & Cuccurullo, C. (2022). *Package "bibliometrix."* CRAN. https://cran.r-project.org/web/packages/bibliometrix/bibliometrix.pdf

Ariel Xu, Q., & Chang, V. (2020). Co-authorship network and the correlation with academic performance. *Internet of Things*, *12*, 100307. https://doi.org/10.1016/j.iot.2020.100307

Arik, S. O., & Pfister, T. (2020). TabNet: Attentive Interpretable Tabular Learning.

ArXiv:1908.07442 [Cs, Stat]. https://arxiv.org/abs/1908.07442

Baumgartner, S. E., & Leydesdorff, L. (2013). Group-based trajectory modeling (GBTM) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims." *Journal of the Association for Information Science and Technology*, *65*(4), 797–811. https://doi.org/10.1002/asi.23009

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-020-09896-5

Bhandari, M., Busse, J., Devereaux, P. J., Montori, V. M., Swiontkowski, M., Tornetta Iii, P., Einhorn, T. A., Khera, V., & Schemitsch, E. H. (2007). Factors associated with citation rates in the orthopedic literature. *Canadian Journal of Surgery. Journal Canadien de Chirurgie*, *50*(2), 119–123. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2384258/

Billah, S. M., & Gauch, S. (2015). *Social network analysis for predicting emerging researchers*. *01*, 27–35.

Biscaro, C., & Giupponi, C. (2014). Co-Authorship and Bibliographic Coupling Network Effects on Citations. *PLoS ONE*, *9*(6), e99502. https://doi.org/10.1371/journal.pone.0099502

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Blei, D., Edu, B., Ng, A., Jordan, M., & Edu, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Borgatti, S. P., & Halgin, D. S. (2011). On Network Theory. *Organization Science*, *22*(5), 1168–1181. https://doi.org/10.1287/orsc.1100.0641

Borghini, E., & Giannetti, C. (2021). Short Term Load Forecasting Using TabNet: A Comparative Study with Traditional State-of-the-Art Regression Models. *The 7th International Conference on Time Series and Forecasting*. https://doi.org/10.3390/engproc2021005006

Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics*, *7*(4), 933–944. https://doi.org/10.1016/j.joi.2013.09.003

Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, *8*(1), 175–180. https://doi.org/10.1016/j.joi.2013.11.005

Bornmann, L., & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, *7*(2), 562–574. https://doi.org/10.1016/j.joi.2013.02.005

Boughorbel, S., & Kadri, A. (2021). *FAIRNESS IN TABNET MODEL BY DISENTANGLED REPRESENTATION FOR THE PREDICTION OF HOSPITAL NO-SHOW A PREPRINT*. https://arxiv.org/pdf/2103.04048.pdf

Brass, D. J. (2002). Social networks in organizations: Antecedents and consequences. In *Unpublished manuscript, University of Kentucky, Lexington*.

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, *11*(1), 13–41. https://doi.org/10.1111/j.1467-839x.2007.00241.x

Callaham, M. (2002). Journal Prestige, Publication Bias, and Other Characteristics Associated With Citation of Published Studies in Peer-Reviewed Journals. *JAMA*, *287*(21), 2847. https://doi.org/10.1001/jama.287.21.2847

Cao, X., Chen, Y., & Ray Liu, K. J. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, *10*(2), 471–484. https://doi.org/10.1016/j.joi.2016.02.006

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. *IEEE/ACM Joint Conference on Digital Libraries*. https://doi.org/10.1109/jcdl.2014.6970190

Chartered Association of Business Schools. (2021, July). *Academic Journal Guide Methodology*. Charteredabs.org. https://charteredabs.org/wp-content/uploads/2021/06/Academic_Journal_Guide_2021-Methodology.pdf

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386–1409. https://doi.org/10.1002/asi.21309

Cimenler, O., Reeves, K. A., & Skvoretz, J. (2014). A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. *Journal of Informetrics*, *8*(3), 667–682. https://doi.org/10.1016/j.joi.2014.06.004

Colladon, A. F., D’Angelo, C. A., & Gloor, P. A. (2020). Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics*, *124*(1), 357–377. https://doi.org/10.1007/s11192-020-03479-5

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, *25*(4), 283–307. https://doi.org/10.1016/s0378-8733(03)00012-1

Das, K., Samanta, S., & Pal, M. (2018). Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, *8*(1). https://doi.org/10.1007/s13278-018-0493-2

De Queiroz, G., & Fay, C. (2021). *Package “tidytext.”* CRAN. https://cran.r-project.org/web/packages/tidytext/tidytext.pdf

Dervos, D., & Kalkanis, T. (2005). *Automatic Ranking of Research Publications* (pp. 668–673).http://eureka.teithe.gr/jspui/bitstream/123456789/10383/1/Dervos_Kalkanis_cc-IFF.pdf

Diallo, S. Y., Lynch, C. J., Gore, R., & Padilla, J. J. (2016). Identifying key papers within a journal via network centrality measures. *Scientometrics*, *107*(3), 1005–1020. https://doi.org/10.1007/s11192-016-1891-8

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064. https://doi.org/10.1002/asi.22806

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Doshi, K. (2021, June 3). *Transformers Explained Visually (Part 1): Overview of Functionality*. Medium. https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452

Features — LightGBM 3.3.2.99 documentation. (2022). Readthedocs.io. https://lightgbm.readthedocs.io/en/latest/Features.html#optimization-in-speed-and-memory-usage

Fonseca, B. de P. F. e, Sampaio, R. B., Fonseca, M. V. de A., & Zicker, F. (2016). Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems*, *14*(1). https://doi.org/10.1186/s12961-016-0104-5

Freeman, L. (1978). /79) 215-239 @Elsevier Sequoia S.A., Lausanne -Printed in the Netherlands. In *Social Networks* (p. 1).

https://www.bebr.ufl.edu/sites/default/files/Centrality%20in%20Social%20Networks.pdf

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270. https://doi.org/10.1007/s11192-010-0160-5

Galli, C., & Guizzardi, S. (2021). The Effect of Article Characteristics on Citation Number in a Diachronic Dataset of the Biomedical Literature on Chronic Inflammation: An Analysis by Ensemble Machines. *Publications*, *9*(2), 15. https://doi.org/10.3390/publications9020015

Garner, J., Porter, A. L., & Newman, N. C. (2014). Distance and velocity measures: using citations to determine breadth and speed of research impact. *Scientometrics*, *100*(3), 687–703. https://doi.org/10.1007/s11192-014-1316-5

Giuliani, E., & Pietrobelli, C. (2011). *Social Network Analysis Methodologies for the Evaluation of Cluster Development Programs*. https://arpi.unipi.it/retrieve/handle/11568/786587/90936/IDB-TN-317%20Social%20Network%20Analysis%20Methodologies%20for%20the%20Evaluation%20of%20Cluster%20Development%20Programs.pdf

Golbeck, J. (2013). Network Structure and Measures. *Analyzing the Social Web*, 25–44. https://doi.org/10.1016/b978-0-12-405531-5.00003-1

Goldenberg, J., Libai, B., & Muller, E. (2006). *AN EGO-CENTERED ANALYSIS OF LARGE- SCALE NETWORKS: THE CASE OF THE MARKETING DISCIPLINE by*. https://tad.colman.ac.il/paper-all/128491.pdf

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., & Murray, S. S. (2006). Effect of E-printing on Citation Rates in Astronomy and Physics. *The Journal of Electronic Publishing*, *9*(2). https://doi.org/10.3998/3336451.0009.202

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hu, Y.-H., Tai, C.-T., Liu, K. E., & Cai, C.-F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. *Journal of Informetrics*, *14*(1), 101004. https://doi.org/10.1016/j.joi.2019.101004

Ibanez, A., Larranaga, P., & Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics*, *25*(24), 3303–3309. https://doi.org/10.1093/bioinformatics/btp585

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H., & Rehman, M. U. (2019). A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting. *IEEE Access*, *7*, 28309–28318. https://doi.org/10.1109/access.2019.2901920

Katz, J. Sylvan., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/s0048-7333(96)00917-1

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Klimek, P., S. Jovanovic, A., Egloff, R., & Schneider, R. (2016). Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks. *Scientometrics*, *107*(3), 1265–1282. https://doi.org/10.1007/s11192-016-1926-1

Lee, J. D., Simchowitz, M., Jordan, M. I., & Recht, B. (2016). Gradient descent only converges to minimizers. *Journal of Machine Learning Research*, *49*. https://collaborate.princeton.edu/en/publications/gradient-descent-only-converges-to-minimizers

LEIMU, R., & KORICHEVA, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32. https://doi.org/10.1016/j.tree.2004.10.010

Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing & Management*, *47*(2), 300–308. https://doi.org/10.1016/j.ipm.2010.09.005

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, *42*(9), 1515–1530. https://doi.org/10.1016/j.respol.2013.06.012

LightGBM: A Light Gradient Boosting Machine. (2018). TechLeer. https://www.techleer.com/articles/489-lightgbm-a-light-gradient-boosting-machine/

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, *126*. https://doi.org/10.1007/s11192-021-04033-7

Marin, A., & Wellman, B. (2010). *Social network analysis: An introduction 1*.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Matveeva, Nataliya; Poldin, Oleg. (2016). Citation of scholars in co-authorship network: Analysis of Google Scholar data. *Applied Econometrics*, *44*, 100–118. https://ideas.repec.org/a/ris/apltrx/0306.html

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19. https://doi.org/10.1016/j.ejor.2015.04.002

Mongeon, P., & Paul-Hus, A. (2015). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Muschelli, J. (2019). *Package "rscopus."* CRAN. https://cran.r-project.org/web/packages/rscopus/rscopus.pdf

Mutoh, A., Imura, Y., Kato, R., Matsui, T., & Inuzuka, N. (2016). A model of friendship networks based on social network analysis. *Artificial Life and Robotics*, *21*(2), 165–170. https://doi.org/10.1007/s10015-016-0275-8

Nalepa, J., & Kawulok, M. (2018). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*, *52*(2), 857–900. https://doi.org/10.1007/s10462-017-9611-1

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Newman, M. E. J. (2014). Prediction of highly cited papers. *EPL (Europhysics Letters)*, *105*(2), 28002. https://doi.org/10.1209/0295-5075/105/28002

Nightingale, J. M., & Marshall, G. (2012). Citation analysis as a measure of article quality, journal influence and individual researcher performance. *Radiography*, *18*(2), 60–67. https://doi.org/10.1016/j.radi.2011.10.044

OPPENHEIM, C. (1995). THE CORRELATION BETWEEN CITATION COUNTS AND THE 1992 RESEARCH ASSESSMENT EXERCISE RATINGS FOR BRITISH LIBRARY AND INFORMATION SCIENCE UNIVERSITY DEPARTMENTS. *Journal of Documentation*, *51*(1), 18–27. https://doi.org/10.1108/eb026940

Parsons, C. (2021, August 16). *What's a Machine Learning Model?* NVIDIA Blog. https://blogs.nvidia.com/blog/2021/08/16/what-is-a-machine-learning-model/

Paruchuri, S. (2010). Intraorganizational Networks, Interorganizational Networks, and the Impact of Central Inventors: A Longitudinal Study of Pharmaceutical Firms. *Organization Science*, *21*(1), 63–80. https://doi.org/10.1287/orsc.1080.0414

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *ArXiv:1503.06462 [Cs]*. https://arxiv.org/abs/1503.06462

Puuska, H.-M., Muhonen, R., & Leino, Y. (2013). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics*, *98*(2), 823–839. https://doi.org/10.1007/s11192-013-1181-7

Quan-Haase, A., & Wellman, B. (2006). *Hyperconnected Net Work, in Charles Heckscher and Paul Adler (eds.) The Firm as a Collaborative Community*.

Rinker, T. (2021). *Package "sentimentr."* CRAN. https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

Rosenzweig, S., Grinstein, A., & Ofek, E. (2016). Social network utilization and the impact of academic research in marketing. *International Journal of Research in Marketing*, *33*(4), 818–839. https://doi.org/10.1016/j.ijresmar.2016.02.002

Rostami, F., Mohammadpoorasl, A., & Hajizadeh, M. (2013). The effect of characteristics of title on citation rates of articles. *Scientometrics*, *98*(3), 2007–2010. https://doi.org/10.1007/s11192-013-1118-1

Royle, P., Kandala, N.-B., Barnard, K., & Waugh, N. (2013). Bibliometrics of systematic reviews: analysis of citation rates and journal impact factors. *Systematic Reviews*, *2*(1). https://doi.org/10.1186/2046-4053-2-74

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, *14*(3), 101039. https://doi.org/10.1016/j.joi.2020.101039

Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, *3*, 1–16.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/bf00116037

So, M., Kim, J., Choi, S., & Park, H. W. (2014). Factors affecting citation networks in science and technology: focused on non-quality factors. *Quality & Quantity*, *49*(4), 1513–1530. https://doi.org/10.1007/s11135-014-0110-z

Sohrabi, B., & Iraj, H. (2016). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, *110*(1), 243–251. https://doi.org/10.1007/s11192-016-2161-5

Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, *9*(3), 642–657. https://doi.org/10.1016/j.joi.2015.06.005

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The Quest for Citations: Drivers of Article Impact. *Journal of Marketing*, *71*(3), 171–193. https://doi.org/10.1509/jmkg.71.3.171

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, *32*, 101084. https://doi.org/10.1016/j.frl.2018.12.032

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, *15*. https://doi.org/10.1186/s12880-015-0068-x

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2

Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network Effects on Scientific Collaborations. *PLoS ONE*, *8*(2), e57546. https://doi.org/10.1371/journal.pone.0057546

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/b:scie.0000018543.82441.f1

van Wesel, M., Wyatt, S., & ten Haaf, J. (2013). What a difference a colon makes: how superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601–1615. https://doi.org/10.1007/s11192-013-1154-x

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need.* https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-

Paper.pdf

Wang, M., Wang, Z., & Chen, G. (2019). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics*, *119*(3), 1575–1595. https://doi.org/10.1007/s11192-019-03052-9

Warren, N. L., Farmer, M., Gu, T., & Warren, C. (2021). EXPRESS: Marketing Ideas: How to Write Research Articles that Readers Understand and Cite. *Journal of Marketing*, 002224292110035. https://doi.org/10.1177/00222429211003560

Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. https://doi.org/10.1017/cbo9780511815478

Web of Science Core Collection - Web of Science Group. (2019). Web of Science Group. https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/

Webster, C. M., & Morrison, P. D. (2004). Network Analysis in Marketing. *Australasian Marketing Journal*, *12*(2), 8–18. https://doi.org/10.1016/s1441-3582(04)70094-4

Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

White, H. D., Wellman, B., & Nazer, N. (2003). Does citation reflect social structure?: Longitudinal evidence from the ?Globenet? interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, *55*(2), 111–126. https://doi.org/10.1002/asi.10369

Willis, D. L., Bahler, C. D., Neuberger, M. M., & Dahm, P. (2011). Predictors of citations in the urological literature. *BJU International*, *107*(12), 1876–1880. https://doi.org/10.1111/j.1464-410x.2010.10028.x

Xu, J., Li, M., Jiang, J., Ge, B., & Cai, M. (2019). Early Prediction of Scientific Impact Based on Multi-Bibliographic Features and Convolutional Neural Network. *IEEE Access*, *7*, 92248–92258. https://doi.org/10.1109/access.2019.2927011

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107–2118. https://doi.org/10.1002/asi.21128

Yan, J., Xu, T., Yu, Y., & Xu, H. (2021). Rainfall Forecast Model Based on the TabNet Model. *Water*, *13*(9), 1272. https://doi.org/10.3390/w13091272

Yan, R., Huang, C., Tang, J., & Li, X. (2012, June 10). *To better stand on the shoulder of giants*. ResearchGate; unknown. https://www.researchgate.net/publication/254006037_To_better_stand_on_the_shoulder_of_giants

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*. https://doi.org/10.1145/2063576.2063757

Yao, Q., Li, R. Y. M., Song, L., & Crabbe, M. J. C. (2021). Construction safety knowledge sharing on Twitter: A social network analysis. *Safety Science*, *143*, 105411. https://doi.org/10.1016/j.ssci.2021.105411

Ye, F. Y., & Bornmann, L. (2017). "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology*, *69*(3), 359–367. https://doi.org/10.1002/asi.23846

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252. https://doi.org/10.1007/s11192-014-1279-6

Yue, W., & Wilson, C. S. (2004). Measuring the citation impact of research journals in clinical neurology: A structural equation modelling analysis. *Scientometrics*, *60*(3), 317–332. https://doi.org/10.1023/b:scie.0000034377.93437.18

Zhang, J., Mucs, D., Norinder, U., & Svensson, F. (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. *Journal of Chemical Information and Modeling*, *59*(10), 4150–4158. https://doi.org/10.1021/acs.jcim.9b00633

Zhang, X., Xie, Q., & Song, M. (2021). Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network. *Journal of Informetrics*, *15*(2), 101140. https://doi.org/10.1016/j.joi.2021.101140

Appendix

A1 R Libraries

Tidyverse - https://www.tidyverse.org/

Tidytext - https://cran.r-project.org/web/packages/tidytext/index.html

Sentimentr - https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

Bibliometrix - https://cran.r-project.org/web/packages/bibliometrix/index.html

Rscopus - https://cran.r-project.org/web/packages/rscopus/index.html

Data.table - https://cran.r-project.org/web/packages/data.table/data.table.pdf

Skimr - https://cran.r-project.org/web/packages/skimr/index.html

Corrplot - https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

A2 Python Libraries

Pytorch – https://pytorch.org/

ScikitLearn – https://scikit-learn.org/stable/

Optuna - https://optuna.org/

Pandas - https://pandas.pydata.org/

Numpy - https://numpy.org/

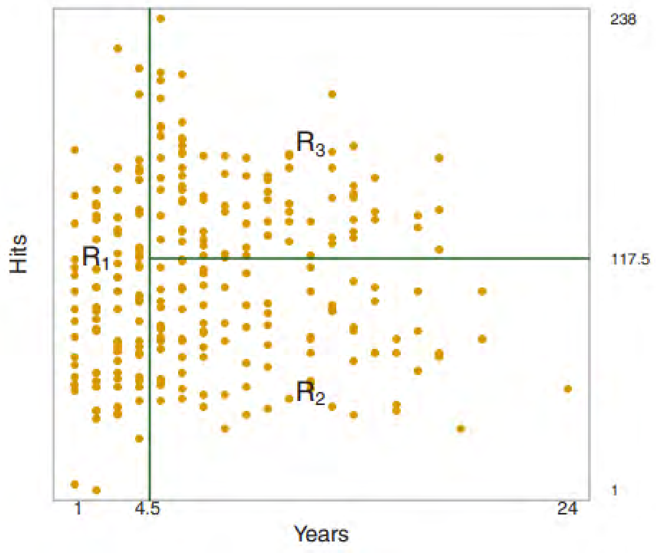
A3 Overview of Drivers of Citations

Table A1: Overview of our literature review. A plus sign signals that it is a driver in the specified study and field, while a minus sign signals that it is not proven to be a driver in the listed study.

Group	Driver	Study	Field
Paper	+ Early-stage citation count	Ma et al. (2021), Ruan et al. (2019), Abrishami & Aliakbary (2019), Abramo et al. (2019), Stegehuis et al. (2015)	Science and Technology, Economics, Library-, Law-, Political-, Social-, Science
	+ Abstract sentiment + Lexical diversity	Colladon et al. (2020)	Chemical Engineering
	- Keyword popularity	Hu et al. (2020)	Marketing
	+ Open access + Number of references + Reference variety + Reference prestige + Editorial Journal Awards + Review-Times + Peer-Reviewed	Tahamtan et al. (2016)	General
	+ Language Complexity	Warren et al. (2021)	Marketing
	Article length	- Royle et al. (2013), + Bornmann et al. (2014)	Medicine, General
	+ Article order - Title length	Stremersch et al. (2007)	Marketing
	+ Abstract length	Van Wesel et al. (2013)	Applied Physics
	+ Novel co-cited reference connections	Chen et al. (2010)	Information-Science
	+ Number of keywords	Chakraborty et al. (2014), So et al. (2014), Rostami et al. (2013)	Science and Technology, Computer Science
Journal	Journal Impact Factor	+ Hu et al. (2020), + Tahamatan et al. (2016), + Garner et al. (2014), - Willis et al. (2011), - Leimu & Koricheva (2005)	Marketing, General, Social-, Natural science, Urology, Ecology
	+ Journal Impact	Callaham (2002)	Medicine
	+ Journal Ranking	Sohrabi & Iraj (2016)	Education
	+ Journal Prestige	Garner et al. (2014)	HSD
	+ Journal Topic Broadness	Bornmann & Williams (2013)	Medical-, Natural-, Engineering Science
Author	+ Previous number of citations + Author productivity	Tahamtan et al. (2016)	General
	+ Number of Authors	Tahamtan et al. (2016), Stremersch et al. (2007)	General, Marketing
	+ Author social network	Chakraborty et al. (2014)	Science and Technology
	+ Number of organizations represented	Puuska et al. (2013)	General (Finland)
	+ Co-Author closeness + Co-Author betweenness + Co-Author degree	Colladon et al. (2020)	Chemical Engineering

A4 Details about LightGBM

Figure A1: Illustration of splitting a feature space. In this case the feature space is two-dimensional, and is made up of two variables, years, and hits. The green line is the splits creating the sections which will get the same predictions. Taken from James et al. (2013).



Ke et al. (2017) explains how the LightGBM model is created. Primarily it uses decision trees to learn a function from the input space X^s to the gradient space G. Imagine a training set with n instances $\{x_1, \dots, x_n\}$, where each x_i is a vector with dimension s in space X^s. The negative gradients of the loss function are denoted as $\{g_1, \dots, g_n\}$ in each iteration. The information gain is measured by the variance after splitting, as defined below.

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\left(\sum_{\{x_i \in O : x_{ij} < d\}} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{\{x_i \in O : x_{ij} > d\}} g_i\right)^2}{n^j_{r|O}(d)}\right),$$

where $n_O = \sum I[x_i \in O], n^j_{l|O}(d) = \sum I[x_i \in O : x_{ij} < d]$ *and* $n^j_{r|O}(d) = \sum I[x_i \in O : x_{ij} > d]$.

O is the training dataset on a fixed node of the decision tree. For feature j, the decision tree algorithm selects $d_j^* = argmax_d V_j(d)$ and calculated the largest $V_j(d_j^*)$. The data is then split according feature j^* at point d_{j^*} into the left and right child nodes.

In the GOSS method, the training instances are ranked according to their absolute values of their gradients in a descending order. Then, the top $a \times 100\%$ instances with the larger gradients are kept in an instance subset A. For the remaining set A^c consisting of $(1 - a) \times 100\%$ instances with smaller gradients, a subset B is randomly sampled with size

$b \times |A^c|$. Lastly, instances are split according to the estimated variance gain $\tilde{V}_j(d)$ over the subset $A \cup B$, as shown below:

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right),$$

Where $A_l = \{x_i \in A: x_{ij} \leq d\}, A_r = \{x_i \in A: x_{ij} > d\}, B_l = \{x_i \in B: x_{ij} \leq d\}, B_r = \{x_i \in B: x_{ij} > d\}$ and the coefficient $\frac{1-a}{b}$ is used to normalize the sum of gradients over B back to the size of A^c.

EFB can be done by first constructing a graph with weighted edges. Then the features are sorted by their degrees in the graph in a descending order. Lastly, each feature in the ordered list is checked and either assigned to an existing bundle with a small conflict, or a new bundle is created. The algorithm can bundle many exclusive features with more dense features, and thus avoiding unnecessary computation for zero feature values. The basic histogram-based algorithm can be optimized towards ignoring the zero feature values by the use of a table for each histogram building for each feature to record the data with nonzero values.

A5 Scopus Query

ISSN({1547-7185}) OR ISSN({1532-7663}) OR ISSN({1537-5277}) OR ISSN({1547-7193}) OR ISSN({1552-7824}) OR ISSN({1526-548X}) OR ISSN({1873-8001}) OR ISSN({1873-3271}) OR ISSN({1758-7123}) OR ISSN({1873-2062}) OR ISSN({1758-6763}) OR ISSN({1557-7805}) OR ISSN({1740-1909}) OR ISSN({1520-6653}) OR ISSN({1547-7215}) OR ISSN({1547-7207}) OR ISSN({1573-059X}) OR ISSN({1741-301X}) OR ISSN({1520-6793}) OR ISSN({1573-711X}) OR ISSN({1477-223X}) OR ISSN({1422-8890}) OR ISSN({1759-3948}) OR ISSN({1470-6431}) OR ISSN({2515-2173}) OR ISSN({1758-6690}) OR ISSN({1479-1803}) OR ISSN({2052-1189}) OR ISSN({1547-0628}) OR ISSN({1745-6606}) OR ISSN({1479-1838}) OR ISSN({1552-6534}) OR ISSN({1472-1376}) OR ISSN({1069-6679}) OR ISSN({1557-7813}) OR ISSN({1873-1384}) OR ISSN({0887-6045}) OR ISSN({1466-4488}) OR ISSN({1352-2752}) OR ISSN({1758-4248}) OR ISSN({1441-3582}) OR ISSN({1356-3289}) OR ISSN({1479-1889}) OR ISSN({1545-0864}) OR ISSN({1758-5937}) OR ISSN({1741-8100}) OR ISSN({1479-103X}) OR ISSN({1758-4248}) OR ISSN({1441-3582}) OR ISSN({1356-3289}) OR ISSN({1479-1889}) OR ISSN({1545-0864}) OR ISSN({1758-5937}) OR ISSN({1741-8100}) OR ISSN({1479-103X}) OR ISSN({1741-8798}) OR ISSN({1466-4402}) OR ISSN({1865-1992}) OR ISSN({1363-254X}) OR ISSN({0736-3761}) OR ISSN({2164-7313}) OR ISSN({1758-7433}) OR ISSN({1479-1846}) OR ISSN({2325-4483}) OR ISSN({1528-6975}) OR ISSN({1525-2019}) OR ISSN({2050-3318}) OR ISSN({1540-7039}) OR ISSN({1466-4445}) OR ISSN({1540-7144}) OR ISSN({1540-6997}) OR ISSN({2054-1643}) OR ISSN({1533-2675}) OR ISSN({2040-7130}) OR ISSN({2042-6771}) OR ISSN({1758-8049}) OR ISSN({1546-5616}) OR ISSN({1533-2977}) OR ISSN({1539-4093}) OR ISSN({1758-7212}) ISSN({1057-7408}) OR ISSN({0093-5301}) OR ISSN ({0167-8116}) OR ISSN({0022-4359}) OR ISSN({0309-0566}) OR ISSN({0019-8501}) OR ISSN({0265-1335}) OR ISSN({0091-3367}) OR ISSN({0265-2323}) OR ISSN({1361-2026}) OR ISSN({1061-0421}) OR ISSN({2040-7122}) OR ISSN({0263-4503}) OR ISSN({0959-0552}) OR ISSN({0885-8624}) OR ISSN({0969-6989}) OR ISSN({00222437}) OR ISSN({1069031X}) OR ISSN({08858624}) OR ISSN({00989258}) OR ISSN({14707853}) OR ISSN({18698182}) OR ISSN({08841241}) OR ISSN({10495142}) OR ISSN({18698182}) OR ISSN({13527266}) OR ISSN({14747979}) OR ISSN({20515707}) OR ISSN({20503318})AND PUBYEAR AFT 1991 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))

AND PUBYEAR AFT 1991 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))

A6 Primary Scopus Data

Table A2: The article and metadata dataset as imported from Bibliometrix

Variable name	Description	Value
AU	The authors who wrote the article	SURNAME INITIALS; SURNAME INITIALS **Datatype:** Character
DE	The author made keywords about the article	KEYWORD; KEYWORD **Datatype:** Character
C1	Author address	ADDRESS; ADDRESS **Datatype:** Character
CR	Cited references in article	REFERENCE; REFERENCE **Datatype:** Character
JI	ISO Source Abbreviation	ISO; ISO **Datatype:** Character
AB	The abstract of the article	TEXT INPUT **Datatype:** Character
RP	Reprint address	REPRINT; REPRINT **Datatype:** Character
DT	Document type	ARTICLE **Datatype:** Character
DI	The DOI of the article	10.1287/mksc.2021.123 **Datatype:** Character
FU	Funding Agency and Grant Number	NEW YORK UNIVERSITY **Datatype:** Character
SN	ISSN Serial number of the journal	00222429 **Datatype:** Character
SO	Journal name	JOURNAL OF MARKETING **Datatype:** Character
LA	The written language of the article	ENGLISH **Datatype:** Character
TC	Total number of citations	115 **Datatype:** Numeric
PN	The number of pages the article has	5 **Datatype:** Numeric
PP	The page numbers where the article is in the journal	48-66 **Datatype:** Character
PU	The name of the publisher	SPRINGER **Datatype:** Character
DB	Database retrieved from	SCOPUS **Datatype:** Character
TI	Title of the article	TITLE **Datatype:** Character
url	SCOPUS URL to the article	https://www.scopus.com/inward/record.uri?eid=2-s2.0-8510 **Datatype:** Character
VL	Journal volume	86 **Datatype:** Numeric
PY	The publication year of the article	2022 **Datatype:** Numeric
FX	Funding text	FUNDING TEXT **Datatype:** Character
AU_UN	The author's affiliations	UNIVERSITY; UNIVERSITY **Datatype:** Character
AU1_UN	Corresponding author's affiliation	AFFILIATION; AFFILIATION **Datatype:** Character
SR_FULL	Short full reference	TSAI YL, 2021, MARK SCI

		Datatype: Character
SR	Short reference	TSAI YL, 2021, MARK SCI **Datatype:** Character

Table A3: Subset of the yealy citation dataset from 2022-2021

Variable name	**Description**	**Value**
doi	DOI of the article	10.1287/mksc.2021.123 **Datatype:** Character
title	Title of the article	Title **Datatype:** Character
X2022	New citations in 2022	7 **Datatype:** Numeric
X2021	New citations in 2021	5 **Datatype:** Numeric
total	Total citations at the retrieved date	12 **Datatype:** Numeric

Table A4: The dataset used to get the publication month of an article

Variable name	**Description**	**Value**
prism.doi	The DOI of an article	10.1177/02761467211062504 **Datatype:** Character
prism.coverDate	Publication date of an article in the journal	2022-03-01 **Datatype:** Unknown
openaccess	Whether an article is available for free	1 or 0 **Datatype:** Numeric

A7 ABS AJG Data

Table A5: The ABS AJG dataset

Variable name	**Description**	**Value**
ISSN	ISSN of the journal	1526-548X **Datatype:** Character
Field	The academic field of the journal	MKT **Datatype:** Character
Journal Title	The title of the journal	MARKETING SCIENCE **Datatype:** Character
Publisher Name	The publisher of the journal	Emerald **Datatype:** Character

AJG 2021	The AJG 2021 score	4 **Datatype:** Numeric
AJG 2018	The AJG 2018 score	4 **Datatype:** Numeric
AJG 2015	The AJG 2015 score	4 **Datatype:** Numeric
ABS 2010	The ABS 2010 score	4 **Datatype:** Numeric
Journal Citescore	The journal citescore	11 **Datatype:** Numeric
SNIP rank	The SNIP rank	9 **Datatype:** Numeric
SJR rank	The SJR rank	3 **Datatype:** Numeric
Citescore rank	The Citescore rank	9 **Datatype:** Numeric

A8 Datset Variable Overview

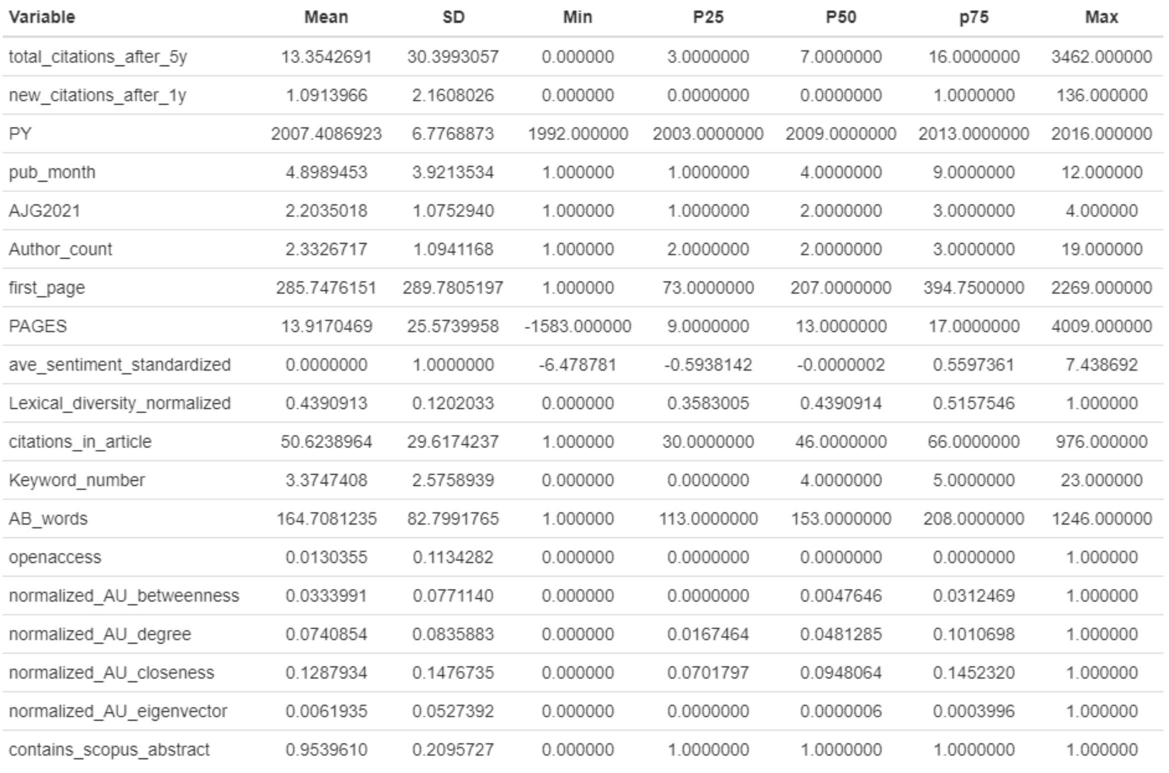
Table A6: Variable overview in dataset

Variable name	**Description**	**Value**
DI	DOI of the article	10.1287/mksc.2021.123 **Datatype:** Character
TI	Title of the article	Title **Datatype:** Character
PY	Publication year	1992 **Datatype:** Integer
pub_month	The month the article was published	5 **Datatype:** Integer
total_citations_after_5y	Total citations five years after the publication year	56 **Datatype:** Integer
new_citations_after_1y	New citations obtained the year after the publication year	6 **Datatype:** Integer
AJG 2021	The AJG 2021 score	4 **Datatype:** Integer
Author_count	The number of authors contributing to the article	2 **Datatype:** Integer
first_page	The starting page of the article in the journal	16 **Datatype:** Integer
PAGES	The articles' number of pages	6 **Datatype:** Integer
ave_sentiment_standardized	Standardized sentiment of the abstract of the article	-0,16 **Datatype:** Float

Lexical_diversity_normalized	The lexical diversity of the articles' abstract	0,04 **Datatype:** Float
citations_in_article	The number of citations in the article	25 **Datatype:** Integer
Keyword_number	The number of author keywords	5 **Datatype:** Integer
AB_words	The number of abstract words	200 **Datatype:** Integer
openaccess	Whether the article is open-access or not	1 **Datatype:** Boolean
normalized_AU_betweenness	The highest betweenness centrality between all authors of the article	0.4 **Datatype:** Float
normalized_AU_degree	The highest degree centrality between all authors of the article	0.3 **Datatype:** Float
normalized_AU_closeness	The lowest closeness centrality between all authors of the article	0.7 **Datatype:** Float
normalized_AU_eigenvector	The highest eigenvector centrality between all authors of the article	0.8 **Datatype:** Float
Contains_scopus_abstract	Whether the article has an abstract on Scopus	0 **Datatype:** Boolean

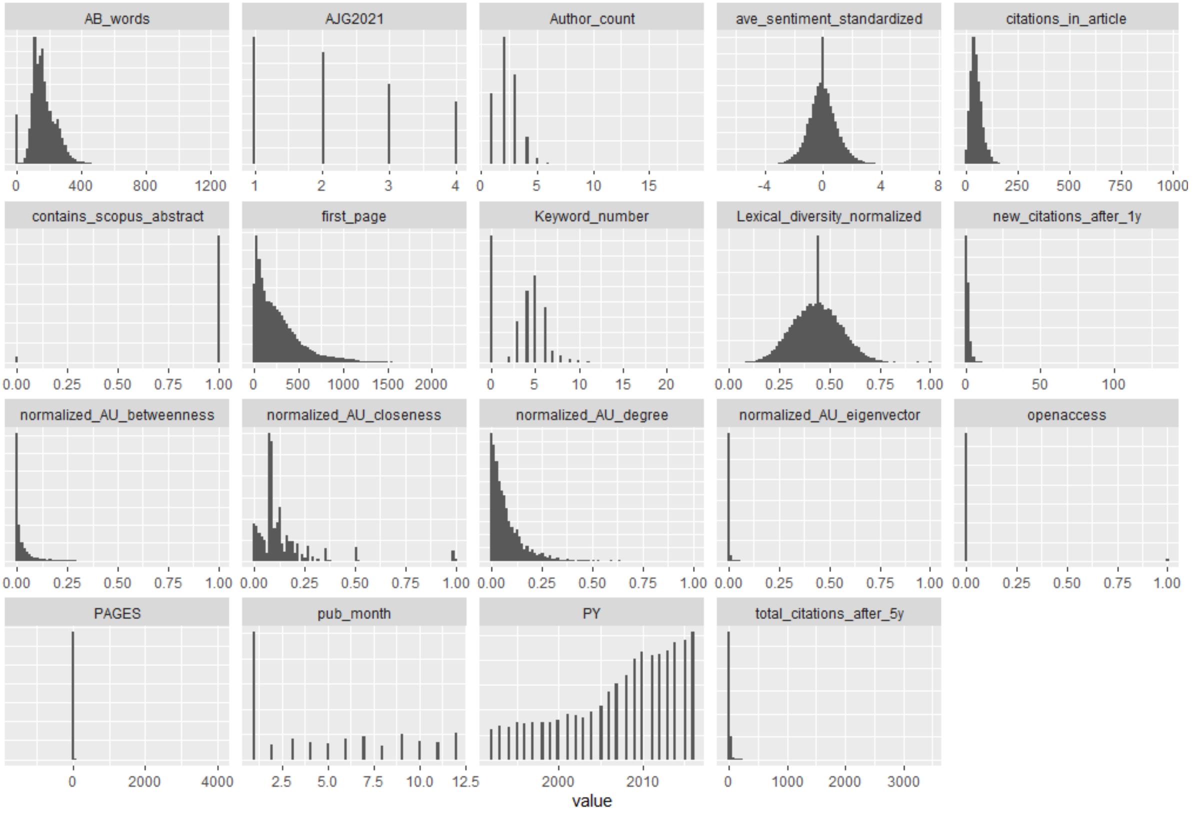
A9 Dataset Statistics Overview

Table A7: Summary statistics of the dataset



Variable	Mean	SD	Min	P25	P50	p75	Max
total_citations_after_5y	13.3542691	30.3993057	0.000000	3.0000000	7.0000000	16.0000000	3462.000000
new_citations_after_1y	1.0913966	2.1608026	0.000000	0.0000000	0.0000000	1.0000000	136.000000
PY	2007.4086923	6.7768873	1992.000000	2003.0000000	2009.0000000	2013.0000000	2016.000000
pub_month	4.8989453	3.9213534	1.000000	1.0000000	4.0000000	9.0000000	12.000000
AJG2021	2.2035018	1.0752940	1.000000	1.0000000	2.0000000	3.0000000	4.000000
Author_count	2.3326717	1.0941168	1.000000	2.0000000	2.0000000	3.0000000	19.000000
first_page	285.7476151	289.7805197	1.000000	73.0000000	207.0000000	394.7500000	2269.000000
PAGES	13.9170469	25.5739958	-1583.000000	9.0000000	13.0000000	17.0000000	4009.000000
ave_sentiment_standardized	0.0000000	1.0000000	-6.478781	-0.5938142	-0.0000002	0.5597361	7.438692
Lexical_diversity_normalized	0.4390913	0.1202033	0.000000	0.3583005	0.4390914	0.5157546	1.000000
citations_in_article	50.6238964	29.6174237	1.000000	30.0000000	46.0000000	66.0000000	976.000000
Keyword_number	3.3747408	2.5758939	0.000000	0.0000000	4.0000000	5.0000000	23.000000
AB_words	164.7081235	82.7991765	1.000000	113.0000000	153.0000000	208.0000000	1246.000000
openaccess	0.0130355	0.1134282	0.000000	0.0000000	0.0000000	0.0000000	1.000000
normalized_AU_betweenness	0.0333991	0.0771140	0.000000	0.0000000	0.0047646	0.0312469	1.000000
normalized_AU_degree	0.0740854	0.0835883	0.000000	0.0167464	0.0481285	0.1010698	1.000000
normalized_AU_closeness	0.1287934	0.1476735	0.000000	0.0701797	0.0948064	0.1452320	1.000000
normalized_AU_eigenvector	0.0061935	0.0527392	0.000000	0.0000000	0.0000006	0.0003996	1.000000
contains_scopus_abstract	0.9539610	0.2095727	0.000000	1.0000000	1.0000000	1.0000000	1.000000

Figure A2: Variable distributions


AB_words
AJG2021
Author_count
ave_sentiment_standardized
citations_in_article
contains_scopus_abstract
first_page
Keyword_number
Lexical_diversity_normalized
new_citations_after_1y
normalized_AU_betweenness
normalized_AU_closeness
normalized_AU_degree
normalized_AU_eigenvector
openaccess
PAGES
pub_month
PY
total_citations_after_5y
value