Quantifying Corporate Culture using Machine Learning and 10-K Filings

Tom Robert Beecken

Dissertation written under the supervision of professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, May 2024.

Abstract

Title: Quantifying Corporate Culture using Machine Learning and 10-K Reports
Author: Tom Robert Beecken
Keywords: Corporate Culture; Machine Learning; Natural Language Processing; Artificial Neural Networks; Word Embedding; Merger and Acquisitions; Cultural Integration

Corporate culture plays a critical role in the success and integration of organizations, particularly during mergers and acquisitions (M&A). This dissertation aims to extract and quantify elements of corporate culture from 10-K textual data and subsequently apply these quantifications to real-world scenarios. To explore this question, a comprehensive dataset consisting of 68,855 Management Discussion and Analysis sections from 10-K Filings stemming from 12,553 companies was utilized. The study employs advanced NLP techniques, including word embedding and sentiment scoring using Term Frequency-Inverse Document Frequency (TF-IDF), to generate a culture dictionary and identify as well as quantify linguistic patterns indicative of corporate culture. The findings demonstrate that NLP-driven quantification of corporate culture can provide valuable insights for addressing cultural integration in M&A scenarios. By scoring the core cultural values innovation, integrity, quality, respect, and teamwork, stakeholders can make more informed decisions, potentially improving the success rates of M&A activities.

Resumo

Título: Quantificação da cultura empresarial utilizando a aprendizagem automática e os relatórios 10-K
Autor: Tom Robert Beecken
Palavras-chave: Cultura empresarial; Aprendizagem automática; Processamento de linguagem natural; Redes neurais artificiais; Incorporação de palavras; Fusões e aquisições; Integração cultural

A cultura corporativa desempenha um papel fundamental no sucesso e na integração das organizações, particularmente durante as fusões e aquisições (M&A). Esta dissertação tem como objetivo extrair e quantificar elementos da cultura empresarial a partir de dados textuais do 10-K e, subsequentemente, aplicar essas quantificações a cenários do mundo real. Para explorar esta questão, foi utilizado um conjunto de dados abrangente que consiste em 68 855 secções de Análise e Discussão da Gestão de Registos 10-K provenientes de 12 553 empresas. O estudo emprega técnicas avançadas de PNL, incluindo a incorporação de palavras e a pontuação de sentimentos utilizando a Frequência de Termo - Frequência Inversa de Documento (TF-IDF), para gerar um dicionário de cultura e identificar, bem como quantificar, padrões linguísticos indicativos da cultura empresarial. Os resultados demonstram que a quantificação da cultura empresarial baseada na PNL pode fornecer informações valiosas para abordar a integração cultural em cenários de M&A. Ao pontuar os valores culturais fundamentais inovação, integridade, qualidade, respeito e trabalho em equipa, as partes interessadas podem tomar decisões mais informadas, melhorando potencialmente as taxas de sucesso das actividades de M&A.

Acknowledgments

This dissertation was undoubtedly one of the hardest tasks I have ever undertaken, thus it only feels natural to end my academic career there (for the time being). I choose the subject to push myself, get outside of my comfort zone, and learn something newall of which I have undoubtedly done. But it is important to acknowledge a few individuals without whom this would not have been possible.

First and foremost, my supervisor Dan Tran, who took the time to hear my wild ideas and tried his best to mentor me so that I could both follow some of those ideas and maintain a realistic perspective. I could not have achieved what I have done without his guidance. Furthermore, I know the advocacy Dan pours into the general topic of Data Science and Machine Learning at Católica, a topic that in my opinion will become more and more relevant. I am very thankful for the opportunity to be part of these courses in conjunction with my Masters in Finance, which would not have been possible to this extent without Dan.

Secondly, I would want to express my gratitude to my friend Ferdinand Muth, who has dedicated a significant amount of time to assist me with the coding, an extremely difficult effort and definitely an area he knows better than I do.

Thirdly, I want to thank my friends and family for their incredibly patient support during this long and difficult journey.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Contextual Analysis

The notion of culture within the corporate context presents a paradox of being immensely influential yet elusive in its concrete assessment and quantification. Whilst central to an organization's identity and operational ethos, it remains a malleable concept, that is challenging to measure with any meaningful precision. This thesis explores the quantification of corporate culture, transforming an abstract concept into comparable and generalizable information. This endeavor is not merely academic; the practical applications of such quantification are vast. To demonstrate this, simultaneously to demonstrating the appropriate methods, this dissertation lays focus on providing practical explanations on how the generated data can be useful in real world scenarios. The focus here lies on the dynamic industry of mergers and acquisitions (M&A), a field where the alignment and appropriate integration of corporate culture bears strong significant on outcomes.

The challenge in quantifying culture has several complications. Firstly, corporate culture encapsulates the ethos, values, and practices that pervade an organization, shaping its interactions both internally and with the external market environment. Despite its significance, corporate culture is inherently qualitative, often described in terms that resist straightforward measurement. Furthermore, culture often has multiple dimentions; it is both a deeply internalized character and a pattern of behaviors that can be observed externally. Companies themselves are often reluctant to externalize or quantify these characteristics, viewing them as intrinsic and unique competitive advantages that are not easily, or advisably, shared. Because of this, the present methods of determining corporate culture might not be the most enclosing path to unveiling a company's true culture. The most accepted method of understanding a companys culture from an external perspective involves analyzing data from resources like the KDL Database, which compiles culture scores based on extensive interviews and questionnaires conducted with the companies. Another issue these methods entail these methods is that they are labor-intensive and not easily scalable across the broader market landscape. Furthermore, they do not always allow for direct comparisons across different organizations, which is a crucial capability in scenarios such as mergers and acquisitions.

The M&A process, it exemplifies a complex landscape where the understanding of cor-

porate culture can play a decisive role. Mergers and acquisitions are intricate transactions characterized by a series of detailed, compartmentalized practices that, despite their interconnection, often fail to achieve synergy. Statistically, the failure rate of M&A transactions hovers around 50%, a figure that underscores the inherent risks involved (Renneboog and Vansteenkiste, 2019). Among the numerous factors contributing to these failures, lack of cultural integration is frequently cited as a significant yet overlooked element. Predicting the success of a merger or acquisition is thus not only about financial and operational alignments but also about understanding the cultural fit or the lack thereof between the entities involved.

1.2 Objectives and Structure

This thesis aims to bridge the gap between the qualitative nature of corporate culture and the quantitative needs of corporate analysis, specifically focusing on M&A as an example. The primary objective is to develop a methodology for the meaningful classification and quantification of corporate culture among listed companies using publicly available data, specifically through the Management Discussion and Analysis (MD&A) section of 10-K filings, the execution of which can be seen at https://shorturl.at/ZVf0t. These documents are rich with linguistic cues and narratives that, when processed through Natural Language Processing (NLP) techniques, can reveal underlying cultural dimensions. Furthermore, the research will discuss how these quantified cultural dimensions can be applied in real-world contexts.

Central to this thesis is the question: How can Natural Language Processing techniques be employed to extract and quantify elements of corporate culture from 10-K textual data, and how can these quantifications be useful in real-world scenarios, particularly in mergers and acquisitions? By addressing this question, the research not only contributes to the academic field of corporate culture study but also offers tangible tools and frameworks that can assist practitioners in navigating the complex dynamics of M&A. The approach taken here is designed not only to explore theoretical aspects of corporate culture but to forge a direct link between theory and practice. Through the lens of M&A, this research aims to show how a deep understanding of corporate culture, grounded in robust quantitative analysis, can lead to more informed decisions and better outcomes in one of the business world's most challenging arenas.

2 Literature Review

To effectively navigate this journey, a contextual framework that illustrates the value of establishing a measurable indicator for corporate culture will be presented. This analysis serves not only as an academic pursuit but also as a practical guide for professionals, in the field of mergers and acquisitions, among others. By providing insights into the quantification of corporate culture, light is shed on the potential for more successful, synergistic, and sustainable mergers.

In order to guarantee that the discoveries presented remain pertinent and up-to-date only literature sourced from highly regarded academic journals will be utilised.

2.1 Corporate Culture Background

2.1.1 What is Corporate Culture?

The scholarly literature on this subject is minimal, which is most likely due to the ineffability of the subject matter, despite the fact that this topic is extremely important. Nevertheless, multiple formal definitions exist. Crémer, 1993 for instance describes culture as a body of knowledge that is shared by the members of an organization. This dynamic provides a common language and sets precedents for behavioral rules. Kreps, 1990 on the other hand describes culture as a mechanism to foster collaboration and managing unforeseen circumstances. Furthermore, many authors describe corporate culture as "a system of shared values and norms that define appropriate attitudes and behaviors for organizational members" (O'Reilly and Chatman, 1996). Unlike national culture, corporate culture is exceedingly malleable by top management and is often shaped with intent (Graham et al., 2022). Most employees value the way they are perceived by their colleagues, forcing them to act in line with the fashioned culture existing within their organization, especially in presence of other employees. This makes corporate culture a social mechanism of control that works through peer influence (Kreps, 1990; O'Reilly and Chatman, 1996). This relationship allows top management to employ the customization of their organizations culture as a complementary tool that supports control systems (e.g. incentives), which finally has a positive effect on organizational effectiveness and efficiency. For the purposes of this study, the concept will be dealt with according to the definition set by O'Reilly and Chatman, 1996, describing culture as a system of shared values and norms. Previous studies that have examined the measurement

of corporate culture have employed the same concept (Li et al., 2021).

2.1.2 Corporate Culture is Hard to Measure

The vague nature of corporate culture has been mentioned before. Since there are no defined parameters in which a companys culture exists, it can be different for each organization. For this reason, it is hard to define, and hard to measure. Furthermore, corporate culture is a very personal topic for many companies, meaning the outside impression of culture is mostly not in line with what happens behind the scenes. This means that the actual outcome of an analysis is always reliant on the truthfulness of the data collected. When culture is determined based on interviews or surveys, which is often the case as will become evident, all information one hopes to receive based on the culture, is contingent on the accuracy of the interviewees understanding of the culture as well as their willingness to disclose that information.

Graham et al., 2022 is one of the most cited academic papers in this field, hence their methodology, which is representative of common methods, will be outlined to demonstrate the conventional way of measuring corporate culture.

Most literature, as well as Graham et al., 2022 makes use of survey-based data, in which the responses are hand-coded to predefined categorizations resembling elements of culture. They start their process by interviewing 18 executives of private as well as public companies, both in early and late life cycle stages to learn about cultural variety. The knowledge gained in these interviews is then used to design the survey instrument. The survey was then sent out to a long list of CEO's and CFO's curated from the Fuqua School of Business at Duke University and the Columbia Business School. This action yielded 1898 responses, which equates to a high response rate of 13.4%. From this sample several responses were eliminated due to companies being outside of the US and Canada, government organizations, non-profit status and general mistakes made filling out the survey. After filtering they were left with 1348 usable responses. Next these responses are hand-coded to the predefined categorizations resembling elements of culture mentioned before. The categories proposed by Graham et al., 2022 are:

1. **Adaptability**: willing to experiment, fast-moving, quick to take advantage of opportunities, taking initiative

2. **Collaboration**: team-oriented, supportive, not aggressive, low levels of conflict
3. **Community**: respectful of diversity, community, and the environment, inclusive, caring, and open
4. **Customer-orientation**: listening to customers, being brand driven, taking pride in service.
5. **Detail-orientation**: paying attention to detail, being precise, emphasizing quality and safety, being analytical.
6. **Integrity**: high ethical standards, being honest, trans- parent Results-orientation: high expectations for performance, focus on achievement, competitive, demanding.

To ensure non-biased encoding, Graham et al., 2022 utilized five researchers (one from their team, and four independent researchers), all of which independently encoded the responses. For a coding to be valid, three of these five researchers had to agree upon the coding.

Surveys bring several advantages when researching culture, especially when compared to other conventional methods. Firstly, they can be rolled out on a large scale, and secondly, they can be structured to extract detailed and precise information tailored to the research question that is attempted to be answered. Graham et al., 2022 for instance go to top management, who are a strong provider of answers in this case, since as previously established, corporate culture is influenced by top management. One major drawback of this methodology however, especially when looking form a far smaller scale, is difficulty of reaching members of top management, particularly when information on large organizations is desired. This problematic is addressed in the next section, which outlines the established applications of Machine Learning approaches to measure culture.

2.2 Machine Learning applications to Measure Culture

Corporate Culture is a highly specialized subject matter that very hard to define and highly specific for each use case. Just as two people are never identical regarding personalities, a similar assumption can be made for two companies never being an exact clone of each other. Furthermore, companies have incentive to keep their own culture private, presenting another obstacle when trying to identify corporate culture using public information (Graham et al., 2022). Whilst it is possible to identify culture using manual approaches such as interviews, these methods are not easily accessible and require a detailed analysis of the responses to yield useful outcomes. For this reason, it is practical to use Machine Learning to

identify corporate culture. If it is possible to come up with a robust way for a machine learning model to comprehend and classify natural language, it may yield promising inferences about corporate culture.

The crossroads of machine learning applications and corporate culture classification is quite unexplored regarding academic research. Li et al., 2021 developed a novel way to quantify and analyze corporate culture using machine learning. They use and adjust previous approaches of textual analysis to cultivate their process. Quantifying words in to vectors as introduced by Mikolov et al., 2013 and the linguistic notion that similar words in meaning share common neighboring words proven by Harris, 1954, were utilized to generate a culture specific dictionary. This collection of word in turn was used in a similar way to E. Henry, 2008's solution to measuring tone and sentiment in texts by counting specific words, with further adjustments to make the model perform as intended. Their process proves to reliably identify culture from earnings calls transcripts, by scoring five cultural values based on the number of times certain words are mentioned in the call. Their method provides an excellent basis for this dissertation to introduce a way to recognize cultural norms from publicly available documents such as the 10-K filing.

2.3 M&A application for Corporate Culture

In the evolving landscape of corporate strategy, mergers and acquisitions have emerged as crucial proceedings for organizations seeking to enhance their competitive edge, expand their market footprint, and harness synergies. The importance of culture in an M&A context for managers is strongly evident from Graham et al., 2022's research, which states that over 54% of executives would abandon an M&A deal due to cultural misalignment. Furthermore, another 33% of the managers value a misalignment of culture as a 10%-30% decrease in value. The M&A process is a complex, multifaceted, and very detailed series of compartmentalized practices, all interconnected, yet far from efficiently harmonious. All however work towards a common goal, ensuring the positive outcome of the transaction. Despite these efforts, the failure rate of M&A transactions averages at around 50% (Renneboog and Vansteenkiste, 2019). Why is that?

The post-merger integration process is an essential component in recognizing the potential advantages of a transaction. Evidence suggests that the post-acquisition phase is more curtail to the success of the transaction than the pre-acquisition phase due to it's importance

in realizing potential value through cultural and organizational integration Haspeslagh and Jemison, 1991. Financial- and legal issues along with anything else that is handled in the pre-acquisition phase is typically handled exceptionally. However, managers as well as consulting agents often neglect the thorough planning of how integration will take place. A common assumption is that M&A advisors undersell the significance of post-acquisition events to their clients since their fee is contingent upon the transfer of ownership.

Intuitively one might assume that to be successful the participating entities need to be fully integrated to be successful, but Weber and Schweiger, 1992 find that over-integration can lead to clash of cultures, which hinders synergy effects. Nevertheless, The consolidated organization will need adjusted operating as well as management pipelines to efficiently merge into one (Inkpen et al., 2000). Here a clear dilemma concerning integration strategy becomes evident, it is posed in the trade-off between managerial attention, knowledge-loss through employee turnover, disruption of functioning organizational routines and the synergy effect resulting from these changes (Puranam et al., 2006; Ranft, 2006; Spedale et al., 2007). Hence, integration of cultures requires careful and thorough preparation and implementation. Mismanagement of this integrate topic may serve as an explanatory factor for high failure rates (Renneboog and Vansteenkiste, 2019).

To summarize, the issue does not necessarily lie in the cultural proximity, but rather in the way companies aim to integrate their cultures. Identical cultures don't automatically yield the most successful mergers, and for some purposes there might be a benefit in deviating cultures. In the end, the integration strategy should closely reflect the transactions purpose, for instance, carveouts should be handled differently than spin-offs, or even full out mergers. This suggests that the importance of preplanning integration using clear strategies cannot be overstated. By planning and aligning integration efforts with the overarching goals of the merger, organizations can navigate the precarious balance between maintaining operational continuity and achieving the desired synergistic benefits, ensuring a smoother transition and greater likelihood for integration success.

This evidence should clearley outline the goal of this dissertation, to identefy a way to assess corporate culture. Due to temporal and monetary constraints, top management usually looses sight of the importance of a companies day-today activities, which is often managed by a sense of culture (Vermeulen and Barkema, 2001). The outcomes of this thesis yield a easy to use analysis tool, which can be used to adequately plan for the integration of two

independent entities. This adresses an issue that is likely to be tha cause for the failiure of over half of the outcome of a $3.1 trillion industry (J. Henry and Van Oostende, 2023).

2.3.1 Real World Use Case

To demonstrate the practicality of the information this dissertation is attempting to provide, real-world mergers that had increased integration difficulties due to cultural integration will be discussed in the following section. As discussed, cultural integration can be difficult due to a several reasons. Deviating national cultures for instance lead to the failure of the merger between Daimler (a German company) and Chrysler (a US company). Major issues within the combined company revolved around the conflicting levels of formality, pay and expense philosophies as well as overall operating styles. Soon, the German culture became dominant, which led to plummeting employee satisfaction on the Chrysler side. Major losses were recorded by 2000, laying off started, and by 2007 Daimler pulled the plug by selling Chrysler to a Financial Investor (Wearden, 2007). A more diplomatic approach could have been implemented if top management had a clearer image of the entity's cultures, and what possible clashes could be. Unfortunately, this information gap led to a full out failure.

A more promising example lies in the merger between Hewlett Packard and Compaq in 2001. At the time of the merger, both HP and Compaq were struggling IT companies, hovering near 52-week lows. The merger rationale was described as a "merger of equals" in which both sides brought different strengths to the table, in theory resulting in a complementary set of products and services with a more favorable ability to serve customers at lower costs. HP claimed the merger would lead to a reduction in expenses amounting to $2.5 billion by mid-2004 (Burke, n.d.). The public had different opinion, stating that the companies were too different regarding culture. HPs orderly approach with a strong emphasis on organized plans for their work, in combination to Compaq's impulsive approach depending heavily on their capability to respond to just-in-time opportunities did not seem to provide a sustainable basis for synergies to be realized. This was met with a 18.7% price drop on HPs stocks, the day after the merger was publicly announced (Burke, n.d.). The speed clash between a process-heavy and deliberative corporate culture and a pressing, reactive corporate culture emerged as a significant and foreseeable area of failure. At first, some of the publics fears were realized. For instance the combined company struggled to identefy a fitting channel strategy, which hurt partners of the firm. Originating from the highly differentiated cultures,

these kinds of inefficiencies led to an estimated loss of $13 billion in market capitalization. This however did not last forever, as evident in Tommy Wald's (CEO of White Glove Technologies HP Partner) statement, who had some issues with the implementation of a direct sales approach; "Eventually HP pulled back from that direct sales push and really listened to partners. To their credit, the folks at HP continued to invest in the channel and in the relationships with partners. That's when we saw the real potential" (Wright, 2011). Looking back today solution providers state the merger was of historic significante, which after a rocky integration phase, ultimately delivered on it's promises a internationally active technology powerhouse with top revenue positions. A clear cunsensus also states that managers may have been able to prevented or mitigated the consequences of disagreeing cultures by accurately predicting it (Wright, 2011).

All this goes to show is that cultural misalignment doesn't necessarily doom a merger for failiure. If the right approach is taken, and both companoies are taken into consideration, it is possible to merge even the most polar opposite cultures. Given this example, and the evidence it conveys, this dissertation aims to provide the information necessary to clearly plan cultural integration, which in turn gives the basis to prevent/minimize losses such as the described $13 billion market cap loss in the future.

3 Data

3.1 Why this Data?

Graham et al., 2022 as well as Guiso et al., 2015 explain that the top management of a corporation has the most significant influence on its current culture. This suggests that information influenced by top management will be able to give an outsider the best estimate of corporate culture. In previous studies various methods have been used, ranging from earnings calls over top management interviews to 10-K reports (Fang et al., 2023; Graham et al., 2022; Li et al., 2021).

The most common and intuitive source of information, as previously mentioned, are interviews with top management. However, these are hard to obtain, and even when such an interview takes place, the answers might be tailored to promote the companies culture as something it is not, given the interviewees awareness of the objective. Additionally, the purpose of this dissertation is to demonstrate an accessible way to determine corporate culture, which would not be achieved through this kind of data collection. Furthermore, some previous literature has used earnings call transcripts to analyze corporate culture, the reasoning being that the main topic of earnings calls lies on business operations and performance rather than on culture. Two issues arise when contemplating earnings calls as a basis for this dissertation. Firstly, the availability of such transcripts is limited, and is usually bound to a subscription such as Thomson Reuters Street Events via WRDS, which again collides with the idea of providing an accessible solution to decypher culture. Secondly, whilst earnings calls are a great source of information, it is not as wide-ranging as other releases made by listed companies.

The 10-K filing is the most extensive and complete source of information that is published annually. It is commonly accessed by a diverse group of stakeholders who closely analyze its contents and respond accordingly (Griffin, 2003). This being the case, companies try to accurately describe the company as a whole to outside investors, whilst other communication such as earnings calls cater more to the existing shareholders and therefore focus mainly on operational performance. Factors regarding value-relevant business relating to products and services, strategies, market segments and organizational structures are likely mentioned and discussed in detail (Fang et al., 2023). Another factor that makes 10-K filings an interesting basis to understand culture lies in the verbal element. Academic literature

provides evidence that the textual content and tone of 10-Ks capture curtail information including capital constraints (Bodnaruk et al., 2015; Hoberg and Maksimovic, 2015), degree of competition (Li et al., 2013) and product life cycles (Hoberg and Maksimovic, 2022). These findings suggest that written text from 10-Ks is unlikely to be irrelevant when it comes to capturing value-relevant information.

10-K filings contain various sections describing Business Operations, Products, Financials, Risk Factors and many more. Keeping in mind that culture is heavily influenced by top management, to analyze corporate culture the textual data from Item 7. Managements Discussion and Analysis of Financial Condition and Results of Operations (MD&A) of the 10-K statements is used in this research. The main purpose of the MD&A section is to help the shareholder understand the results of operations and financial condition of the company in question. Given the previous discussion, 10-Ks hold curtail information within the textual linguistic and tone, this is the most likely place to find clear implication to unearth the organizations culture. Furthermore, analyzing company culture through 10-K reports is advantageous due to easy accessibility via the SEC EDGAR system. A detailed walkthrough of how the data was collected can be found in the Data Retrieval section.

3.2 Data Retrieval

10-K reports were obtained using the United States Securities and Exchange Commission Electronic Data Gathering, Analysis, and Retrieval (SEC EDGAR) system in combination with the SEC EDGAR API. EDGAR is the primary system for listed companies to submit documents required by the SEC. Containing millions of company filings, EDGAR "processes about 3,000 filings per day, serves up 3,000 terabytes of data to the public annually, and accommodates 40,000 new filers per year on average" ("SEC.gov", 2023).

Since it is impractical to download every filing one by one, the process of data collection was facilitated through web scraping by using the python library Beautiful Soup 4 (BS), "a Python library for pulling data out of HTML and XML files. It works with different parsers to provide idiomatic ways of navigating, searching, and modifying the parse tree" ("Beautiful Soup Documentation", 2023).

To retrieve the described datapoints, web links to the corresponding documents are necessary. The SEC provides filings in document as well as in HTML form (see Appendix for different presentation formats). For scraping, links navigating towards the HTML document

will be generated, as the HTML versions have proven to work better with BS. Fortunately, the required links follow a generic format that can be populated with information accessible through the SEC API. Information needed is outlined in Table 1, with an example of the data in Table 2.

Name	**Description**	**Format**
Accession Number	The accession number is a unique identifier assigned automatically to an accepted submission by EDGAR. The first set of numbers (0001193125) is the CIK of the entity submitting the filing. This could be the company or a third-party filer agent. Some filer agents without a regulatory requirement to make disclosure filings with the SEC have a CIK but no searchable presence in the public EDGAR database. The next two numbers (15) represent the year. The last series of numbers represent a sequential count of submitted filings from that CIK. The count is usually, but not always, reset to zero at the start of each calendar year ("SEC.gov", 2023).	XXXXXXXXXX-XX-XXXXXX
Central Index Key (CIK)	The CIK is a unique identifier assigned by the SEC to all companies and people who file disclosures with the SEC("SEC.gov", 2023).	XXXXXX

Table 1: Explanation of Variables needed to generate links to 10-K Filings

CompName	**Ticker**	**accessionNumber**	**CIK**	**reportDate**
MICROSOFT CORP	msft	0000950170-23-035122	320193	2023-06-30
MICROSOFT CORP	msft	0001564590-22-026876	320193	2022-06-30
MICROSOFT CORP	msft	0001564590-21-039151	320193	2021-06-30
...	...	...	...	...

Table 2: Data needed to generate links to 10-K Filings

These identifying components are used to generate lings in the following format

https://www.sec.gov/Archives/edgar/data/[cik]/[accessionNumber_nodash]/[accessionNumber].txt

an example would look as follows:

https://www.sec.gov/Archives/edgar/data/320193/000032019323000106/0000320193-23-000106.txt

To get the correct information from the SEC API, company tickers need to be matched with CIK numbers through a database available from the SEC. For this study, 12,553 tickers were curated through the Compustat database. This constitutes all companies that were at any time present on the AMEX, NYSE and/or NADSAQ exchanges in a period between January 2014 and December 2024. From these tickers it was possible to generate 68,855 links. This number is considerably larger than the number of tickers, since multiple filings were pulled per company.

The next step entails parsing the document and extracting the MD&A section. This proved to be a challenge, whilst there is a standardized format for these filings, it is noticeable that many companies make it a mission to distinguish themselves through format and/or structure. These stylistic and sometimes even technical adjustments make it difficult to program a generalizable parsing algorithm to find the desired section with 100% accuracy. Due to this issue, as well as some duplicate links, of the 68,855 links generated only 42,130 were successfully processed and used to score corporate culture. If it is desired to find the data used without utilizing web-scraping, link generation, or any other form of coding, all 10-K filings can be found through the EDGAR full text search engine (https://www.sec.gov/edgar/search/).

4 Methodology

4.1 ML Applications How can culture be determined using ML?

To quantify corporate culture, a series of steps need to be followed. This process is outlined in Figure 1 below. Simply but, the goal is to use a dictionary of words to analyze the MD&A section of a company's 10-K, and determine a culture score based on weighted word counts in relation to said dictionary. As will be explained in greater detail, it is impractical to use an existing, pre-defined dictionary, since this dissertation is dealing with a very specialized subject matter Loughran and Mcdonald, 2016. Hence, a custom word dictionary will be generated through a series of Machine Learning Techniques.

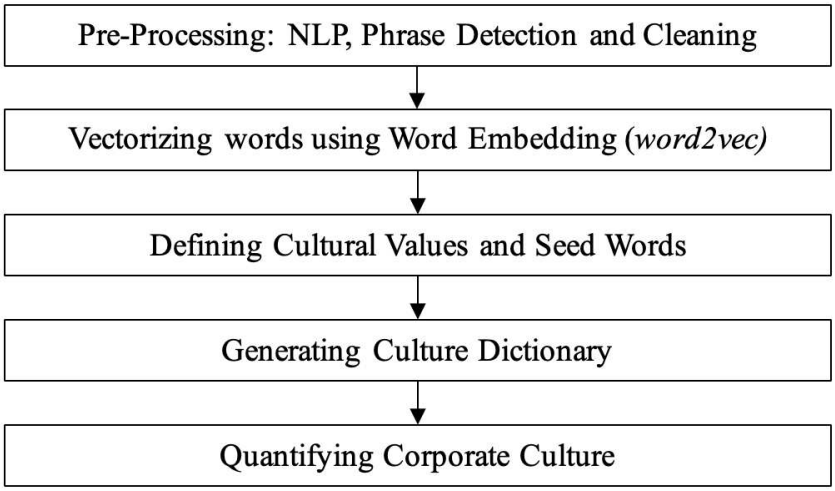


Figure 1: Outine of how to Quantify Corporate Culture

4.2 Pre-Processing: Natural Language Processing, Phrase Detection and Cleaning

As described in the previous section, the MD&A sections were obtained through the SEC EDGAR portal through the utilization of web scraping techniques. Each file is matched to its filing id (Accession Number) as well as the filing company and the filing period. The content of each file consists of the pure textual data that can be found in the filing, any tables, images, or formulas containing financial data have been filtered out in the data collection process.

The first step of preprocessing is done with a Natural Language Processing (NLP) toolkit called Stanford CoreNLP. First established by Manning et al., 2014, CoreNLP is an open-source NLP package, capable of running a variety of tasks associated to NLP. For this dis-

sertation the CoreNLP toolkit was imported through the stanza library, which allows the use of the latest CoreNLP client in the native Python language.

The NLP modules utilized for this dissertation are sentence segmentation and tokenization, lemmatization, name entity recognition and dependency parsing, the function of which will be described in this passage, and can be reviewed in Figure 2.

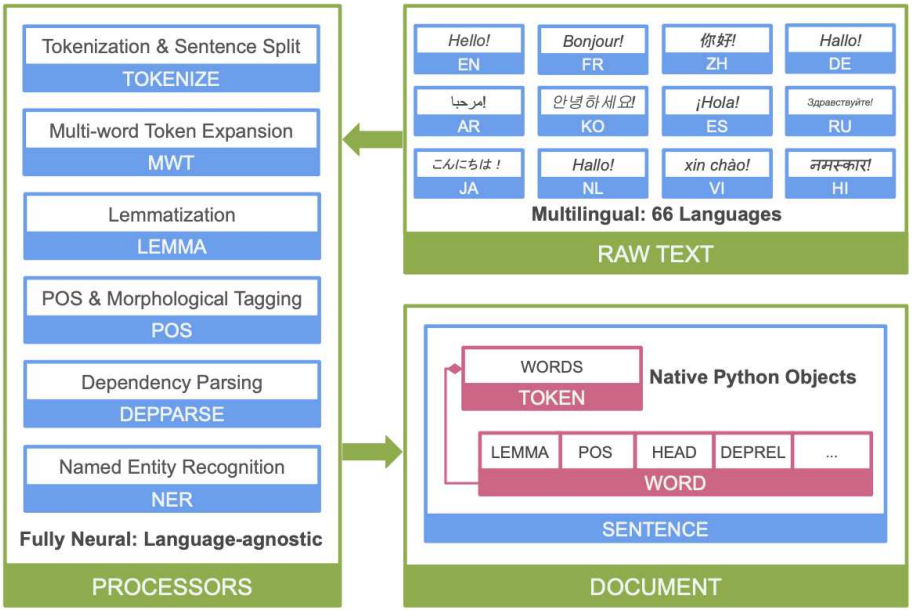


Figure 2: CoreNLP Toolkit Outline (Manning et al., 2014)

Since *word2vec* operates at the sentence level, the text is segmented into sentences. Within each segment, the words are tokenized. Tokenization refers to the procedure of dividing text into individual tokens. For example, the line *Marie was born in Paris.* might be tokenized into the following list: "Marie", "was", "born", "in", "Paris", "." (Manning et al., 2014). Next, each word is individually lemmatized, meaning they are returned to their base forms. The base form for the word was for example is be. This is done to minimize the number of words to be analyzed, as well as prevent the duplicate generation of vectors for words with similar meaning through *word2vec*. Furthermore, named entities such as person and company names but also numerical entities such as dates and times need to be handled. Named entities don't hold cultural value within the name, so they will be mapped to a predefined tag. For example, "We repurchased 71.7 million Apple shares" is transformed to "we repurchase [NER:NUMBER] [NER:ORGANIZATION] share". Multiword named entities, such as Wells Fargo, are also recognized (Li et al., 2021). Finally the texts are run through the dependency parsing module. This type of parser learns grammatical relationships. Figure 3 shows parsed dependency relationship for scentence *So we usually get a build up and then very strong sales*

during that period.

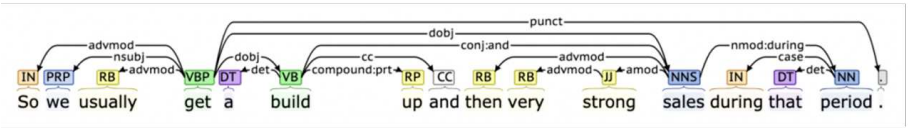


Figure 3: Visualisation of Dependency Parsing (Manning et al., 2014)

CoreNLPs dependency parser can identify multi-word expressions (eg. break the ice and cutting-edge technology), as well as compounds (eg. profit margin and intellectual property). These will be concatenated using _ and treated as single words (break the ice to break_the_ice).

Additionally, to CoreNLP, the phraser module of the gensim library was used to learn phrases that are specific to the data used. The module was employed to detect clusters of phrases that exhibit notable frequent co-occurrences resulting from real-world customs and events, rather than linguistic rules. (Sag et al., 2002). The importance to learn these data-specific phrases will be discussed in a later section. Mikolov et al., 2013 introduced a learning algorithm that can be used to learn two- and three-word phrases. This algorithm scores pairs of consecutive words using the following Formula.

$$score(w_i, w_j) = \frac{(count(w_i w_j) - \delta) * |V|}{count(w_i) * count(w_j}$$

Here, δ represents the minimum frequency for the phrase to be considered (this dissertation uses δ = 50), w_i and w_j represent the consecutive words, and $|V|$ indicates the size of vocabulary (Li et al., 2021). The subjects are considered a phrase if the score assigned by Formula 1 is greater than 10, in which case they are treated the same as multi-word phrases. Some phrases identified through this process are: private equity fund and forward-looking statement.

Finally, the parsed text is cleaned by removing punctuation, stop words and single-letter words. Loughran and Mcdonald, 2011 have extensively been used to measure tone and sentiment, they also provide a list of generic stop words including 121 appropriate words such as and, the, and of. Stop words are generic, and don't pass information to the context of the text analyzed, hence, they are removed (Loughran and McDonald, 2024). They are removed at the end since they could potentially be part of multi-word expressions.

This concludes the pre-processing phase. All steps are necessary to maximize efficiency

and minimize processing time when running the full model, as will be explained in the next sections.

4.3 Vectorizing words using Word Embedding (*word2vec*)

4.3.1 Word Embedding?

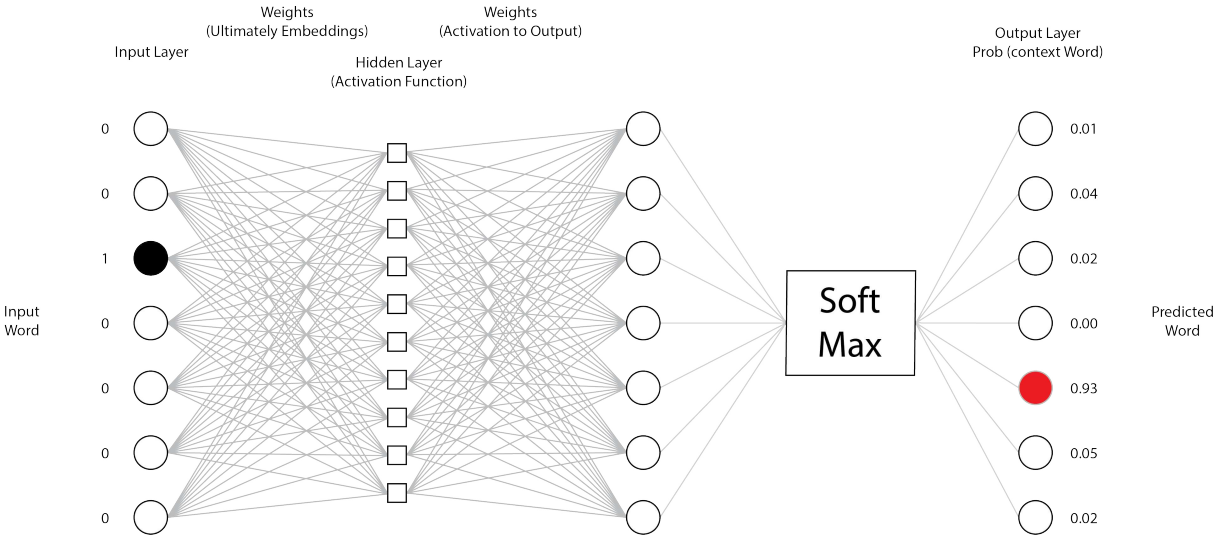


Figure 4: Traditional Neural Network used for Word Embedding

Word embedding is a technique used in natural language processing (NLP) to represent words as vectors. This method is employed to transform words into numerical forms that can be processed by machine learning models. Word embeddings capture semantic relationships between words, allowing algorithms to understand and interpret text more effectively. Specifically, vectorized neatural language enable the determination of relationships between words, usually using simple vector arithmetic. The core principle behind word embedding is based on a linguistic concept that words appearing in similar contexts tend to have similar meanings. Word embeddings work by training on large text corpora. During training, the model learns to map words to vectors such that the distances between vectors reflect semantic similarities. This is achieved through methods like neural networks that optimize word vector representations based on the context in which words appear. Essentially, the embedding process involves learning from word co-occurrence patterns, resulting in vectors that encode both syntactic and semantic information. The most basic process to embed words is to create a count vector that records the number of times other words appear close to the focal word within a specified distance in the text corpus. Using this method to construct vec-

tors for each word in the corpus, it is hypothetically possible to find the association between any two words by calculating the cosine similarity.

As an example, if the relationship between the word's innovation, collaboration and leadership needed to be determeined neighboring word count can be used. This yields that creativity, vision and success often appear next to innovation and collaboration, and strategy and vision often appear near leadership. The vector representation can be seen in Table 3.

	Neighboring Word Count				
	creativity	*teamwork*	*strategy*	*vision*	*success*
innovation	5	4	1	0	6
advancement	4	5	0	1	5
leadership	1	0	8	9	0

Table 3: Example on Terms and Neighboring Word Counts

If the cosine similarity is calculated for each of these vectors, a similarity of 0.968 between innovation and advancement, a similarity of 0.122 between innovation and leadership and a similarity of 0.131 between advancement and leadership is determined. The higher the cosine similarity, the closer the association between the two words.

This example demonstrates how word embeddings are done, but also unearths some major shortcomings of this method of doing it. One limitation is the requirement for extensive computational resources to train effective embeddings. This example only shows five constituents, in practice however the number of combinations of all words in a corpus is colossal, thus the implementation of the basic count-based word embedding method becomes challenging, necessitating the adoption of an alternative strategy.

4.3.2 word2vec

word2vec (Mikolov et al., 2013) is an algorithm that significantly decreases computational resources while effectively producing word embeddings. Levy and Goldberg, 2014 illustrate the similarity between word2vec's vectorization and a singular value decomposition of the matrix that counts neighboring words. The fundamental approach of the algorithm involves learning semantic understanding by sequentially processing textual content and learning to predict surrounding words in each iteration. The model utilizes a conventional neural network that is trained via backpropagation to obtain optimum parameters. After training, these parameters, specifically the weights connecting the input layer to the hidden layer (Figure 5), then serve as precise vector representations of the word being embedded. The vector

representation of a word captures its meaning and context by encoding the features of its co-occurrence connection with neighboring words (Li et al., 2021). *word2vec* predominantly makes use of two tactics to create word embeddings. The continuous bag of words model enhances contextual understanding by leveraging neighboring words to anticipate the content in the middle. Skip-gram on the other hand enhances contextual understanding by utilizing the central word to forecast the adjacent words. The drawback of both these methods is analogous to simple count-based word embeddings; it would be highly inefficient and time-consuming, as each word is compared against every other word. For this dissertation, the corpus comprises around 700,000 words. Each word will be represented by a set of 500 vectors resulting in a requirement to optimize 700,000,000 weights in every iteration. Figure 4 visualizes the amount of connections this Neural Network were to have provided the corpus consisted of 7 words using 10 Activation Functions per word.

An effective method to accelerate this procedure is to employ negative sampling. Negative sampling does not concentrate on predicting the neighboring words (a problem of multi-class classification), but instead seeks to forecast the likelihood of randomly produced word pairs appearing in the training corpus. The weights computed for each iteration using this approach are indicated in green in Figure 5. In this dissertation, the approach utilized is a combination of skip gram and negative sampling, which has been introduced and demonstrated to be effective by Li et al., 2021. The initial step involves randomly generating focal-word-neighboring-word pairs. Subsequently, the model acquires the ability to differentiate between negative samples (word pairs with low likelihood of being connected) and focal-word-neighboring-word pairings that are present in the corpus (Li et al., 2021).

4.4 Defining Cultural Values and Seed Words

In order to create a precise cultural dictionary, it is necessary to have a clear starting point. In their study, Guiso et al., 2015 discovered that the corporate websites of companies listed in the S&P 500 frequently emphasize certain values. Specifically, innovation is cited 80% of the time, integrity 70% of the time, quality 60% of the time, respect 70% of the time, and teamwork 50% of the time. These values also exhibit resemblances to the cultural categories proposed by Graham et al., 2022. Although it is technically feasible to utilize these values as seed words, doing so would be impracticable. Words can possess varying meanings depending on the context. For instance, the term "quality" can refer to either the quality of life or the

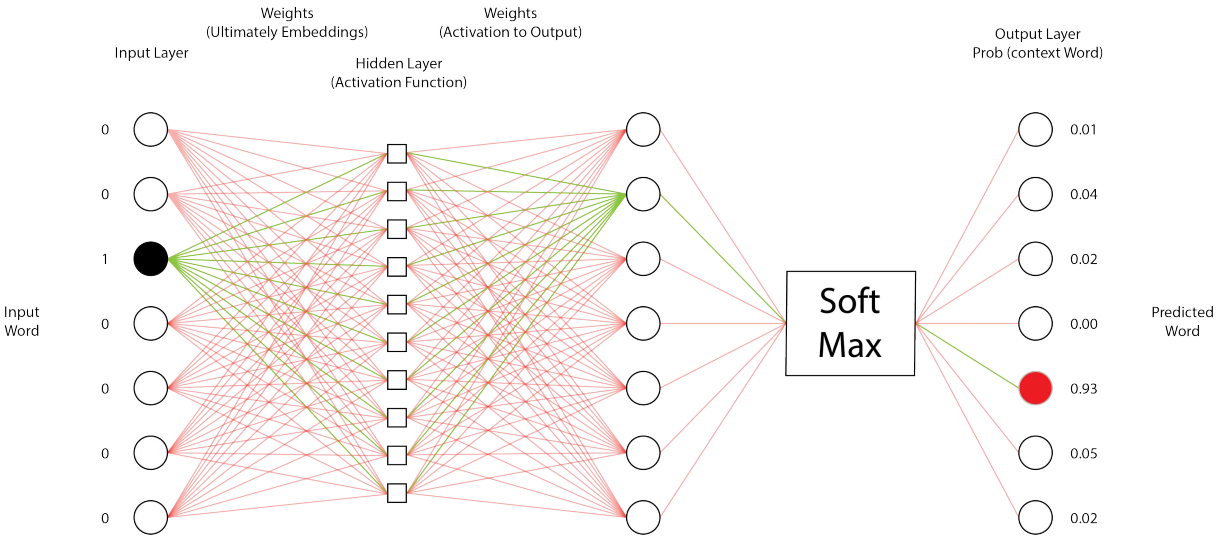


Figure 5: Neural Network for Word Embedding using *word2vec* and Negative Sampling

quality of management. While these notions may be interconnected, they do not share their cultural implications or even the fundamental definition of the word. In order to address this issue, a collection of seed words corresponding to each value introduced by Guiso et al., 2015 will be employed.

There is evidence to indicate that using seed word lists from different disciplines can lead to flawed conclusions, particularly when evaluating financial texts (Loughran and Mcdonald, 2011). To circumvent this constraint, the word vectors corresponding to each value/seed word were retrieved from the trained *word2vec* model and inspected to ensure a precise definition of the cultural value. Appendix 2 contains a list of the cultural values, along with the accompanying seed words.

4.5 Generating Culture Dictionary

To be able to assess culture, the method employed necessitates a dictionary consisting of clearly defined and categorized phrases. This vocabulary is used to detect tendencies towards specific values in the text. The process of quantifying the frequency of specific word combinations is frequently used for automated textual analysis. In order to accomplish this, researchers often employ pre-established lexicons like Harvard's General Inquirer for the purpose of tone and sentiment analysis (E. Henry, 2008; Loughran and Mcdonald, 2011). However, this method would not be adequate for cultural analysis. Culture is a highly nuanced and intricately conveyed idea that necessitates a more tailored assortment of phrases to fully comprehend (Loughran and Mcdonald, 2016). Traditionally, it has been required to

have experts manually recognize and organize the corpus in order to create a suitable cultural dictionary. However, along with the issues of high monetary cost, time consumption, and susceptibility to human mistake, additional concerns exist. Firstly, as previously mentioned, corporate culture is conveyed in a nuanced manner. The vocabulary, consisting of phrases, abbreviations, and even idioms, are easily comprehensible when placed in a specific context. Yet, they become quite difficult to identify as they appear among millions of other words and phrases. Furthermore, Culture is characterized by its varied nature. Correctly and objectively classifying the extracted terms from a corpus is just as challenging, if not more so, than initially identifying them. Finally, dictionaries are subject to the passage of time, as both business language and natural language continually evolve. A dictionary that is not up-to-date will not have the capability to accurately identify contemporary terms like "artificial intelligence". The maintenance of such dictionaries is unrealistic when carried out through humans.

Due to these complexities, it is imperative to compile a tailored corporate dictionary specifically for the objectives of this dissertation. This can be accomplished by employing the cultural values and seed words previously specified, the conversion of words into vectors using the *word2vec* model, and a basic mathematical notion known as cosine similarity. The *word2vec* model that has undergone training has a total of 659,832 words that have been converted into vectors. Cosine similarity is a useful method for measuring the distance between two vectors, which in this context represent indevidual words. A well-established linguistic phenomenon, first described by Harris, 1954, explains that words that share neighboring terms possess comparable meanings. Based on this assumption, the model will determine the similarity between two words by analyzing the words that commonly appear near them in the corpus. By employing this functionality, a roster of counterparts will be produced for each initial word by evaluating the resemblance between these words and phrases stored in the *word2vec* model. In order to create the ultimate culture dictionary, the top 500 words that have the strongest connection (highest cosine similarity) to each cultural value will be included, therefore describing that particular word. In the case where words are present in the top 500 for several values, they will be included in the value that has the highest cosine similarity and removed from the others.

4.6 Scoring Corporate Culture

Ultimately, it is now possible to utilize the cultural lexicon that has been created over a extensive process and draw initial conclusions regarding corporate culture. The measurement of each cultural value is determined by calculating a weighted count of occurrences of seed words. This is done by dividing the number of occurrences by the total number of words in the document. This strategy aligns with previous work that employed word lists to evaluate papers such as Li et al., 2021. The weight utilized is tf.idf (term frequency-inverse document frequency) as recommended by Loughran and Mcdonald, 2011. In this context, the count of the word "i" in document "d" is initially adjusted using the formula (1 + Log(tfi,d)). Afterwards, it is multiplied by the idf weight of Log(N/df), where N represents the total number of documents and df represents the number of documents that contain the word "i". The tf.idf weight is additionally modified based on the similarity between each dictionary word and the seed words. More precisely, the words in the dictionary for each value are arranged in order of similarity, and their similarity weights are calculated as 1/Log(1 + rank). This approach assigns lower weights to terms that have a higher frequency of occurrence.

Given that all conclusions regarding culture are derived from extensively revised and publicly available materials, which have a strong motivation to present a positive image to interested parties, it is probable that specific cultural characteristics may be emphasized in an exaggerated manner that deviates from reality, influenced by market trends. Respect has gained greater importance in contemporary society compared to a decade earlier. Hence, the dataset consists of a panel that measures cultural scores on an annual basis for a period of 10 years, where the greatest and lowest score for each cultural value will be excluded. This is to prevent the inclusion of outliers when assessing an organization's culture. The average of the remaining values will serve as the definitive score for the cultural value in question.

5 Results, Discussion and Implications

5.1 Results

The core focus of this thesis has been the development and application of a methodological approach to quantify corporate culture by categorizing it into five distinct dimensions: innovation, integrity, quality, respect, and teamwork. The outcome is a simple list of companies with their corresponding scores in each of the cultural categories. The results of applying this methodology have underscored its potential utility rather than concentrating on the exact precision of the outcomes.

For further elaboration, as will be discussed later in this section, the scores are statistically similar among different companies. An intriguing aspect, however, arises from the relative nature of the scores within a company. Appendix 3 presents the scores for Meta, Amazon, Apple, Netflix, and Alphabet (FAANG) exhibited in Radar Plots. This enables us to analyze all cultural values inside a singular representation. It is evident that Alphabet has lower scores compared to the other corporations, but their relative positions are accurate. Alphabetăis highly innovative, yet its culture is renowned for being centered on camaraderie and respect. Therefore, it is logical that their rankings in teamwork, respect, and integrity are relatively high in comparison to their lower scores in quality and innovation. Undoubtedly, these scores should be stronger considering that Google is an exceptionally innovative firm. However, it is important to note that the presence ofădefects in the scoring system does not invalidate the viability of the procedure.

In practice, the real-world applicability of this methodology is explored through its potential role in the planning of post-merger integrations. This application is critical as it often presents significant challenges concerning organizational and cultural integration between merging entities. As previously mentioned, evidence suggests that 10-K Forms are an appropriate source to make inferences about cultural values whilst the process presented is proven to quantify natural language (Li et al., 2021).

In evaluating corporate culture, it's essential that the scores across different categories are comparable. This comparability ensures that the assessments of the various cultural dimensions are meaningful and can be reliably compared against each other. Without a common scale or standard, the interpretations drawn from these scores could be deceptive, leading to faulty conclusions and potentially flawed strategic decisions. Comparability al-

lows organizations to accurately gauge their strengths and weaknesses in different cultural areas. For instance, if integrity and respect are measured on different scales, it would be challenging to determine if a high score in integrity truly compensates for a lower score in respect, or vice versa. Therefore, maintaining a uniform scoring scale is fundamental to making balanced and informed assessments. Upon examining the raw scores provided by the methodology employed in this dissertation, several key statistics (Table 4) as well as the distribution graph (Figure 6) indicate that these scores are not immediately comparable.

	innovation	integrity	quality	respect	teamwork
count	5411	5411	5411	5411	5411
mean	47.823	16.018	47.817	11.267	32.344
std	37.06	15.768	30.78	15.565	22.98
min	0	0	0.334	0	0.114
25%	27.786	7.28	30.1	4.623	19.111
50%	39.913	11.849	41.344	7.886	27.689
75%	56.863	19.465	56.241	12.606	38.655
max	539.321	276.111	418.935	300.708	276.17

Table 4: Score Summary Statistics before Processing

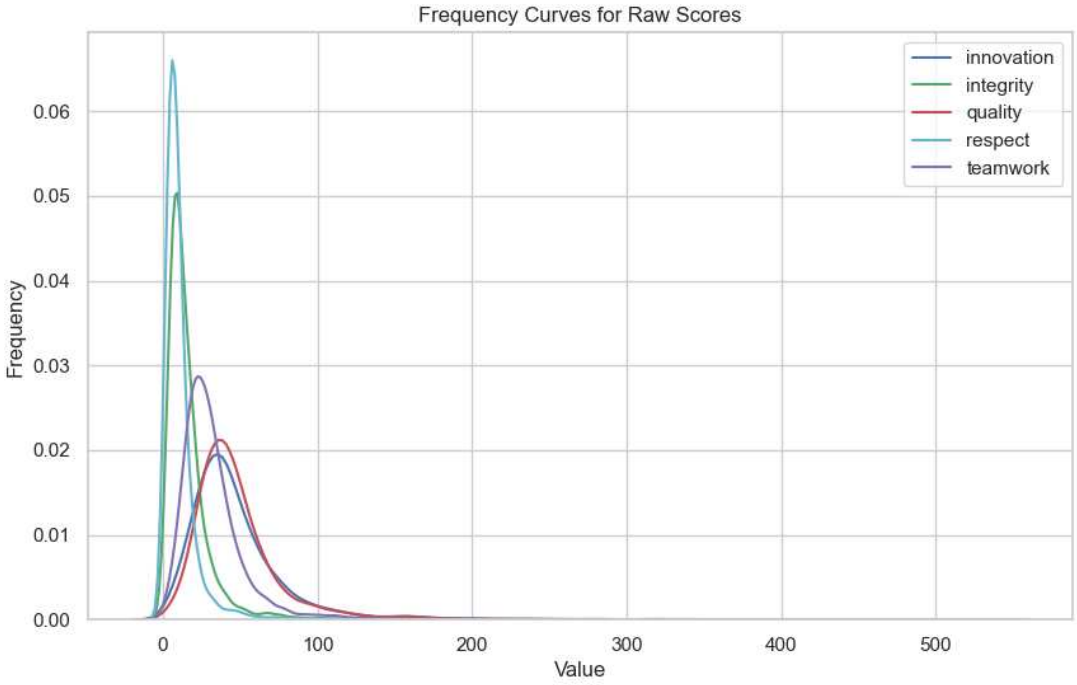


Figure 6: Distribution Curve before Processing

The mean score as well as the Standard Deviation for each cultural value have variability

towards the next. Innovation for instance has a mean score of 52.58 with a standard deviation of 34.65, whilst respect has a mean of 11.80 with a standard deviation of 16.98. Furthermore, a clear skewness issue is present. The interquartile ranges are inconsistent, with the 25th percentiles ranging from 5.09 (Respect) to 30.92 (Innovation). Additionally, the difference between the 75th percentile and the maximum score is minimal for categories like Innovation (65.32 to 277.19) and Respect (12.97 to 206.83), suggesting positive skewness. The high maximum scores are also an indicator of outliers. These discrepancies highlight that the distribution of scores varies across categories. These differences suggest varying averages, score dispersion, skewness, and the presence of possible outliers. This indicates that the raw scores are not on a uniform scale, making direct comparisons potentially misleading.

To ensure the scores across innovation, integrity, quality, respect, and teamwork are comparable, a multi-step normalization process was applied. This process involved logarithmic transformation, outlier trimming, standard scaling, and adjusting scores to eliminate negative values. The first step involved taking the logarithm of each score. Log transformation helps to stabilize the variance across the dataset and reduces the impact of extreme values, making the data more homogenous. Next, the outliers were dealt with. Outliers can skew the data significantly. Trimming the outliers ensures that the scores better represent the central tendency and variability of most of the data. This step involves removing data points that fall outside a specified range, often determined by a certain number of standard deviations from the mean. Next, the data was scaled using StandardScalar from the sklearn library. This method of scaling involves rescaling the data so that it has a mean of 0 and a standard deviation of 1. This process, also known as z-score normalization, transforms the data into a standard normal distribution. This step is crucial for comparability as it ensures that each category has the same mean and variance, allowing for meaningful comparisons across categories. Finally, after standard scaling, some scores have become negative. To ensure all scores are positive and more intuitive, the lowest value across all scores is added to the entire dataset. This adjustment shifts the distribution to the positive range without altering the relative distances between the scores.

This normalization process transformed the raw, non-comparable scores into a standardized, comparable set of scores. This meticulous approach ensures that each category can be assessed on an equal footing. The comparable distributions can also be observed in the distribution curves (Figure 7), as well as in the Box and Whisker Plot (Figure 8) and

	innovation	integrity	quality	respect	teamwork
count	5229	5229	5229	5229	5229
mean	4.3	4.3	4.3	4.3	4.3
std	1	1	1	1	1
min	0.576	0.05	0.333	0.892	1.033
25%	3.715	3.7	3.707	3.755	3.675
50%	4.295	4.309	4.291	4.345	4.299
75%	4.894	4.927	4.866	4.872	4.894
max	8.662	8.309	8.681	8.497	8.394

Table 5: Score Summary Statistics after Processing

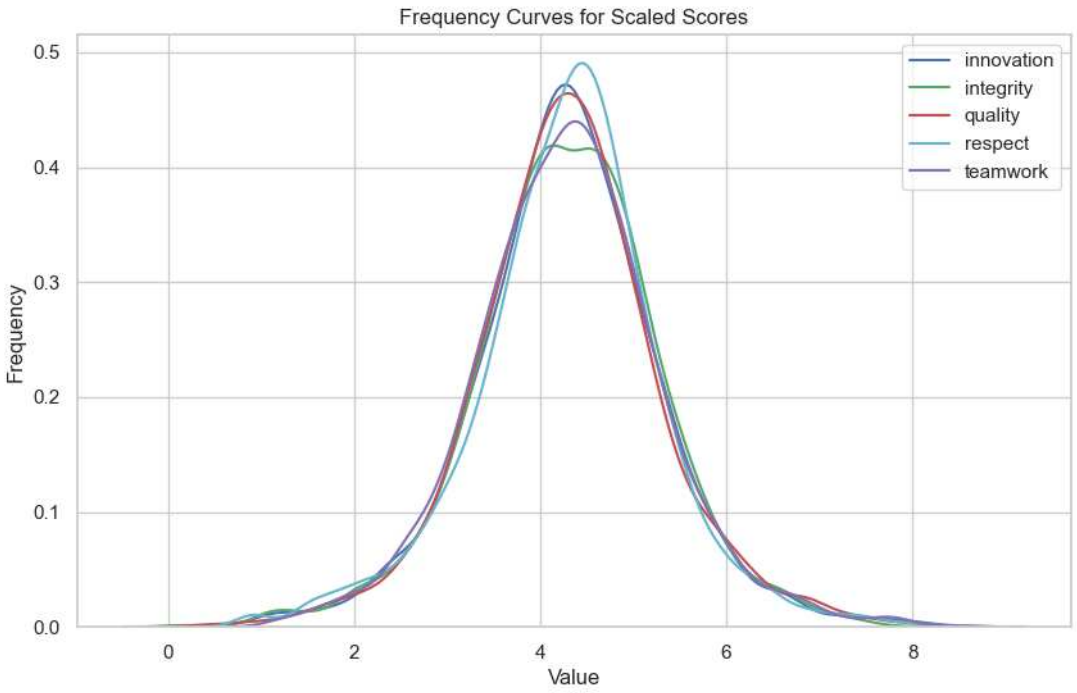


Figure 7: Distribution Curve after Processing

the summary statistics after processing (Table 5). As a result, organizations can make more informed and balanced evaluations of their cultural dimensions, facilitating better strategic planning and organizational development. By applying log transformation, trimming outliers, standard scaling, and adjusting scores to avoid negatives, the resulting data provides a clean, uniform scale that accurately reflects the underlying cultural attributes.

To validate the scaling of the results, several tests were made, the outcomes of which validate that the scores are now comparable. The ANOVA F-Statistic of 0.00 indicates no variance between the group means, implying identical score distributions for innovation, integrity, quality, respect, and teamwork. Similarly, the Permutations Test Statistic of 0.00

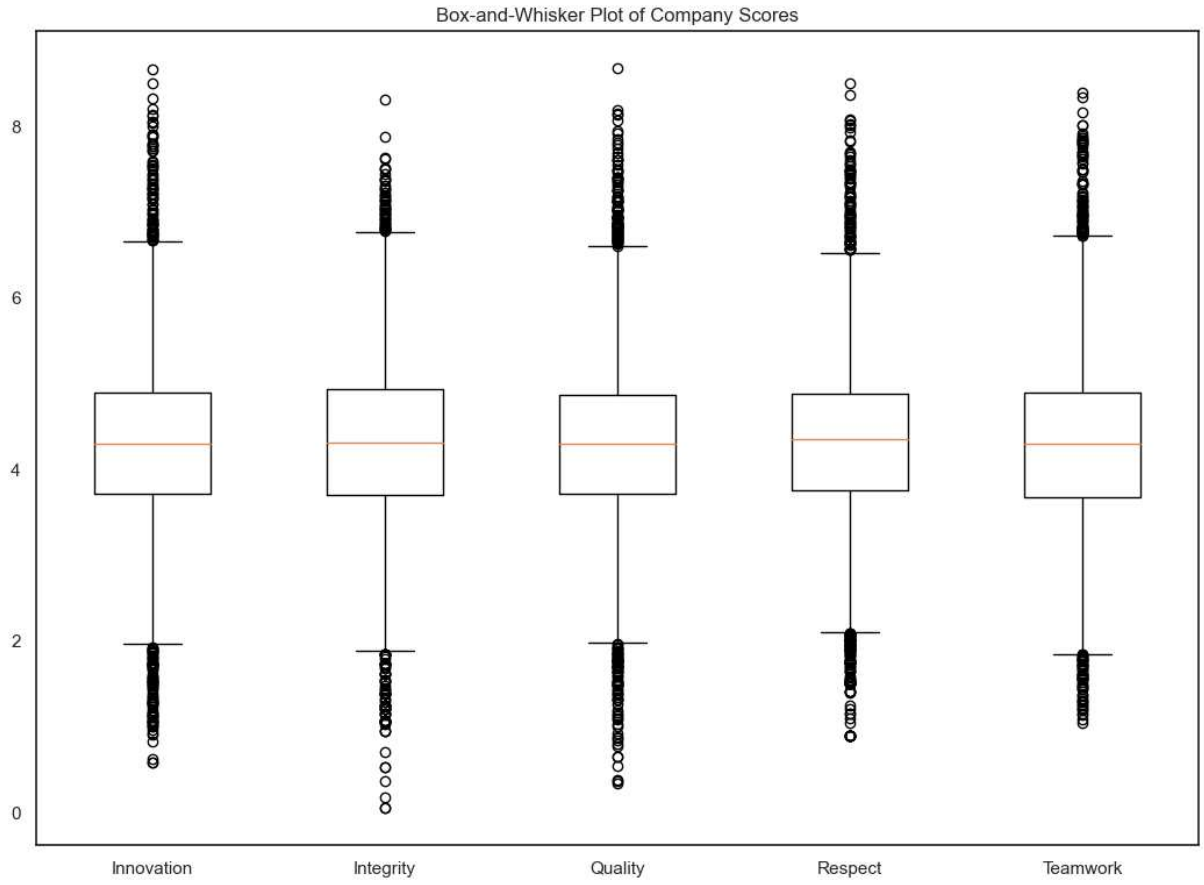


Figure 8: Box-and-Whisker Plot after Processing

shows no observable difference when the data is shuffled, further confirming that the scores for these attributes are not significantly different. Thus, the scores are on a uniform scale and can be reliably compared.

5.1.1 Validation of Results

In order to validate the scores generated by the machine learning model expected relationships between attributes will be confirmed through correlation analysis (Figure 9). This alignment with organizational theories ensures the model's output reflects real-world dynamics, making the scores meaningful. The correlations among the different cultural values are generally high (between 0.66 and 0.87), with enough variance to make insightful statements about their relationships.

The high correlation between quality and innovation (0.87) underscores the importance of fostering both attributes to drive organizational success. High-quality standards provide a strong foundation for innovative practices. When organizations prioritize quality, they establish reliable processes and products, which can support and enhance creative exper-

imentation. This stable base allows teams to take calculated risks and explore new ideas confidently, knowing that the underlying systems are robust. Consequently, a commitment to quality can drive continuous improvement and foster an environment where innovation thrives. On the other hand, the lower but still significant correlation between integrity and innovation (0.66) highlights the complex interplay between maintaining ethical standards and fostering creative growth. A focus on rapid innovation can sometimes lead to cutting corners or unethical behavior. In highly competitive or fast-paced environments, the pressure to innovate quickly can overshadow adherence to ethical standards, leading to compromised integrity. However, innovation does not exist in a vacuum and often requires diverse perspectives and collaborative problem-solving. Teams that work well together can more effectively brainstorm, prototype, and implement new ideas, leading to the hypothesis that high innovation might correlate with high teamwork. This interplay between innovation and teamwork underscores the necessity for organizations to cultivate strong collaborative cultures to support innovative endeavors.

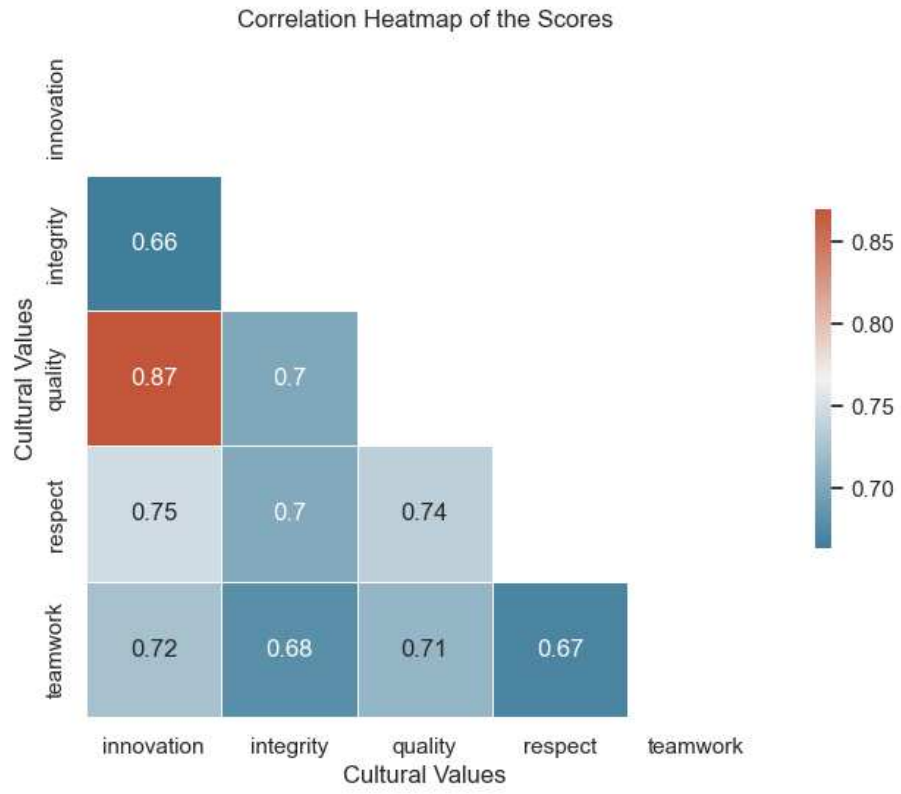


Figure 9: Correlation Heat Map of Scoring

On a different note, organizations that emphasize integrity often have strict standards and ethical guidelines, which can translate into higher quality products and services, as con-

firmed by a correlation of 0.70 between quality and integrity. Employees in such organizations are likely to follow protocols and ensure that their work meets high standards, supporting the hypothesis that high integrity might correlate with high quality. Similarly, delivering high-quality products or services often requires effective collaboration. Teams that work well together can share knowledge, provide feedback, and support each other in maintaining high standards, suggesting a high correlation between quality and teamwork (0.71). Integrity fosters a culture of trust and ethical behavior, which can lead to increased mutual respect among employees. When people act with integrity, they are more likely to treat others respectfully, supporting the hypothesis that integrity correlates with respect (.70).

Whilst the correlations don't paint a clear picture due to the small window in which they exist, these relationships confirm that the model effectively captures the nuanced interconnections between different cultural values, providing a robust foundation for further analysis. The validation of these scores through correlation analysis not only reinforces the reliability of the model but also provides actionable insights for organizations aiming to balance and optimize their cultural attributes.

5.2 Discussion

This methodological approach addresses a significant gap in the existing literature on corporate culture by providing a structured, quantifiable means to assess an often intangible aspect of corporate entities. Several applications are viable for this kind of information. The main application this paper focuses on lies in the improvement of merger success. As stated previousley, M&A deals often fail due to a mismatch between the two entities cultures and the integration strategy, or even a lack of integration strategy. The exploration of the topic along with the outcomes of the method have led to several key insights.

First off, the methodology effectively highlights areas of cultural alignment and disparity between merging companies. For instance, if one company scores highly on 'innovation' while another scores lower, strategic integration efforts can focus on bridging this gap to foster a more innovative unified culture. Furthermore, by quantifying aspects of corporate culture, the methodology aids in making informed decisions regarding integration strategies. It provides a clear picture of where companies might face resistance or where they might leverage cultural strengths, thus supporting smoother integration processes. This enables merging entities to tailor their integration approaches based on their specific needs. For ex-

ample, if both companies score highly on 'teamwork,' this strength can be used as a foundational base for integrating other less aligned cultural areas. As for synergistic opportunities, the presented methodology provides predictive insights into potential challenges during the merger process, enabling proactive rather than reactive integration planning. Needless to say, the discussions and findings presented here primarily demonstrate the applicability and adaptability of the methodology. They show how quantified corporate culture can significantly enhance the planning and execution phases of merger integrations. The emphasis has been on the method's utility and flexibility, showcasing its potential as a tool for strategic organizational development rather than focusing solely on the accuracy of cultural assessments.

5.3 Implications for Research

The quantification of corporate culture, as demonstrated in this thesis, opens numerous avenues for further research that could significantly enhance understanding of corporate culture and its impacts on real world applications. Utilizing the data generated through the methodology of analyzing textual data from 10-K filings via Natural Language Processing (NLP), various additional dimensions of how corporate culture influences and interacts with other aspects of business operations and success could be explored, which reach beyond the scope of this research. An intriguing kind of research can be conducted on cultural relationships as demonstrated earlier.

A different approach entails exploring the understanding of different cultural values by corporations. The model developed in this thesis identifies specific linguistic cues associated with cultural dimensions such as respect. A deeper understanding of how respect in this case is framed within corporate communications, examining the closest words and contexts associated with this value could help in understanding whether respect is predominantly discussed in terms of compliance and authority, or if it leans more towards diversity, inclusion, and employee empowerment. Finally, the methodologies used in this thesis could be applied to conduct cross-industry comparisons to uncover industry-specific cultural profiles. By doing so, researchers can identify whether certain cultural attributes are particularly prevalent in specific industries and how these cultural traits correlate with industry-wide trends and challenges.

Each of these directions could not only validate the usefulness of the data generated in

this thesis but also expand our understanding of the intricate ways in which corporate culture impacts key business outcomes. It is important to note that the limitations don't lie within the cultural values proposed in this thesis. The model is described in a way that it can be altered and augmented to the requirements of the problem at hand. If future researchers decide to name a sixth value, or exclude one, or even come up with an entirely new set of values, this model allows them to do just that. Through these explorations, stakeholders from academics to industry leaders can gain actionable insights that drive more nuanced strategies for corporate management and development.

6 Conclusion

This thesis embarked on a journey to bridge the gap between the qualitative nuances of corporate culture and the quantitative rigor required for corporate analysis, specifically within the context of mergers and acquisitions (M&A). The central research question posed was: How can Natural Language Processing techniques be employed to extract and quantify elements of corporate culture from 10-K textual data, and how can these quantifications be useful in real-world scenarios, particularly in mergers and acquisitions? The findings presented herein not only affirmatively answer this question but also illustrate the transformative potential of integrating advanced machine learning models with linguistic analysis for corporate assessments. A machine learning model incorporating NLP techniques was successfully implemented as a proof of concept to derive quantitative data regarding corporate culture from MD&A sections of 10-Ks. This approach enabled the conversion of textual dataoften dense and overlooked in traditional analysesinto a structured format that quantifies different cultural attributes of organizations. The effectiveness of this model was validated by comparing the derived cultural scores with Glassdoor reviews, which provided an internal viewpoint of the company culture as experienced by employees. This validation process confirmed that the NLP-derived scores were mostly consistent with perceived company cultures, thereby supporting the reliability of the model.

Moreover, the usefulness of the generated data was demonstrated through various implications for further research and practical applications in the real world. The quantified cultural data offers a novel lens to better plan post-merger integrations and thereby improve of M&A outcomes. The notion that purely cultural compatibility can significantly influence the success rates of mergers has been reconsidered and reposed as a need to plan integration by identifying the appropriate strategy based on the cultures of the merging corporations. By providing a measurable way to assess and compare the cultural profiles of potential merger partners, this research contributes valuable insights that can inform strategy and decision-making processes in M&A activities.

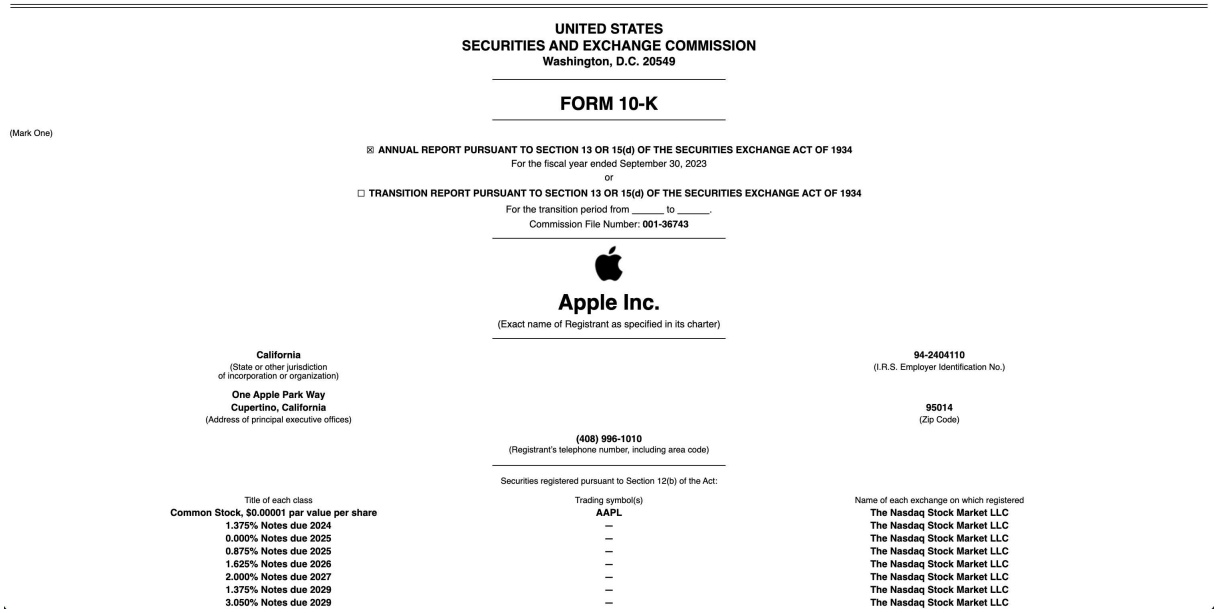
The implications for further research are profound. They suggest avenues for future investigations into the relationships between specific cultural attributes and various business performance metrics, such as innovation output, employee satisfaction, and profitability. Each of these factors is critically relevant in today's fast-evolving business environment, where organizational culture is increasingly recognized as a driver of long-term sustainabil-

ity and competitive advantage.

In conclusion, this thesis not only answers its initial research question by demonstrating a successful application of machine learning and NLP in quantifying corporate culture from corporate disclosures but also sets the stage for subsequent empirical and practical explorations. It underscores the potential of such quantitative measures to enhance our understanding of corporate dynamics and to improve the strategic execution of corporate mergers and acquisitions. The foundation laid by this research is poised to catalyze further innovations in the interdisciplinary fields of corporate governance, strategic management, and organizational psychology.

Appendix

Appendix 1 - 10-K Formats



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the fiscal year ended September 30, 2023
or
☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the transition period from _____ to _____.
Commission File Number: **001-36743**

Apple Inc.
(Exact name of Registrant as specified in its charter)

California	**94-2404110**
(State or other jurisdiction of incorporation or organization)	(I.R.S. Employer Identification No.)
One Apple Park Way **Cupertino, California**	**95014**
(Address of principal executive offices)	(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading symbol(s)	Name of each exchange on which registered
Common Stock, $0.00001 par value per share	**AAPL**	**The Nasdaq Stock Market LLC**
1.375% Notes due 2024	—	**The Nasdaq Stock Market LLC**
0.000% Notes due 2025	—	**The Nasdaq Stock Market LLC**
0.875% Notes due 2025	—	**The Nasdaq Stock Market LLC**
1.625% Notes due 2026	—	**The Nasdaq Stock Market LLC**
2.000% Notes due 2027	—	**The Nasdaq Stock Market LLC**
1.375% Notes due 2029	—	**The Nasdaq Stock Market LLC**
3.050% Notes due 2029	—	**The Nasdaq Stock Market LLC**

10-K Filing in Document Form

```
<SEC-DOCUMENT>0000320193-23-000106.txt : 20231103
<SEC-HEADER>0000320193-23-000106.hdr.sgml : 20231103
<ACCEPTANCE-DATETIME>20231102180827
ACCESSION NUMBER:               0000320193-23-000106
CONFORMED SUBMISSION TYPE:      10-K
PUBLIC DOCUMENT COUNT:          96
CONFORMED PERIOD OF REPORT:     20230930
FILED AS OF DATE:               20231103
DATE AS OF CHANGE:              20231102

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 Apple Inc.
                CENTRAL INDEX KEY:                      0000320193
                STANDARD INDUSTRIAL CLASSIFICATION:     ELECTRONIC COMPUTERS [3571]
                IRS NUMBER:                             942404110
                STATE OF INCORPORATION:                 CA
                FISCAL YEAR END:                        0930

        FILING VALUES:
                FORM TYPE:              10-K
                SEC ACT:                1934 Act
                SEC FILE NUMBER:        001-36743
                FILM NUMBER:            231373899

        BUSINESS ADDRESS:
                STREET 1:               ONE APPLE PARK WAY
                CITY:                   CUPERTINO
                STATE:                  CA
                ZIP:                    95014
                BUSINESS PHONE:         (408) 996-1010

        MAIL ADDRESS:
                STREET 1:               ONE APPLE PARK WAY
                CITY:                   CUPERTINO
                STATE:                  CA
                ZIP:                    95014

        FORMER COMPANY:
                FORMER CONFORMED NAME:  APPLE INC
                DATE OF NAME CHANGE:    20070109

        FORMER COMPANY:
                FORMER CONFORMED NAME:  APPLE COMPUTER INC
                DATE OF NAME CHANGE:    19970808
</SEC-HEADER>
<DOCUMENT>
<TYPE>10-K
<SEQUENCE>1
<FILENAME>aapl-20230930.htm
<DESCRIPTION>10-K
<TEXT>
<XBRL>
```

10-K Filing in HTML Form

Appendix 2 - Cultural Values and Seed Words

This Table lists the Cultural Values along with the seed words for those cultural values used to compile the culture dictionary, which in turn is later used to score corporate culture. This set of terms is proposed by Guiso et al., 2015.

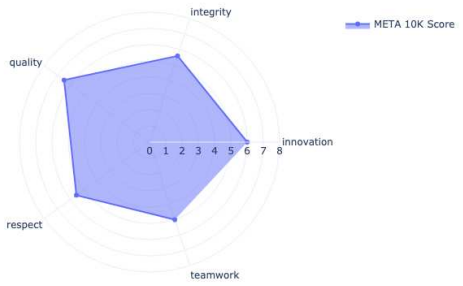
Begin of Table	
Cultural Value	**Seed Words**
integrity	integrity ethic ethical accountable accountability trust honesty honest honestly fairness responsibility responsible transparency transparent
teamwork	teamwork collaboration collaborate collaborative cooperation cooperate cooperative collaboration cooperation unity

Continuation of Appendix 2	
Cultural Value	**Seed Words**
temwork cont.	partnership coordination alliance cohesion harmony collective solidarity
innovation	innovation innovate innovative creativity creative create passion passionate efficiency efficient excellence
respect	respect respectful talent talented employee dignity empowerment empower

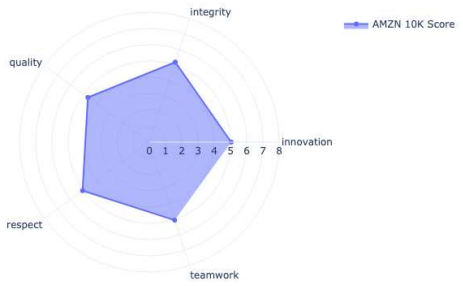
Continuation of Appendix 2	
Cultural Value	**Seed Words**
quality	quality customer customer_commitment dedication dedicated dedicate customer_expectation
End of Table	

Appendix 3 - Cultural Radar Plots for FAANG Companies

Meta Platforms Inc.



Amazon.com Inc.



Apple Inc.

Netflix Inc.

Alphabet Inc.

References

Beautiful soup documentation. (2023). Retrieved May 22, 2024, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, *50*(4), 623–646. https://doi.org/10.1017/S0022109015000411

Burke, S. (n.d.). *HP to buy compaq in $25 billion stock deal | CRN*. Retrieved May 22, 2024, from https://www.crn.com/news/channel-programs/18836553/hp-to-buy-compaq-in-25-billion-stock-deal

Crémer, J. (1993). Corporate culture and shared knowledge *. *Industrial and Corporate Change*, *2*(3), 351–386. https://doi.org/10.1093/icc/2.3.351

Fang, Y., Fiordelisi, F., Hasan, I., Leung, W. S., & Wong, G. (2023). Corporate culture and firm value: Evidence from crisis. *Journal of Banking & Finance*, *146*, 106710. https://doi.org/10.1016/j.jbankfin.2022.106710

Graham, J. R., Grennan, J., Harvey, C. R., & Rajgopal, S. (2022). Corporate culture: Evidence from the field. *Journal of Financial Economics*, *146*(2), 552–593. https://doi.org/10.1016/j.jfineco.2022.07.008

Griffin, P. A. (2003). Got information? investor response to form 10-k and form 10-q EDGAR filings. *Review of Accounting Studies*, *8*(4), 433–460. https://doi.org/10.1023/A:1027351630866

Guiso, L., Sapienza, P., & Zingales, L. (2015). The value of corporate culture. *Journal of Financial Economics*, *117*(1), 60–76. https://doi.org/10.1016/j.jfineco.2014.05.010

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(1), 146–62.

Haspeslagh, P., & Jemison, D. (1991). Managing acquisitions : Creating value through corporate renewal. *New York: Free Press*. Retrieved May 22, 2024, from https://library.wur.nl/WebQuery/titel/1869872

Henry, E. (2008). Are investors influenced by how earnings press releases are written? [Publisher: SAGE Publications]. *The Journal of Business Communication (1973)*, *45*(4), 363–407. https://doi.org/10.1177/0021943608319388

Henry, J., & Van Oostende, M. (2023). *Top m&a trends in 2024: A blueprint for success | McKinsey*. Retrieved May 22, 2024, from https://www.mckinsey.com/capabilities/m-and-a/our-insights/top-m-and-a-trends-in-2024-blueprint-for-success-in-the-next-wave-of-deals

Hoberg, G., & Maksimovic, V. (2015). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies*, *28*(5), 1312–1352. https://doi.org/10.1093/rfs/hhu089

Hoberg, G., & Maksimovic, V. (2022). Product life cycles in corporate finance. *The Review of Financial Studies*, *35*(9), 4249–4299. https://doi.org/10.1093/rfs/hhab134

Inkpen, A. C., Sundaram, A. K., & Rockwood, K. (2000). Cross-border acquisitions of u.s. technology assets. [Publisher: California Management Review]. *California Management Review*, *42*(3), 50–71. https://doi.org/10.2307/41166042

Kreps, D. M. (1990, September 28). Corporate culture and economic theory [Google-Books-ID: wKJu6g5ovhcC]. In *Perspectives on positive political economy* (pp. 90–110). Cambridge University Press.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. https://

proceedings.neurips.cc/paper_files/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf

Li, K., Griffin, D., Yue, H., & Zhao, L. (2013). How does culture influence corporate risk-taking? *Journal of Corporate Finance*, *23*, 1–22. https://doi.org/10.1016/j.jcorpfin.2013.07.008

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning [Publisher: Oxford University Press / USA]. *Review of Financial Studies*, *34*(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079

Loughran, T., & McDonald, B. (2024, February). *Stopwords*. Retrieved May 22, 2024, from https://sraf.nd.edu/textual-analysis/stopwords/

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x]. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & Mcdonald, B. (2016). Textual analysis in accounting and finance: A survey [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12123]. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

O'Reilly, C. A., & Chatman, J. A. (1996). Culture as social control: Corporations, cults, and commitment. In *Research in organizational behavior: An annual series of analytical essays and critical reviews, vol. 18* (pp. 157–200). Elsevier Science/JAI Press.

Puranam, P., Singh, H., & Zollo, M. (2006). Organizing for innovation: Managing the coordination-autonomy dilemma in technology acquisitions [Publisher: Academy of Management]. *Academy of Management Journal*, *49*(2), 263–280. https://doi.org/10.5465/amj.2006.20786062

Ranft, A. L. (2006, January 1). Knowledge preservation and transfer during post-acquisition integration. In C. L. Cooper & S. Finkelstein (Eds.), *Advances in mergers and acquisitions* (pp. 51–67, Vol. 5). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-361X(06)05003-4

Renneboog, L., & Vansteenkiste, C. (2019). Failure and success in mergers and acquisitions. *Journal of Corporate Finance*, *58*, 650–699. https://doi.org/10.1016/j.jcorpfin.2019.07.010

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45715-1_1

SEC.gov. (2023, April 23). Retrieved May 22, 2024, from https://www.sec.gov/edgar/about

Spedale, S., Van Den Bosch, F. A. J., & Volberda, H. W. (2007). Preservation and dissolution of the target firm's embedded ties in acquisitions [Publisher: SAGE Publications Ltd]. *Organization Studies*, *28*(8), 1169–1196. https://doi.org/10.1177/0170840607075672

Vermeulen, F., & Barkema, H. (2001). Learning through acquisitions [Publisher: Academy of Management]. *Academy of Management Journal*, *44*(3), 457–476. https://doi.org/10.5465/3069364

Wearden, G. (2007). From $35bn to $7.4bn in nine years. *The Guardian*. Retrieved May 22, 2024, from https://www.theguardian.com/business/2007/may/14/motoring.lifeandhealth

Weber, Y., & Schweiger, D. M. (1992). TOP MANAGEMENT CULTURE CONFLICT IN MERGERS AND ACQUISITIONS: A LESSON FROM ANTHROPOLOGY [Publisher: MCB UP Ltd]. *International Journal of Conflict Management*, *3*(4), 285–302. https://doi.org/10.1108/eb022716

Wright, R. (2011, September 8). *The HP-compaq merger: Partners reflect 10 years later | CRN*. Retrieved May 22, 2024, from https://www.crn.com/news/mobility/231601009/the-hp-compaq-merger-partners-reflect-10-years-later