WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Bachelor Thesis

Evaluating Citation Quality and Relevance Using Text Embeddings

Matthias Prokesch

Date of Birth: 31.07.2000
Student ID: 11848129

Subject Area: Information Business

Studienkennzahl: 033 561

Supervisor: Dr. Amin Anjomshoaa

Date of Submission: 09.09.2025

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK
UND OPERATIONS MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND OPERATIONS
MANAGEMENT

Contents

List of Figures

Abstract

Traditional metrics like citation counts are a cornerstone for evaluating scientific work, yet they often provide an incomplete picture of true research quality. Because the motivations for citing are complex, ranging from core intellectual acknowledgment to strategic persuasion, simple counts of citations can be a limited proxy for a work's actual impact.

To complement existing metrics, this thesis explores the development of an automated "Citation Score" that uses AI & Large Language Model (LLM) technology to evaluate the quality and relevance of an article's citations. Unlike volume-based metrics, this score treats each citation individually, aiming to provide a more granular and context-aware measure of an article's scholarly foundation. A fully functioning system must identify the relevant citation context, infer the author's motivation, and assess how well a cited work supports the claims being made, and in this thesis we explore the foundation for such a tool.

This score is explored in part through an extensive analysis of relevant literature. Specifically, literature relevant to traditional citation analysis is analyzed for implications for a possible high-complexity implementation of an article scoring system utilizing AI & Large Language Model (LLM) technology. Furthermore, a simplified prototype was developed that uses document embeddings to measure the semantic similarity between the abstracts of a citing paper and the articles it references. The prototype showed mixed results; its Citation Score has a statistically significant positive relationship with the established metric of the citation count (how many times the paper was cited); however, its positive relationship to the journal H-Index does not reach statistical significance. The work validates the potential of a semantic-based approach and provides groundwork for more sophisticated tools for evaluating scholarly communication.

1 Introduction

Citations are a fundamental mechanism in the scientific process. They are supposed to make an article's conclusions transparent, fostering the trust necessary to continually build upon prior knowledge, and they play a crucial role in distributing credit to scientists for their work. For decades, quantitative indicators based on citations have been used not just to evaluate articles, but also the researchers and institutions behind them, a practice that has been influential yet highly controversial (Garfield, 1979).

A central claim of this thesis is that this quantitative approach is flawed. Simple citation counts are an unreliable proxy for intellectual influence because they treat all references as equal, ignoring the complex motivations behind them (Teufel et al., 2006). The reasons for citing are nuanced; they can include social factors, passing mentions, or even negative critiques, all of which dilute the meaning of a simple count. This creates significant issues for accurately assessing the quality of research (MacRoberts & MacRoberts, 1996).

1.1 An Opportunity with AI

The limitations of traditional metrics point to a need for more advanced techniques, such as Content-based citation analysis (CCA), which moves beyond mere counts to analyze the meaning behind a citation (Ding et al., 2014). Recent advances in Artificial Intelligence (AI), particularly in Large Language Model (LLM) technology, have made such an analysis possible at a large scale. Unlike older computational methods that relied on surface-level features, modern LLMs can capture rich semantic relationships, enabling a more nuanced judgment of how and why a paper cites prior work.

This technological shift creates an opportunity to develop different types of evaluation metrics, which are more meaningful. In this thesis, in particular, we explore this opportunity by proposing and developing a "Citation Score", an automated method for analyzing a scientific article based on the quality and relevance of the citations it makes.

1.2 Research Question

The central research question guiding this thesis is:

How can a scientific paper automatically be assigned a "Citation Score', reflecting the quality & relevance of its citations?

To address this primary question, the following sub-questions are explored:

- *What state-of-the-art literature on citation analysis is necessary to develop an automated quality & relevance based Citation Score?*
- *How can a simplified prototype of a Citation Score system be implemented?*

1.3 Methodology and Thesis Structure

To answer these questions, a two-phase research methodology has been adopted.

First, we conduct an analysis of literature to build a conceptual framework. This involves reviewing existing and state-of-the-art literature on citation analysis to identify the key factors that define citation quality. We explore the prevailing theories on citation (such as Normative Theory and Social Constructivism) and formulate a comprehensive taxonomy of citation motivations to inform the design of a potential complex implementation of a citation quality & relevance score model, as well as to inform the design of a simplified implementation.

Second, we develop the prototype implementation to operationalize some of these theoretical insights. The development of a full automated system for the task would generally be complex; it requires identifying the relevant citation context, inferring the author's motivation, and evaluating how well the cited work supports the claims of the citing paper (Hernández-Alvarez & Gomez, 2016). The core of our simplified prototype involves using a state-of-the-art model to generate document embeddings from paper abstracts, allowing for a quantitative measure of their semantic similarity. An "embedding" is a numerical representation of text in a high-dimensional vector space, where the geometric distance between two vectors corresponds to the semantic similarity of the texts they represent.

The following chapters will detail this process, presenting the results from the prototype and concluding with a discussion of the findings, limitations, and directions for future work.

For the analysis of the literature, the important (full) search terms to discover the literature on the topic were: “Citation Function”, “Citation Motivation”, “Citation Analysis”, “Citation Analysis LLM”, “Citation Analysis NLP”, “Citation Analysis Machine Learning” and “Citation Context Analysis. Most of the literature searches were performed through Google Scholar. Scholarly papers were more closely selected on the basis of the title, partially the number of times it was cited, the year of release, and finally the abstract. Some foundational papers were discovered through repeated mentions in the relevant literature and as a result were analyzed independently. Tools for the practical implementation of a simple Citation Score were discovered independently to the previous search process, and in second order the published literature related to those tools was brought into the synthesis and review of literature.

2 Research Background

We now explore the discussed necessary research background and an analysis thereof. Generally, citation analysis as an area of research is situated as a signature technique of bibliometrics, whereas scientometrics adopts those same techniques to investigate the broader system of science. (Hood & Wilson, 2001)

2.1 Quantitative Citation Analysis

As the first literature behind what is needed to understand citing behavior, and the research aims related to a potential Citation Score Model, that we now try to understand, is on how traditional, quantitative citation analysis is conducted. This includes a "Citation Count", where for a paper the number of times it was cited is counted and then used as a metric to judge the paper. (Szomszor et al., 2021)

2.1.1 Importance of Quantitative Analysis of Scientific Works

We begin by understanding arguments on how widespread and important quantitative analyses of certain factors, such as how often academic works are cited, are.

As producing a quantitative result for given articles is also explored by us this thesis, this chapter can help us contextualize the importance, and the place in literature, of the investigated Citation Score.

A paper by Aksnes et al. (2019) talks about this quantitative analysis of scholarly work, and about indicators related to scientific citations and references in particular.

It also talks about how this analysis has, to many, extended what is understood to be of assured quality, and what isn't, beyond whether a scientific article is peer-reviewed or not. They cite research which claims that quantitative citation analysis and other quantitative tools can even be superior predictors to the binary of an available peer-review in determining certain factors about a research paper, such as how impactful it may be, although they added caveats to making conclusions about evaluating aspects of the paper, such as its citation quality.

Several other works such as by Teufel et al. (2006) and MacRoberts and

MacRoberts (1996) also point out how "Bibliometric metrics" will commonly be used to understand the relevance and importance of a scientific paper or, in Ding et al. (2014), how thoroughly entire diverse research disciplines and fields have been investigated, through connections and recurring themes of research output, by utilization of such measures. Even how not just scientific articles, but by extension researchers and groups of researchers may be evaluated through such metrics has been described and discussed, and this discussion has been led more emotionally charged, relative to other uses of such metrics, according to an article by Garfield (1979), showing how these metrics have been controversial for the span of decades. A review paper by Bornmann & Daniel (2006) talks about how such indicators have been pioneered in 1927 and since then came to influence even the governmental politics and incentives around science, as well as listing other influences already discussed.

2.1.2 Criticism of Quantitative Analysis of Scientific Works

Such quantitative indicators used to investigate the quality & relevance of individual scholarly works are generally not without criticism and controversy throughout the myriad of uses available. This chapter will now begin to investigate the extensive theories and repeated, claimed "shown to be true", criticisms of such measures, which has relevance first in understanding what must be avoided both when using metrics such as a Citation Count for a post-hoc analysis of the Citation Score, but also, secondly and importantly, what must be avoided in creating and using such a Citation Score, when viewed from with a perspective on traditional metrics based on related citation-data, and thirdly how this underlines the relevancy of exploring new avenues of (quantitative or otherwise relatively automated) analysis of scientific works.

Teufel et al. (2006) caution the use of the quantitative indicator of citation tallies by citing articles showing them to be an unreliable proxy for intellectual influence, since many references are made for social reasons, and because negative or passing mentions should not be credited on-par with citations that genuinely extend earlier work. This is also important for the flip-side: For citations made in an article, theses criticisms remain true, just like they would for citations made of an article. More importantly, for a Citation Score, these criticisms, such as that a citation might have been made for social reasons, are the point.

Others argue as well, for a multitude of factors, that citations should not be treated equally and are too complex for this. (Ding et al., 2014) (Garfield, 1979)

Because citation counts are shaped by many technical and social factors, pure quality signals are described as easily drowned out by this enormous complexity, these intrinsic limitations, the inherent complexity of such a task, will remain the challenge even, as they argue, with different, ever more complex or sophisticated, metrics. (Aksnes et al., 2019)

The paper by Garfield (1979) concludes that many articles, such as the articles mentioned, whilst not explicitly mentioned in the paper, talk about "negative citations", "self-citations" and "methodological papers" as factors that dilute any real analysis of pure quality in the case of quantitative metrics. Yet, these criticisms, particularly "self-citation", further bolster the argument for the proposed Citation Score analysis.

In the article "Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories" by Aksnes et al. (2019) the authors, again, list a body of work criticizing higher level versions of citation metrics, such as evaluating journals through them. As more relevant for our purposes they also again stress, through their own through analysis of a wide body of research, that citations only partly register scientific impact, and this being only one facet of research quality, whereas other equally important dimensions such as "plausibility", "originality", and "societal value" (Aksnes et al., 2019, p. 8) are escaping detection. The concept of plausibility, an article being "plausibly" cited, coming to mind for an analysis of a Citation Score based primarily on the comparison of abstracts.

MacRoberts & MacRoberts (1996) had written with conviction about the need to test the "empirically testable" hypotheses behind bibliometrics and the lack thereof. Claims in articles cited by them around how "objective" and "concrete" the analysis of such bibliographic metrics potentially is can and should be tested, they argue.

They describe a design for an empirical study, where a human reads a citing paper, understands its context and evaluates if it is meaningfully and conclusively cited from the context of citing papers, and how much is missing in the information from the cited papers. This description in "Problems of Citation Analysis" by MacRoberts and MacRoberts (1996) matches an underlying, yet inverse, idea behind an automated analysis of citations, that scientific articles can be empirically evaluated if citations/citation contexts are semantically compared to semantic information found in the citing articles, well.

They went on to empirically analyze the origins, traces and transmissions of facts and information, and how a lot of this information is systematically missing in metrics based on citations, and show how in certain fields citing is not as much of a rigorous practice, even when the information is used. (MacRoberts and MacRoberts, 1996)

For other specific empirical criticisms of citation-based metrics Aksnes et al. (2019) talk of "band-wagon" effects for citation metrics, skewed distributions, both between fields as well as between papers, that often gain traction from this aforementioned "band-wagon" effect, and other issues. They claim there to be not much evidence pointing in favor of choosing any such "citation metrics", such as a Citation Count, over established peer-reviewed processes, when the option is presented. (Aksnes et al., 2019)

2.2 Direct Citation Analysis

This chapter will now investigate the analysis of citations not as a mere "count", with each citation of equal value, but dive into the theory of what citations are, what they represent, what motivates scientists to use them and what attempts at analysis there are, that treat each citation differently, based on citation context or other factors.

2.2.1 Theories on Scientific Citations

In the scientific literature we find articles depicting two kinds of theories on citations. These have differing designations, such as the "normative theory" vs. "social constructivist view" (Ding et al., 2014, p. 6), or "traditional scientific view" vs. "social constructivism" (MacRoberts & MacRoberts, 1996, p. 4), but the contrast drawn between each binary proposed has strong overlap in the articles examined.

Therefore, these many binaries on theories of citations are now examined as a singular binary of theories on citations. The competing theories will now be called "Social Constructivist View" and "Normative Theory" in the following pages. A groundbreaking paper for the Traditional Scientific Theory is the paper by Merton (1973) (The paper was originally published in 1942).

For the purposes of this thesis an important conclusion to draw is that Traditional Scientific Theory describes citations as good data to draw further conclusions without adjustments, whereas the Social Constructivist View implies a necessity to ask questions about such conclusions, with our analysis in this thesis aiming at how and why to ask precisely these questions. (MacRoberts & MacRoberts, 1996) (Ding et al., 2014)

2.2.1.1 Normative Theory

This theory, which is originally rooted in the mentioned sociology of science pioneered by Merton (1973), posits that the scientific community operates under shared norms which guide behavior. Both theories are situated in social theories of science. (Bornmann & Daniel, 2008)

Within this theory citation is viewed as a formal mechanism for correctly distributing credit, making it a sort of, non-literal, currency credibly used to pay intellectual respect to colleagues whose work has been used or has had an influence on the citing work. (Aksnes et al., 2019) According to this theory, which Erikson & Erlandson (2014) outline as well, the act of citing is a rational act, a behaviour governed by mentioned shared norms, a behavior that is essential to the system of rewards and incentives in science and ensures that recognition is appropriately distributed for contributions to knowledge.

2.2.1.2 Social Constructivist View

In contrast, the Social Constructivist view, according to Bornmann & Daniel (2008) strongly associated in early conception with sociologist Gilbert, in a foundational work in 1977, argues citing to be primarily a rhetorical act of persuasion. (MacRoberts & MacRoberts, 1996).

This makes a scientific paper not a neutral report of findings, but rather a document designed to persuade and convince others, mainly in the scientific community, of validity, importance, novelty, .. of the author's claims. (Erikson & Erlandson, 2014) (Lyu et al., 2021)

For another aspect of the Social Constructivist View, that has remained largely unmentioned, Gilbert (1977) also succinctly highlights, in his foundational paper for the Social Constructivist View, that "..efforts at persuasion usually depends not only on the intrinsic quality of the arguments put forward, but also on the parties' relative power." (p. 8).

As Brooks (1985) concluded, in a first study, already after interviews with authors themselves, "authors can be pictured as intellectual partisans of their own opinions, scouring the literature for justification". (Brooks, 1985, p. 227)

Authors strategically select to strengthen their own arguments and align their work with established figures of authority, they discredit rival approaches, and they carve out a niche for their own contribution, if they select some material for citation but not other material, or leave components of the material systematically unmentioned. Brooks concludes this after further interviews, in an article that presents a more formal and comprehensive theory of motivations yet. (Brooks, 1986)

Moving beyond the binary, in modern conceptions these two theories are largely not viewed mutually exclusive. Rather they are viewed as describing different facets of a highly complex, and yet precisely definable, single, common behavior.

As Cozzens (1989), which Bornmann and Daniel (2008) cite, summarizes, citations stand at the intersection of being a rhetorical system for persuading peers and being a reward system for allocating credit, with this inherent complexity and multiplicity of motives as a crucial finding, one that we already saw discussed with Brooks (1986). These more formal and comprehensive theories of citation motivations we could already read about in relation to the article by Brooks (1986) and by Bornmann and Daniels summary of Cozzens (1989), or “Taxonomies” of citation motivations, are what we will look at in the next chapter. The complexity and multiplicity of motivations confirms that monolithic interpretations of citations are inadequate and reinforce the need for analytical methods that can deconstruct and classify these different intentions, with empirical evidence that supports this complexity, as the study by Brooks (1986) found that over 70% of references were attributed by their authors to more than one motive, and with motivations clustering best into groups related to “persuasion”, “negative credit”, and providing useful information to the reader. Beyond deconstructing those intentions, the existence and widespread occurrence of them, and the implications that some common intentions are more scientific in nature than others, makes it plausible to evaluate an article by its citations and to draw conclusions about the quality of the article from that process.

2.2.2 A Taxonomy of Citation Motivation for Automated Classification

The paper by Bornmann & Daniel (2008), that has already been investigated in other respects, also, as a focal point of their work, comprehensively reviews studies investigating citing behavior of scientists, using papers up until 2005. It seeks to understand the different motivations of scientists to make citations, specifically also for non-scientific reasons, apart from intellectual and cognitive influences, as the paper seeks to understand "What do citation counts measure?". For this analysis into the theoretical "high-complexity" Citation Score it is relevant to investigate citation motivation generally, it is also valuable implicitly as it pertains to how relevant the abstract-similarity is for the underlying quality of citations, and in which cases it may be more or less relevant, citation motivation is due to this also explicitly relevant for both the methodological pipeline of the prototype, as well as for the theoretical methodological pipeline of a "high-complexity" version of such a model implementation.

To analyze citation motivation is also often a form of "semantic" citation analysis, as referred to by Ding et al. (2014), differentiating different types of citation analysis will be explored further in later sections of our analysis.

There is, once again, a large number of different schemes available in research. Schemes of classification-options of citation motivation & function, or schemes of full taxonomies on citation motivation & citation function, that researchers have built on the previously described theoretical foundations of understanding citations. As this is integral for a part of the suggested theoretical methodological pipeline in this thesis, and highly relevant for other sections of the literature analysis, the aim now is to aid the choice of an appropriate taxonomy of citation function and motivation, using as input the multitude of (relevant) overlaps between the different classification schemes found in research.

One of the early and most influential schemes was proposed by Moravcsik and Murugesan (1975), as described by Bornmann & Daniel (2008), Swales (1986) and Teufel et al. (2006). They analyzed citations on 4 distinct, unordered levels, categorizing them into two different groups, as well as a "neither" for each level:

- They have a level contrasting immaterial, more abstract concepts in the cited article vs. functional concepts and objects, "conceptual or operational". (Moravcsik and Murugesan, 1975, p. 4)

- A level asking if the citation is needed to understand the citing paper, or rather to give credit to the cited paper, whether "organic or perfunctory". (Moravcsik and Murugesan, 1975, p. 4)
- They identify a level contrasting whether the citation expands on the cited material or is merely showing a differing perspective on the matter, "evolutionary or juxtapositional". (Moravcsik and Murugesan, 1975, p. 4)
- And a level asking whether the citation shows agreement or disagreement with the cited paper, whether "confirmative or negational". (Moravcsik and Murugesan, 1975, p. 4)

Here we again see reflected the established complexity and the established many dimensions on which we can meaningfully analyze citations.

The paper by Teufel et al. (2006) is one of the various expansions and additions of these ideas that happened over time. They created a "12-Category" scheme that is designed for computational analysis, not much unlike to what we are suggesting. It goes into detail for this purpose, with specific functions of citations such as:

- "Weak": The citing paper points out a weakness in the cited work.
- "CoCo": The work in the citing paper is said to be superior to the cited work.
- "PUse": The citing paper utilizes a procedure from the cited work.
- "PModi": The citing paper modifies a procedure from the cited work.
- "PBas": The cited work is used as a starting point for understanding something.

A paper by Liu (1993), "A study of citing motivation of Chinese scientists," shows that scientists in their analyzed cases had often cited based on their available home library, and generally cited more if they did so more. They also show that citation, as it relates to internal motives, often is done to demonstrate expertise and to signal prestige. They also argue, once again, that citations are an imperfect proxy of impact due to various internal and external reasons, and that more nuanced, individualized (per citation) processes are needed.

Jha et al. (2016) developed a multi-category taxonomy as well, also basing it on previous work, actually including the paper by Teufel et al. (2006) and others and coming up with 6 categories. Their research specifically deals with newer models and therefore gaining higher relevancy for the purposes of our analysis.

"The six categories proposed by Jha et al. (2016) are:"

Criticizing This category refers to discussing what a cited paper did right or wrong, or whether it is trustworthy. This can include pointing out that a paper attempted something but did not achieve significant outcomes.

Comparison The citing sentence compares the work in the cited paper to the author's own work, for instance by presenting the author's work as an alternative approach.

Use The citing paper uses a method, tool, or dataset from the cited paper.

Substantiating A result from the citing work verifies or supports a claim made in the cited paper.

Basis The cited work is used as a foundation or starting point that the citing paper expands upon.

Neutral The citing sentence provides a neutral description of the cited work or does not fit into the other categories.

They also showed a strong correlation between these purpose labels and the options of "positive", "negative" and "neural", their analysis e.g. showed that 99% of citations falling in "Substantiating" were positive, while 67% of citations falling under "Criticizing" were negative. Here as well that same seemingly slightly unique use of "negative" and "positive" occurred. (Jha et al., 2016)

More recently Lyu et al. (2021) conducted a very extensive meta-synthesis of articles related to citation motivation, in which their search process for relevant articles was a formalized multi-step process itself, with some of the articles found and ultimately used being articles we have described here. (Interestingly, they identified 1771 studies and of those "38 met the inclusion criteria", (Lyu et al., 2021, p. 5) and many of those appear to overlap with the results on the first pages of Google Scholar, if one searches for some of their search terms. This is known to us because many of the 38 articles also appear in our reference list.)

They then identified "Thirty-five descriptive concepts" which they grouped into "13 analytic themes". Through this they identified that the themes belong to the larger groups of either "scientific motivations" and "tactical motivations". (Lyu et al., 2021, p.1)

So far the analysis has not distinguished much between the function and motivation of a citation, Lyu et al. (2021) point out a study that argues for this distinction, arguing that motivation is the broader term to refer to the decision to use or not use an entire source or not. They give the example of not even having "read a particular source" (Lyu et al., 2021, p. 6), which is a highly relevant "motivation" for our purposes. They chose to use "motivation", and the term we should use for this analysis is also "motivation", as it has much higher relevance for the topic, according to the information the article the analysis by Lyu et al. (2021) is citing provides.

Now the motivations are described, which Lyu et al. (2021) listed and it's described how they themselves describe those motivations as having been used in the literature. After this we will have enough knowledge on citation motivations, both the general motivations identified in the literature as well as the motivations identified and described when an automated analysis is involved, that we can take a look at an informative table in order to aid making a final decision on the categories to find potentially useful for a theoretical, fully implemented ("high-complexity" version), Citation Score.

Lyu et al. (2021) begin with "**Background**", which is a category describing how authors will often cite in order to do one of three things, and has been used as a category in a majority of the synthesized articles: 1. To give a historical overview on integral topics. 2. To give an overview of current recognized or established thought on the topic. 3. To give an overview over the background of alternative topics, that could have been integral to the research, but aren't.

They proceed with "**Gap**", this category is related to the background, from the perspective that it tries to understand the background that is missing and should be there. These citations establish and build an argument for how and why the citing article should construct their argument, based on this missing background.

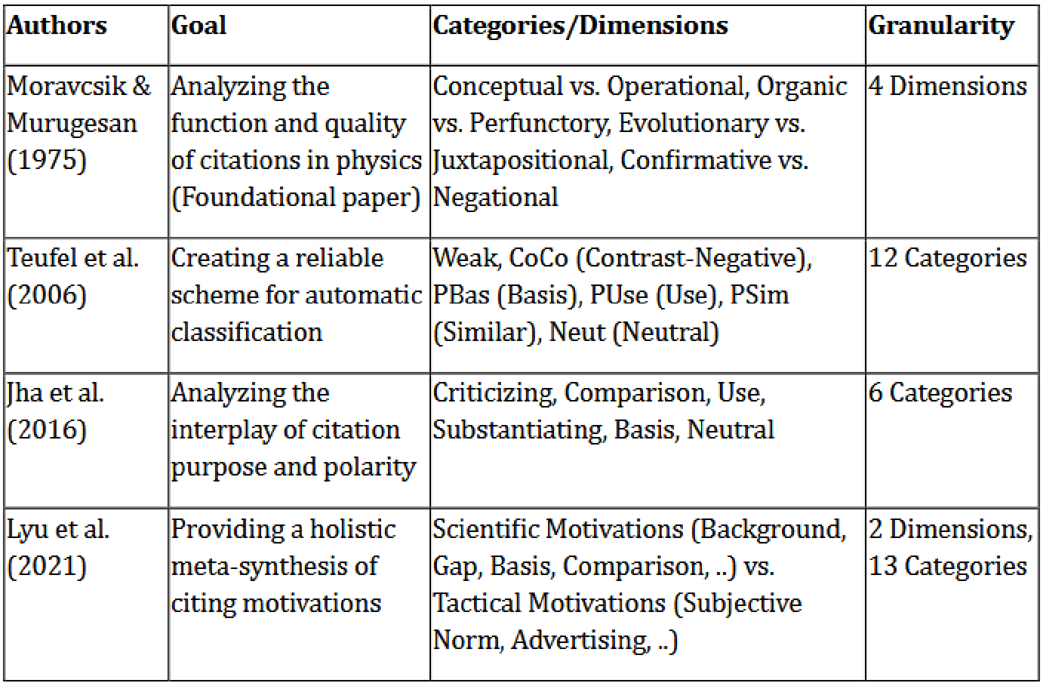
“**Basis**” is another motivation that some articles reviewed have identified, it is referring specifically to that background literature that has been most influential and important.

The next category that a majority of examined articles have identified as well is “**Comparison**”, which refers to talking about similarities and differences both between the citing article and the cited article as well as between cited articles.

They identify several more categories they found to varying degrees in the literature, that relate to motivations based on traditional scientific grounds such as “**Application**”, “**Improvement**”, “**Evidence**”, “**Further Reading**”, “**Assessment**” and "**Identification of the Originator**” (Lyu et al., 2021, p. 13).

The other groups of motivations they identified, aligning more with the Social Constructivist View or what they call “Tactical Motivation”, since the other group of motivations aligns more with the Normative View, although it is not a perfect overlap, are “Advertising”, “Profit-Seeking”, and “Subjective Norm” (Lyu et al., 2021, p. 14). “Subjective Norm” in particular refers to the authors’ (of citing papers’) perceptions of an expectation placed on them by outside (social) forces, irrespective of whether it is true or not. So, feeling like somebody in a position to judge may or may not expect a particular cited article, or feeling like it may or may not be expected to cite an article from a specific topic, or to give acknowledgement even if not furthering the goal of the research, all fall under this category of citing motivation (Lyu et al., 2021).

The content of potential further work we suggest in this thesis is directly influenced by these created taxonomies. In order to implement the suggested theoretical (“high-complexity” version) implementation of a Citation Score, the task of classifying extracted citation contexts should rely on the overlap of these established schemes. The following table provides an overview of these studies categories, and helps inform a potential final taxonomy for a classification scheme.



Authors	Goal	Categories/Dimensions	Granularity
Moravcsik & Murugesan (1975)	Analyzing the function and quality of citations in physics (Foundational paper)	Conceptual vs. Operational, Organic vs. Perfunctory, Evolutionary vs. Juxtapositional, Confirmative vs. Negational	4 Dimensions
Teufel et al. (2006)	Creating a reliable scheme for automatic classification	Weak, CoCo (Contrast-Negative), PBas (Basis), PUse (Use), PSim (Similar), Neut (Neutral)	12 Categories
Jha et al. (2016)	Analyzing the interplay of citation purpose and polarity	Criticizing, Comparison, Use, Substantiating, Basis, Neutral	6 Categories
Lyu et al. (2021)	Providing a holistic meta-synthesis of citing motivations	Scientific Motivations (Background, Gap, Basis, Comparison, ..) vs. Tactical Motivations (Subjective Norm, Advertising, ..)	2 Dimensions, 13 Categories

Figure 1: *A Comparison of Citation Motivation Classification Schemes. It also reflects the evolution of broad dimensions of motivations to a more fine-grained approach that is more suitable for an automated analysis.*

2.2.3 Manual Early Citation Analysis

After having understood the theories and possible motivations behind individual citations we can move towards understanding a direct analysis of each citation, the result of which does not treat every citation equally. A paper by Ding et al. (2014, p. 1) calls this "Content-based citation analysis (CCA)".

First we look at a manual analysis of citations, this includes some of the work we have seen around manually categorizing citations by motivation, but it also extends beyond this into areas of quality & relevance.

We are more interested in works with an analysis of citations that treat every citation differently, but still, ultimately, take aim at the entire citing paper as the sum of its citations, similarly to the aims in the practical component of the analysis later on.

For an analysis of individual citations, generally CCA is a relevant term, Ding et al. (2014) are suggesting to distinguish between “semantic” and “syntactic” citation analysis, where semantic citation analysis describes an analysis of citations that directly considers the meaning of the citation. Utilizing LLMs for citation analysis can be a form of semantic citation analysis, and syntactic citation analysis considers features about the citation not related to meaning, such as how frequent it is cited in the article or where it can be found within the structures of a scientific paper or even if it was cited more or less formally, such as how the year may have been included in the reference or if it is to be found more in introductory, methodological or conclusory sections of the paper.

Some early CCA, focusing on syntactic aspects, have shown that in the introduction-sections of a scientific paper more impactful articles will typically be found as sources, whilst suggestions have been made that a citation of a paper should be given more weight if it is e.g. found to be cited multiple times, especially in the methodological section as well, as these are typically more integral to what the citing paper is attempting to achieve. These have been among the most repeated, conclusive findings in the early stages of the syntactical citation analysis. (Ding et al., 2014)

The early, manual semantic citation analysis consists of work that has already been extensively explored in “A Taxonomy of Citation Motivation” in this analysis. This is because the challenge of manually identifying a citation to be of a certain category of “citation motivation” has heavy overlap with (first) establishing a suitable, well-defined, taxonomy of citation motivation.

2.2.4 Computational and Modern Citation Analysis

For the practical application of the theoretical understanding for automated citation analysis, Content-Based Citation Analysis (CCA) as described by Ding et al. (2014) is again relevant. The development of these and related methods has progressed from early, feature-engineered, machine learning systems to the very complex deep learning models that are central to the analysis in this thesis as well. The success of early computational methods was fundamentally limited by their reliance on surface-level features that could not incorporate the mentioned complexity and semantic understanding, they could however learn correlations of those features. This bottleneck is the main motivation for moving to deep-learning models.

2.2.4.1 Citation Context

Hernandez-Alvarez and Gomez (2016) write about the importance of how the rest of the paper, the broader citation context, relates to the citation itself, with respect to its analysis.

They explicate how many possible definitions of what the context of a citation is, from the full-text to much narrower definitions, have been tried before.

They cite several papers that deal with the question of what citation context windows might be more or less optimal, suggesting context windows that can dynamically adapt to the content. An interesting finding of one paper referenced by Hernandez-Alvarez and Gomez (2016) is that this "adapted window" is selected best, when compared with the entire context of the cited paper. This has implications for comparing abstracts of a citing and cited paper, as we now may be able to establish that it can be reasonable to assume the entire cited paper is relevant for comparison. And the semantic content of the cited abstract, understood as a shortened version of a paper, could therefore be relevant for the prototype of a Citation Quality & Relevance Score. (Hernandez-Alvarez and Gomez, 2016)

It is however important to note that Hernandez-Alvarez and Gomez (2016) themselves suggest identifying the function the citation is serving, or the argument it is making, as the best way to identify citation context.

There is a possibility that a combination of both approaches, identifying the argument around a citation **through** utilizing the full context of the cited paper, would be a more ideal approach. This approach however, as described in the methodological section, is of high-complexity and out of scope for the prototype.

One approach Hernandez-Alvarez and Gomez (2016) rule out, based on most of their reviewed literature, is to keep the citation context too narrow and short, such as to only one sentence. This would also miss citation context that is neither explicitly identifiable as such through human eyes. Another challenge is that lines of argumentation will often be followed by multiple sources, which complicates allocating the components of the argumentation to each individual citation.

Based on the reviewed arguments in the paper by Hernandez-Alvarez and Gomez (2016) it does not appear conclusive whether the most abstracted form of the citing argument and full citing paper, the "abstract", could be relevant-enough citation context or not. Furthermore, they state that research is “limited to experimental methods to find optimal fixed-sized windows of context” and that “Most of the current efforts use supervised learning algorithms. Unsupervised methods are less commonly used to define context, mainly because of the complexity of the task”. This underlines arguments for the simplified and, through this simplicity, potentially insight-gaining approach to Citation Analysis. (Hernandez-Alvarez and Gomez, 2015, p.8)

For the extraction of citation context tools like GROBID, which stands for “GeneRation Of Bibliographic Data” could play an important role. GROBID is an open-source library that uses machine learning models, specifically Conditional Random Fields, to parse scholarly documents in PDF format (Lopez, 2009). It excels at segmenting documents and extracting structured metadata, including header information, like “title”, “authors”, “abstract”, and, crucially for this thesis, parsing reference strings and associating each in-text citation marker with its surrounding textual context (Lopez, 2009). By leveraging GROBID, the proposed “high-complexity” methodology could reliably automate the extraction of the raw textual data needed for a subsequent analysis, which would form a robust foundation for such a project.

2.2.4.2 Automated Classification of Citation Motivation

After a citation context may be extracted, a potential next step is to automatically classify its citing motivation. Again, citing motivation identified through this approach may be still be relevant for Citation Quality & Relevance, because the importance of semantic similarity to sections of the cited paper may differ depending on the stated motivation to cite.

Early and established approaches to this task have relied on supervised machine learning models, most notably Support Vector Machines (SVMs), trained on manually annotated corpora (Jha et al., 2016).

The performance of these models is quite dependent on the quality and ingenuity of the features extracted from the text to serve as input, which is a clear negative of this approach in respect to Citation Quality & Relevance, with these features designed to be proxies for meaning and designed to include a wide range of linguistic and other cues.

As citation context may depend on the identified motivations, the citation context between the task of identifying the motivation and the task of identifying the subsequent semantic similarity for a Citation Score may ideally differ, especially with models based on these features:

- “Lexical” and “N-gram” Features:
  The presence of specific words or phrases acting as cues. For example, words like "however" or "contrast" could signal a comparison, whilst "builds upon" or "extends" can signal a “Basis” function (Athar & Teufel, 2012).

- Syntactic and Structural Features:
  Information about the grammatical structure of the sentence, such as “part-of-speech” tags and other dependency relations (e.g., to identify the verb closest to the citation marker/reference), which could provide strong signals (Jha et al., 2016). The already mentioned important location of the citation within the document for syntactical analysis, e.g., in Introduction, Methods or Discussion, is also a powerful feature, as already stated, as certain functions are more common in specific sections (Jha et al., 2016). The implications for grading a citing paper on Citation Quality & Relevance through “location” being fully parallel to the implications of traditional citation analysis of cited papers, as in both a citation that may be more/less important/integral to the citing text should have more weight.

- Sentiment and Polarity Features:
  The presence of words from so called sentiment lexicons, which describe lists of positive and negative words, and cues for negation or speculation are particularly useful for classifying citation polarity, which describes the overall positive, negative or neutral sentiment expressed towards the cited work (Jha et al., 2016). Research has demonstrated a strong correlation between functional categories and polarity, for instance, citations with the purpose “Criticizing” are almost always negative, while those with the purpose “Use” or “Substantiating” are majority positive (Jha et al., 2016). This link between function and sentiment, and further potential links between sentiment and similarity, the former being validated in experiments by the study of Jha et al. (2016) or other related work, could be a key insight which underpins the concept of a Citation Quality or Relevance score.

2.2.4.3 Leveraging LLMs for Citation Quality & Relevance

The limitations of feature-engineered NLP methods have shifted approaches towards deep learning models and LLMs in particular. These models learn to somehow capture the semantic meaning of text, more so than other technology, directly, from vast amounts of data, and move beyond shallow, surface-level patterns. This semantic frontier offers powerful new tools for quantifying the relevance between scientific documents and for performing nuanced, qualitative analysis at an unprecedented scale. LLMs can operate at a scale while performing tasks that require a (not quite) human-like "understanding" of language and context.

2.2.4.4 Measuring Semantic Relatedness with Document Embeddings

The core of the modern semantic approach is the concept of document embeddings, which refer to dense vector representations of documents in a high-dimensional space. For these vectors, geometric proximity corresponds to semantic similarity and documents that discuss similar concepts will have vectors that are "close" to each other. (Cohan et al., 2020) (Lagopoulos & Tsoumakas, 2021).

This allows semantic relatedness to be approximated by a calculation of the distance, e.g. the calculation of Cosine Similarity, as in our case, or the calculation of Euclidean distance, between two embedding-vectors, also depending on the specifications of the embedding model.

A state-of-the-art model for generating these embeddings for scientific documents, which will be examined more closely for the embeddings of the abstracts, is SPECTER, which stands for "Scientific Paper Embeddings using Citation-informed TransformERs". (Cohan et al., 2020). SPECTER is built on a Transformer architecture, initialized with the weights of SciBERT, a language model pretrained on scientific text (Beltagy et al., 2019). SPECTER is then fine-tuned on a very large amount of scientific papers using the citation graph as a signal of how related documents are.

This training objective that the model uses, whilst not closely examined in this thesis, is referred to as "triplet loss pretraining objective", which means that for each "query" paper the model is trained to produce an embedding that is closer to the embedding of a "positive" paper, which is defined as one that it cites, than to the embedding of a "negative" paper, which is a random paper, not cited. This forces the model to organize its embedding space according to the structure of the scientific influence, which makes it exceptionally well-suited for tasks involving document relatedness, and arguably even more well-suited still for the task that is attempted for use in this thesis. To provide an even more nuanced training signal, the model is also trained with "hard negatives", which, without explaining further, means papers that are cited by a paper that is itself cited by the “query paper” and are not directly cited by the “query paper”. (Cohan et al., 2020). What is difficult to predict with this model is how it would behave once citation context is introduced, which is not the case for our abstract-based analysis. As the model is already trained on being closer to cited papers, aspects of the described citation context analysis might already be reflected in how the model creates its embedding vectors, so a real analysis of citation quality through the consideration of citation context may involve either more generalized models or models that are otherwise better suited for the analysis of Citation Quality and Relevance.

The methodology of the analysis which is comparing the abstracts of citing and cited papers is directly supported by this technology and a reason to decide for this model, apart from “testing” it across relevant metrics, as explained later. By generating SPECTER embeddings for the abstracts of a citing-cited-paper-pair, one can obtain a very reasonable, in regards to other state-of-the-art technology, quantification of their semantic similarity. The utility of the general approach has been demonstrated in related work, such as the mentioned study by Lagopoulos and Tsoumakas (2021), which used sentence embeddings and cosine similarity to analyze the relevance of journal self-citations. They could successfully distinguish between justifiable and potentially unethical citation practices by utilizing this technology.

3 Methodology of Prototype

This chapter details the methodology used to implement the simplified prototype of a "Citation Score". The goal was to create a functional example to illuminate the core challenges and opportunities of using Large Language Models (LLMs) for a more nuanced, automated citation analysis.

3.1 Overall Process

The methodology was structured as a multi-step pipeline, as visualized in the following figure. The process moves from raw data acquisition and preprocessing to semantic analysis and the final score calculation. The following sections give a first look on the implementation details for each stage of this pipeline, which will be described in the *following sections.*

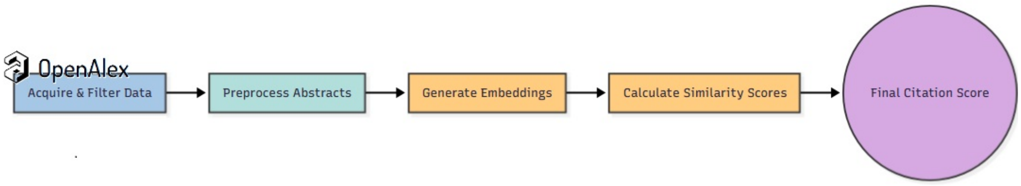


Figure 2: *A simplified flowchart of the prototype methodology, from data acquisition to the final score.*

3.2 Data Acquisition from OpenAlex

The initial data for the prototype was collected from the OpenAlex platform (OurResearch, 2025), an openly licensed scholarly graph containing over 160 million records. While its citation indicators correlate well with expert quality ratings (Alperin et al., 2024) (Thelwall & Jiang, 2025), one must account for its still-incomplete reference graph and occasional misclassifications.

To define the project's scope, the API query was filtered. The dataset was restricted to the **Computer Science** research field to ensure comparability of results, as a paper by Crespo et al. (2013) explores how different academic fields have significantly different citation practices. The query was further limited to journal “articles” published in “2019” that were listed as “Open Access”.

From the resulting papers, the following metadata fields were extracted for the prototype:

- DOI
- OpenAlex ID
- Title
- The source of the primary location
- The name of the primary location
- How often it was cited
- How often it was cited, normalized by the field the paper is in
- What the primary topic of the paper is, as well as the secondary and tertiary topics
- Research Field classifications (e.g., field, subfield)
- Where to best find an Open Access version of the paper
- How many works the paper has referenced
- Which works the paper has referenced, by Name, OpenAlex ID and a URL to a downloadable PDF
- The abstract, provided as an inverted index

3.3 Data Preprocessing

The primary preprocessing step involved reconstructing the abstracts from the “inverted index” format provided by the API into clean, readable text strings suitable for the SPECTER model.

The initial plan for a more complex implementation involved analyzing the full text of PDFs to extract precise citation contexts using tools like GROBID. However, this approach was deemed too unreliable for this prototype, so the decision was made to use abstracts as a simplified proxy for a paper’s content.

3.4 Semantic Analysis and Score Calculation

After comparing models, SPECTER was chosen for the semantic analysis phase (Cohan et al., 2020). An “embedding” is a numerical representation of text in a high-dimensional “vector space”; in this space, the geometric distance between two vectors corresponds to the semantic similarity of the texts they represent. The SPECTER model was used to convert the preprocessed abstract of each paper in our dataset into one of these vectors.

The core of the score calculation involves comparing a citing paper to the papers it cites. For each citing paper in the dataset, its abstract embedding was compared against the abstract embedding of *each paper it referenced.* The cosine similarity was calculated between these vector pairs to produce a numerical value representing their semantic relevance.

A preliminary “Citation Score” for each citing paper was then calculated by averaging these individual similarity scores. This average score represents the overall semantic relevance of a paper’s citations. Finally, these scores were normalized with min–max scaling to $[0, 1]$ using dataset-wide bounds $r' = \frac{r - r_{\min}}{r_{\max} - r_{\min}}$ to produce the final score for analysis.

4 Results of Prototype

First, one can begin to visually compare the described vector embeddings of the papers, to see if there appears to be a clear relationship between a paper and its citations versus other papers and their citations.

Exemplary Visual Vector Representation of two Citing Papers

To visualize the high-dimensional embeddings, their complexity was first reduced to three “principal components” (x, y, and z) using Principal Component Analysis (PCA). It is important to note that PCA is used here strictly for the purpose of dimensionality reduction to enable visualization; the actual score calculation relies on the cosine similarity metric applied to the original, high-dimensional vectors. It is also a question of how meaningful those principal components actually are for the "meaning" within the text.

Then, for a first look, two random papers were taken, and their simplified embeddings were plotted alongside the embeddings of the papers they cite. As Figure 3 shows, there appears to be a clear semantic clustering: a citing paper is visually closer to its own citations than to a random set of other papers.

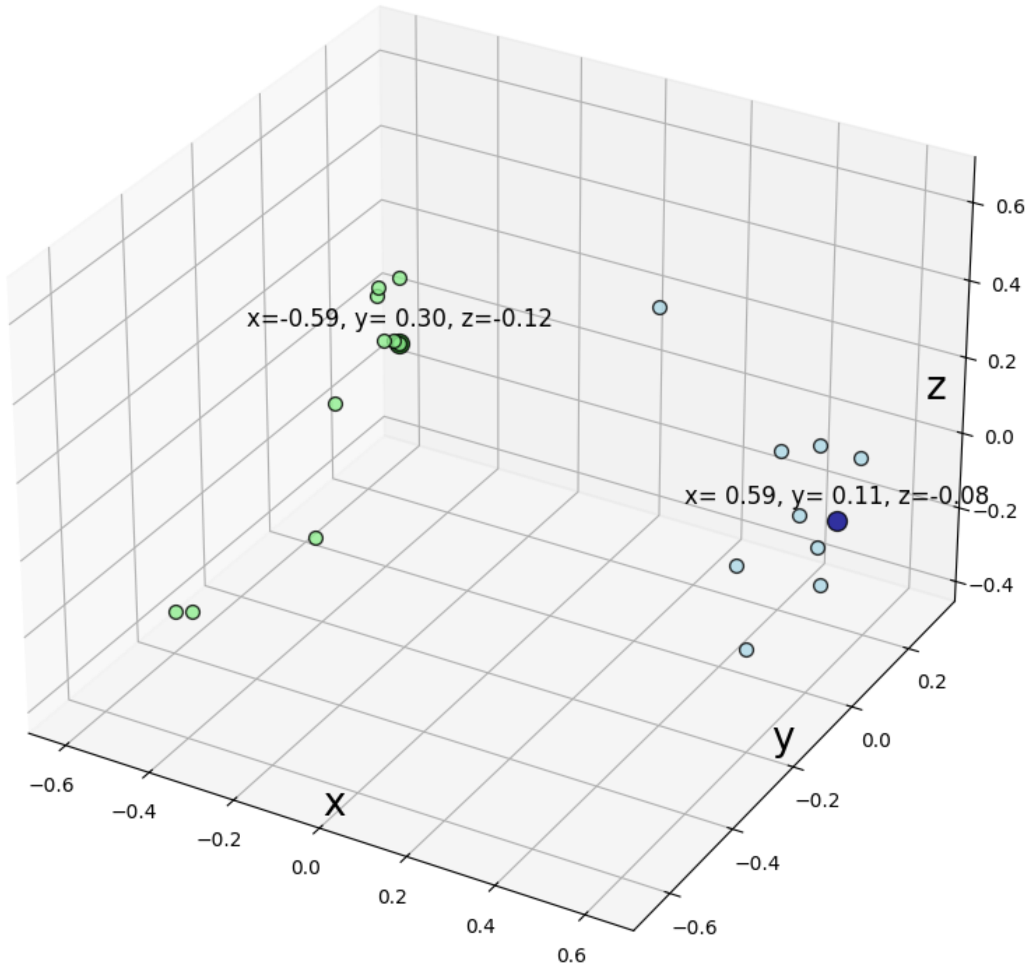


Figure 3: *PCA-Based 3-D Mapping of Paper Embeddings*

Exemplary Visual Vector Representation of several Citing Papers

This same procedure can be repeated with multiple papers at once. Figure 4 plots several citing papers and their respective citations in a single graph, with corresponding papers assigned matching colors. This plot offers a qualitative intuition that papers tend to cluster semantically with the works they reference. While this visualization demonstrates the general principle of semantic proximity, a diagram of scores or a comparison across disciplines, will be a valuable next step to further explore these trends.

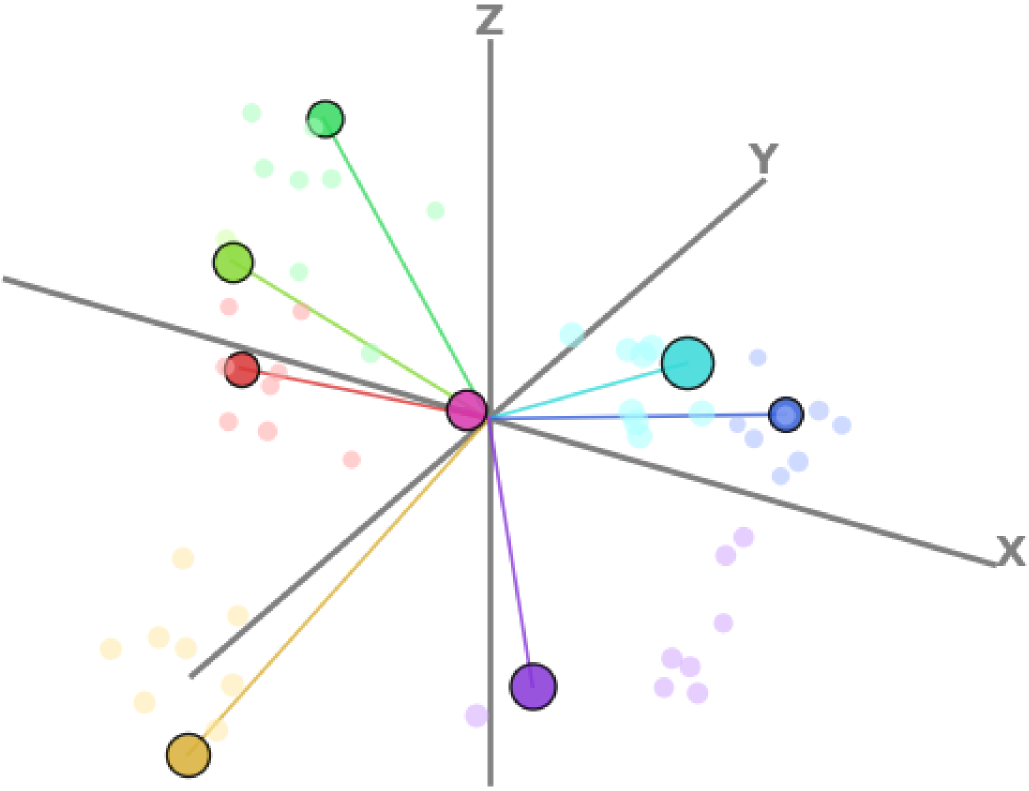


Figure 4: *PCA-Based 3-D Map of Multiple Papers and Their Citations*

Graphical Representation: Citation Score and Citation Count

To validate the prototype, its output must be compared against existing benchmarks. Therefore, a key point of interest is the relationship between the new, semantics-based Citation Score and the traditional, volume-based metric of Citation Count. The Citation Count represents the number of times the paper itself has been cited. The following analysis explores this relationship. The Citation Scores are plotted against the Citation Count, represented as a normalized percentile. A regression line is added to show the general trend.

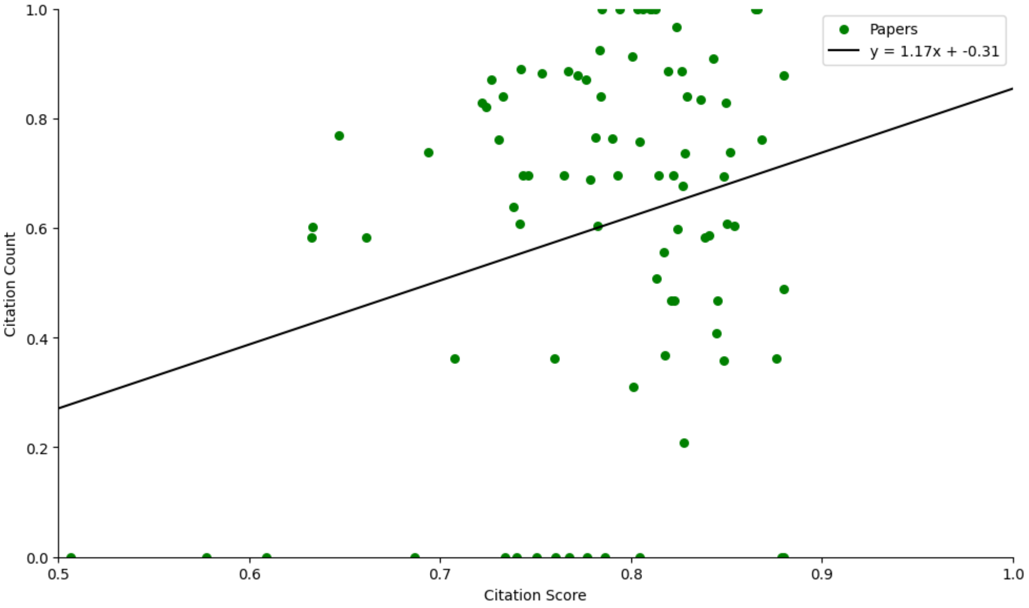
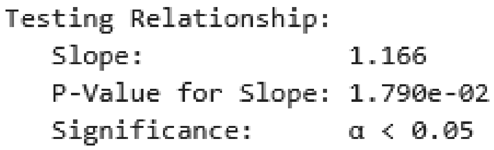


Figure 5: *Scatterplot of Citation Score Percentile Versus Citation Count*

Statistical Significance of Relationship: Citation Score and Citation Count

A simple linear regression analysis shows a positive and statistically significant association between the Citation Score and the percentile of the citation count. The positive slope of the regression line ($\beta = 1.17$) indicates a strong positive association, where higher Citation Scores are correlated with higher citation count percentiles. The p-value ($p = 0.018$) falls below the usual alpha threshold of 0.05, confirming that the correlation is statistically significant.



```
Testing Relationship:
   Slope:             1.166
   P-Value for Slope: 1.790e-02
   Significance:      α < 0.05
```

Figure 6: *Linear Regression of Citation Score on Citation-Count Percentile*

Graphical Representation: Citation Score and Journal Rank

Another option is to plot the Citation Score of each citing paper against the H-Index of the journal in which it was published. A regression line again visually represents the trend in the data.

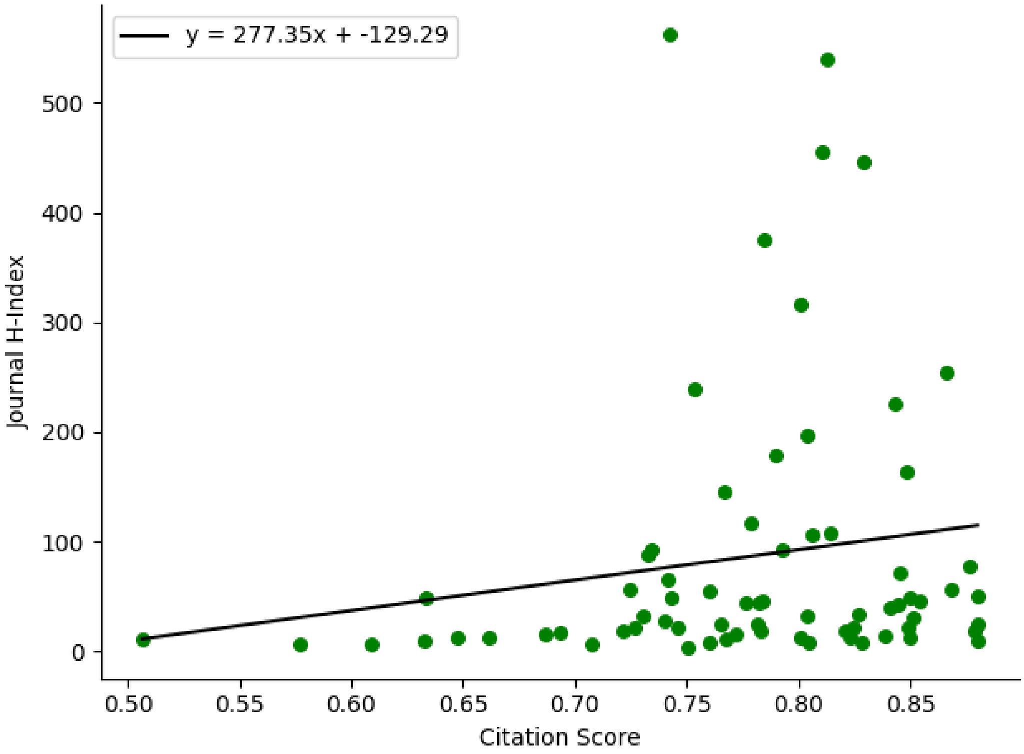
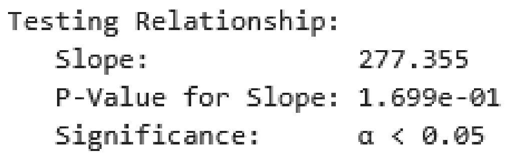


Figure 7: *Scatterplot of Citation Score Versus Journal H-Index*

Statistical Significance of Relationship: Citation Score and Journal Rank

Further analysis of the linear regression shows a positive but statistically non-significant association between the Citation Score and the H-Index of the journal. The p-value of 0.1699 is higher than the usual significance threshold of 0.05. This suggests that while a positive trend exists in the sample data, it is not strong enough to conclude that there is a clear, non-random correlation in the broader population. The observed relationship could be due to random chance, and the analysis does not imply a causal link in either direction.



```
Testing Relationship:
   Slope:              277.355
   P-Value for Slope: 1.699e-01
   Significance:       α < 0.05
```

Figure 8: *Linear Regression of Journal H-Index on Citation Score*

Further Investigation of Score vs. Research (Sub-)Field

We can further investigate our score and relate it to different scientific (sub-)fields. This plot visualizes the distribution of the calculated Citation Scores across the different subfields of Computer Science present in the dataset.

Each box represents the interquartile range (the middle 50%) of scores for a subfield, with the internal line marking the median. The individual dots overlaid on the plot represent single papers, showing the full spread and density of the data, revealing visually which sub-fields may tend to have higher or more varied Citation Scores. It e.g. shows a very high visual variance in the field of Information Systems.

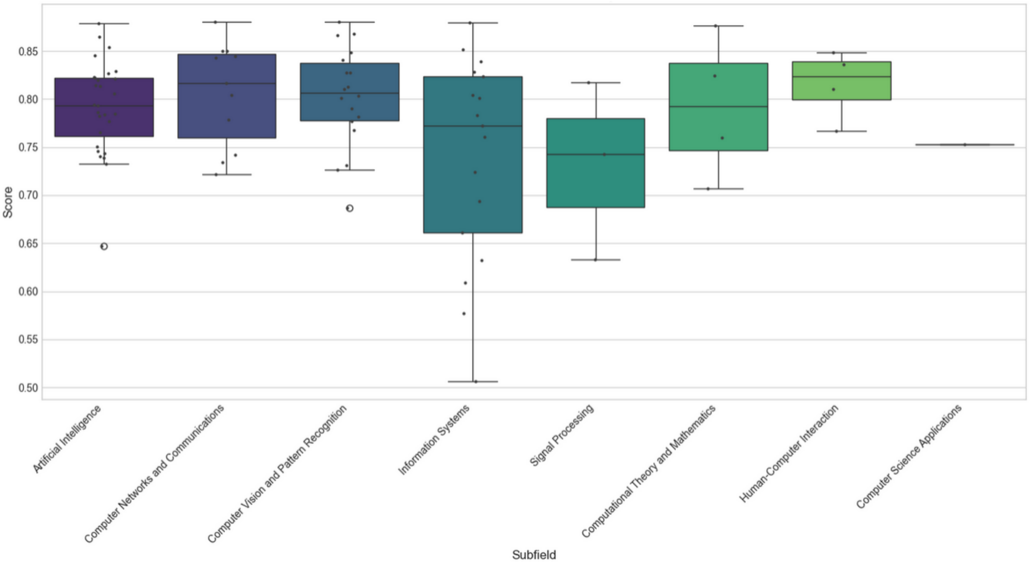


Figure 9: *Boxplot of Citation Score over Research Fields*

5 Discussion

The goal of this analysis was to map the conceptual territory needed to create an automated, context-aware model for evaluating scientific articles and to test these concepts with a simplified prototype using semantic embeddings. A key finding from the prototype is the demonstrated relationship between the new Citation Score and existing metrics. The results show a mixed but promising outcome: there is a statistically significant positive correlation between the semantic-based Citation Score and the citation count; however, the relationship with the journal H-Index did not reach statistical significance. This suggests that the semantic relevance of an article's citations is meaningfully associated with conventional measures of impact, though its connection to a frequent metric of journal prestige was not confirmed in this analysis.

The analysis of literature and the prototype's performance yielded several other insights. It is crucial to have a full, multilayered view of citations, as treating all references as equal is not defensible. An automated metric that ignores the different dimensions of citation motivation risks interpreting noise. Existing taxonomies provide a workable foundation for a future, more complex model that could classify citation roles before aggregating a score. The prototype showed that semantic proximity is a measurable and relevant factor and that using abstracts as a proxy provides a surprisingly strong baseline. This supports the finding that information-dense sections of a paper are important context for its citations.

Limitations

There are several important limitations to these findings, such as:

- Abstracts omit many cues that signal citation intent and are not the ideal citation context.
- The landscape of LLMs is shifting rapidly, so foundational knowledge may need significant adjustment over time.
- Empirical tests comparing alternative classification taxonomies are still scarce.
- The prototype's scope was limited to a single discipline and year, which reduces generalizability.

- SPECTER is not fully up to date, and newer terminology may receive degraded representations.
- The evaluation benchmarks used (citation count and journal rank) are useful but not perfect stand-ins for research "quality." A stronger test would involve expert judgement.

Future Work

For future work, a full-scale, "high-complexity" pipeline could be developed to build upon the findings of the prototype. The following steps outline a potential path for creating such a system.

First, the process would begin by fetching the full text of each paper and parsing it with a robust system like GROBID to recover a clean, machine-readable structure and identify every citation "anchor" within the text.

Next, the analysis would move to a more granular level. The citation context should be identified based on the entire argument the citation makes in the citing paper, ensuring that implicit context is included. For the contents of the cited paper, the context that is closest to the argumentation of the citation context in the citing paper should be used, and this context should be chosen to be cohesive rather than fractured. Then, each citation's motivation could be classified against an optimized taxonomy (e.g., Basis, Comparison, Critique, Use, etc.) before its semantic relatedness is calculated via embeddings.

A further model could then incorporate the syntactic context of each reference, such as its location within the paper and how frequently it is mentioned, combining these features with the semantic analysis to produce a final, normalized Article Score.

Crucially, this system would need a dynamic layer to adjust for evolving terminology, new insights into citation practices, and even potential patterns of manipulation as researchers adapt to new evaluation metrics. Finally, any such system must undergo rigorous empirical testing. A crucial validation step would involve comparing the resulting scores against the qualitative judgments of a panel of expert reviewers who rank the same set of papers.

Although extensive further research would be necessary to fine-tune and expand these suggestions, the prize of adding new, useful, metrics to the widely used citation count is large.

6 References

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), 2158244019829575. https://doi.org/10.1177/2158244019829575

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses. *arXiv preprint.* https://doi.org/10.48550/arXiv.2404.17663

Athar, A., & Teufel, S. (2012, June). Context-enhanced citation sentiment detection. *In Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 597-601).

Beltagy, I., Cohan, A., & Lo, K. (2019). SciBERT: A pretrained language model for scientific text. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615–3620). https://doi.org/10.48550/arXiv.1903.10676

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? *Journal of Documentation*, 64(1), 45–80. https://doi.org/10.1108/00220410810844150

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37(1), 34–36. https://doi.org/10.1002/(SICI)1097-4571(198601)37:1<34::AID-ASI5>3.0.CO

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Document-level representation learning using citation-informed transformers.
arXiv preprint arXiv:2004.07180.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLOS ONE*, 8(3), e58727.

https://doi.org/10.1371/journal.pone.0058727

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(5), 887–900. https://doi.org/10.1002/asi.23256

Erikson, M. G., & Erlandson, P. (2014). A taxonomy of motives to cite. *Social Studies of Science*, 44(4), 625–637. https://doi.org/10.1177/0306312714522871

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4), 359–375. https://doi.org/10.1007/bf02019306

Nigel Gilbert, G. (1977). Referencing as Persuasion. *Social Studies of Science*, 7(1), 113-122. https://doi.org/10.1177/030631277700700112

Hernandez-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22(3), 327–349. https://doi.org/10.1017/S1351324915000388

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2), 291-314. https://doi.org/10.1023/A:1017919924342

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2016). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1), 93–130. https://doi.org/10.1017/S1351324915000443

Lagopoulos, A., & Tsoumakas, G. (2021). Self-citation analysis using sentence embeddings. *arXiv preprint* arXiv:2105.05527. https://doi.org/10.48550/arXiv.2105.05527

Liu, M. (1993). A study of citing motivation of Chinese scientists. *Journal of Information Science*, 19(1), 13–23. https://doi.org/10.1177/016555159301900103

Lopez, P. (2009, September). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *In International conference on theory and practice of digital libraries* (pp. 473-474). Berlin, Heidelberg: Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-04346-$8_6$2

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: A meta-synthesis. *Scientometrics*, 126(4), 3243–3264.
https://doi.org/10.1007/s11192-021-03908-z

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36(3), 435–444.
https://doi.org/10.1007/bf02129604

Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), The sociology of science: Theoretical and empirical investigations (pp. 267–278). *University of Chicago Press.* (Original work published 1942)

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86–92.
https://doi.org/10.1177/030631277500500106

OurResearch. (2025). OpenAlex – a fully open catalog of the global research system. *OpenAlex.* Retrieved August 31, 2025, from https://openalex.org

Swales, J. (1986). Citation analysis and discourse analysis. *Applied linguistics*, 7(1), 39-56.

Szomszor, M., Adams, J., Fry, R., Gebert, C., Pendlebury, D. A., Potter, R. W. K., & Rogers, G. (2021). Interpreting bibliometric data. *Frontiers in Research Metrics and Analytics*, 5, Article 628703.
https://doi.org/10.3389/frma.2020.628703

Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 103-110).

Thelwall, M., & Jiang, X. (2025). Is OpenAlex Suitable for Research Quality

Evaluation and Which Citation Indicator is Best?. *arXiv preprint* arXiv:2502.18427.
https://doi.org/10.48550/arXiv.2502.18427