Universiteit
Leiden
The Netherlands

Bachelor Computer Science & Economics

Language and Deal Characteristics

in Mergers and Acquisitions:

A Subgroup Discovery Study

Olav Witvliet S2642964

Supervisors:
Dr. A.J. Knobbe
Dr. M. van Leeuwen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/11/2025

Abstract

Introduction
Mergers and Acquisitions (M&A) are critical events in financial markets, significantly influencing stock prices and company value. However, understanding how quantitative deal metrics and qualitative language features from company communications interact to predict deal outcomes remains limited. Motivated by advances in Natural Language Processing (NLP) and data mining, this research aims to clarify which linguistic and deal characteristics are most strongly associated with M&A deal completion and market reactions.

Methodology
This research proposes a pipeline that combines the standardised SDC Platinum database and textual data from SEC DEFM14A filings. Transformer-based NLP models extracted sentiment scores covering, e.g., positivity, risk, and overconfidence. Subgroup Discovery, a descriptive data mining technique, was applied to identify nuanced patterns linking quantitative deal metrics with qualitative textual sentiment features across 1754 US-based M&A deals from 2002 to 2017.

Results and Discussion
Findings indicate that language features such as positivity and uncertainty, alongside deal characteristics such as smaller deal size, non-complexity, and cash-based structures, are strongly associated with successful deal completion and a higher than 7% uplift in target stock price at announcement compared to the average target stock price. The combined analysis of textual and financial variables presents more nuanced descriptive insights than either alone, uncovering non-linear relationships that traditional models often miss. Limitations include the study's data preselection, missing entries in the dataset, and model token constraints in combination with the selected standard Subgroup Discovery strategies.

Conclusion
This research presents a novel, reproducible pipeline that advances the understanding of M&A deal dynamics by integrating quantitative and qualitative data analysis. It offers practical implications for financial analysts, investors, and policymakers and already presents meaningful findings, such as subgroups that characterise deals that complete successfully or exhibit a high target stock price increase. Future work should expand datasets, incorporate more diverse qualitative sources, and enhance NLP methodologies to deepen insights into the complex interplay shaping M&A outcomes.

Keywords: Subgroup Discovery, M&A, Sentiment Analysis, Market Reactions, NLP, Proxy Statements, DEFM14A, Cosine Similarity, Stock Price

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Arno Knobbe and Matthijs van Leeuwen, for their invaluable support, guidance, and flexibility throughout the process of writing this thesis. Their insights and constructive feedback have been essential to completing this work. I would also like to thank my study advisor, Ilja van den Brandt, for her assistance during the graduation phase. Finally, I am deeply grateful to my friends and family for their continuous support, patience, and understanding during the entire thesis-writing period.

Contents

1 Introduction

Mergers and Acquisitions, or M&A for short, are a key topic in the study of economics and the analysis of financial market activities. M&A are complex, high-impact market events that enable companies to strengthen their market position by consolidating with other companies, generating more revenue, and gaining access to new technologies and markets. These activities have an influence that extends beyond the companies directly involved. They can shift the market's sentiment, competitive landscape, and stock prices, driving innovation and value creation in the economy. The measurement of value creation with M&A deals is described as *synergies*. *Synergies* capture the incremental value created when two firms combine, resulting in a combined entity that is worth more than the sum of the separate companies. The main types of synergies are *cost synergies*, which are created by the efficiency of consolidation, where expenses can be lowered by minimising redundancies, and *revenue synergies*, which are gained by a stronger market position and the ability to sell more. The financial relevance is further highlighted by an empirical analysis conducted by Feng and Wang (2020), which substantiates the general trend that, in the long term, M&A activities and global economic growth influence each other positively. Since 2000, more than 790 million M&A transactions have been announced worldwide, with a value exceeding $57 trillion, as noted by the Institute for Mergers, Acquisitions and Alliances (IMAA) (2025).

Each day, the corpus of M&A-related texts is growing with new business reports, news articles, regulatory filings, and academic findings. These texts are outgrowing human capacity to process and interpret them all, creating an overload that renders the traditional method of financial and sentiment analysis infeasible. Meanwhile, the abilities of analytical techniques are advancing rapidly, especially in the field of Natural Language Processing (NLP). This further develops the capability to utilise the most recent techniques, such as Transformers and Large Language Models (LLMs), to bridge the gap between the vast amount of unexplored qualitative text surrounding M&A deals, the extensive dataset of quantitative market data, and analysis insights from subgroup analysis.

The financial field is rich in literature, dating back to the first pursuit of a mathematical approach to quantifying the stock exchange, undertaken by Bachelier (1900). Following Bachelier's foundational work on quantifying the financial market, new approaches evolved towards economic models and machine learning. With current advancements, as seen in recent fundamental works on textual analysis in finance, more methods of financial research are being explored. A key example of these advancements is the domain-specific sentiment groundwork by Loughran and McDonald (2011), which introduced the ability to study financial language with general language models. Recent advancements in computer science have also enabled the research of underlying relations and connections within complex data sources. A notable example of this is Subgroup Discovery (SD), which further illustrates how algorithmic pattern detection can reveal hidden, non-linear relationships within such data (Atzmueller, 2015). This research utilises an SD methodology proposed by Knobbe et al. (2021). The convergence of these advances in financial language analysis methods and data analysis methods demonstrates the growing capacity to quantify the complex nature of aspects of financial communication and market behaviours.

1.1 Problem Statement

Despite these advances, the intersection between advanced data analysis methods and the domain-specific context surrounding M&A deals remains understudied, and a reproducible analysis framework that integrates qualitative text and quantitative financial data has yet to be developed. This knowledge gap is further highlighted by Berens et al. (2023), who note that data complexity is the most significant scientific challenge of our time. The integration of AI and data analysis into various application domains, such as economics, is necessary to bridge this gap and drive new scientific insights and innovations. The gap described can also be found in research in the computer science domain, where the breadth and depth around the economic context are missing, and therefore, this intersection is understudied. To address this gap, a designated pipeline for data analysis in finance is needed to facilitate interdisciplinary research, such as identifying relationships between textual sentiments, deal information, and M&A outcomes.

Early studies, such as Hajek and Henriques (2024), demonstrate the value of a structured analysis pipeline and methodology, showing that sentiment derived from news articles can significantly influence M&A outcomes. Similarly, Morgan (2018) highlights the predictive strength of textual sentiment analysis by leveraging established datasets, such as SDC Platinum and standardised SEC filings. However, a common challenge across these works is either inconsistent standardisation of data, exemplified by the 298,134 news headlines in Hajek and Henriques (2024), or limited explanatory power due to the omission of critical factors such as managerial intent, market dynamics, and regulatory influences. These gaps contribute to lower model performance, reflected in modest R-squared values, underscoring the need for enhanced variable inclusion and methodological refinement.

1.2 Motivation

Given my academic background in economics and the application of computer science methods, I wanted to integrate these two disciplines and propose an interdisciplinary solution to address current analysis complexities, providing an actionable pipeline for future work. The methodology of this research utilises a standardised data analysis pipeline for M&A deal information. The research field of M&A is a perfect domain for the application of advanced data-driven methods, as it combines strategic decisions, behavioural economics and a large and diverse corpus of financial and textual data. The rapid development of AI-driven methods also offers the opportunity for a deeper empirical examination of the relationship between theoretical concepts from the economic sciences, such as synergies and market behaviour, and textual sources. By aligning these methods with financial theory, this research contributes to a more holistic understanding of M&A dynamics.

The timing of this research is relevant, as the analysis models have reached a point where qualitative information, such as sentiment and textual tone, can be utilised in non-linear data analysis and financial decision-making. In current literature, the focus is shifting from linear and purely numerical analyses to a new dimension where the strategic language of companies, reflecting their intentions and expectations, is considered. By applying NLP to M&A texts, new insights are gained into the underlying factors, including perception, tone, and confidence surrounding mergers and acquisitions. These factors are under-represented in the current literature due to their non-numerical and complex

nature.

This research creates relevance as the relationship between language, deal information, and deal outcomes can help investors and researchers address M&A deals on a deeper level, considering risk and potential financial returns. Companies and policymakers may also reconsider their approaches to communicating about M&A deals, and the methodological design used in this research will aid future applications.

1.3 Research Scope

Given the exploratory nature of this research and the scarcity of prior related studies on the application of advanced analysis methods on M&A data, the scope is designed to produce reproducible insights that are applicable and relevant to broader financial contexts and applications. To be able to be reproducible and have applicable findings to different domains within the finance and M&A context, the generally established data source of SDC Platinum dataset from the London Stock Exchange Group (LSEG) is used for M&A information. Then, to add an additional qualitative layer, the publicly available DEFM14A proxy statements from the United States Securities and Exchange Commission (SEC) Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) filing database have been added for sentiment analysis alongside the quantitative SDC Platinum data. The focus of these data sources is on publicly traded deals with a value exceeding $100 million involving US-based companies during the period from 2002 to 2017. The period was chosen because, after 2001, the SEC introduced a standardised format for companies to document their regulatory filings. Additionally, all deals within this period were completed by the time this research began. The longer period helps reduce bias from macroeconomic shifts confined to shorter time windows. The companies being US-based and the chosen deal size are typical within the M&A literature, providing a large dataset of deals to research. By establishing these specific criteria, a streamlined process is created for collecting and analysing deal outcomes and market data, supporting the generation of meaningful and reproducible insights aligned with the research objectives and methodologies outlined.

Within this scope, the subgroup discovery analysis explicitly focuses on one nominal target or one numeric target at a time, rather than modelling several outcome variables together as a single complex target. Deal completion, announcement-day stock price reactions and valuation changes are therefore treated as separate targets, each with its own subgroup search configuration. Analysing these outcomes simultaneously would require an Exceptional Model Mining setup with multi-target quality measures and more intricate model classes in order to capture joint patterns across binary and numeric targets in a principled way. Given the size and diversity of the M&A dataset, such a multi-target configuration would also require an additional round of systematic experimentation to understand the trade-off between model complexity, stability and interpretability. By staying within the formulated research scope, a clear baseline design for the analysis pipeline is adopted with a single nominal and a single numeric target in order to correctly explore single-variable associations in the data. This will aid an extension to a joint, complex target setting as a natural direction for future work.

1.4 Research Goals

The goal of this research is to provide novel insights into the associations of variables within M&A deals and to establish a reproducible and standardised analysis pipeline that invites future research to delve further into the topic. This will be achieved by feature engineering the SDC Platinum dataset and the feature extraction method with the SEC.gov DEFM14A filings. This will create a final quantitative dataset on which recent methodological innovations, such as Subgroup Discovery, are applied to convert the complex patterns hidden within the data into understandable features and insights. The research question formulated to achieve these goals is:

> *Which language features and deal metrics are most strongly associated with M&A deal completion and related market reactions, such as stock price or company value changes?*

The insights gained through this research will be compared with existing literature to evaluate their significance and relevance. Combined with the reproducible analysis pipeline developed, these insights will enable the identification of new dynamics within the complex M&A data landscape, such as how sentiment expressed in company communications before a deal influences both short-term market reactions and long-term deal outcomes. This research will also deepen understanding of which general deal metrics and sentiment categories (including positivity, risk, and confidence) hold the strongest associations with deal outcomes. Furthermore, it will clarify how sentiment indicators can be effectively extracted from M&A filings and integrated with traditional financial variables to enhance the accuracy and depth of outcome predictions.

1.5 Research Outline

This research has been structured into several chapters. It begins with Chapter 2, which reviews related topics and literature to establish a foundation for this research. The economic landscape, market, and M&A deals will be described in context. Then, a deep dive into sentiment analysis and subgroup discovery in the financial context will be done. Chapter 3 describes the methodology, from the data collection steps involving both the SDC Platinum database and the DEFM14A filings, to feature engineering and extraction, and finally, gathering the full quantitative dataset that will be used in the described data analysis with subgroup discovery. Chapter 4 presents the results of the data analysis. Chapter 5 discusses the results and their implications. The final chapter provides a summary of the research and concludes the study with concluding remarks.

2 Theoretical Background

Chapter 2 offers the theoretical foundation for this research by reviewing key literature and clarifying the reasoning behind the chosen methodology. It begins with an overview of the economic landscape of mergers and acquisitions (M&A), emphasising the financial and scientific impact of market reactions and introducing the industry-standard SDC Platinum dataset as a cornerstone for quantitative M&A analysis. The discussion then turns to quantitative data analysis methods in economic research, with a focus on integrating sentiment analysis into financial studies. Additionally, the chapter explains the principles and relevance of subgroup discovery as a data mining method, highlighting its unique suitability for analysing M&A data. Throughout, a theoretical context is provided, major research gaps are identified, and methodological decisions to address these gaps are made clear. Collectively, these elements establish a strong foundation for the combined quantitative and qualitative analysis pipeline used in this study, supporting robust insights and informing subsequent results and discussion chapters.

2.1 M&A Deals and Market Reactions

For more than a century, M&A activities have been a vital source for analysing business strategy and market behaviour. Almost 1 million transactions have been announced worldwide with a value of over 57 trillion U.S. dollars since the year 2000 (Bollaert and Delanghe (2015)). These activities are a significant driver for the global economy.

Work from Martynova and Renneboog (2008) highlights the diverse M&A waves that exist, characterised by economic and technological expansion, market regulation or managerial decision-making. The work of Gaughan (2010) further examines the intricacies of these M&A activities, illustrating the restructuring of markets and business structures that these M&A waves cause, as well as their effects on market efficiency, competitive changes, and value creation.

To address the effects of M&A activities on the market, two methods of assessment are used: short-term event studies and long-term synergy analysis. The short-term event studies examine the market effects in the days leading up to and following the deal, including stock price changes, changes in company value, and deal completion metrics. The foundational work of Andrade et al. (2001) provided empirical insights into M&A deals, revealing key trends and characteristics. Notably, mergers create shareholder value, with the gains mostly accruing to the acquired company involved in the merger, yielding a significant return of more than 16%. These acquired companies are also referred to as the targets of a deal. The literature suggests that acquiring companies exhibit neutral to slightly negative returns. However, these results are challenging to substantiate due to external factors influencing post-deal measures, such as integration challenges, costs, and the diverse financing types associated with the deals, which can impact the overall company stock. Andrade et al. further describe the importance of M&A industry clustering and the effects of deregulation on M&A activities. However, they also challenge their findings by highlighting the inability to describe deeper mechanisms in place and the long-term impact on what makes M&A deals successful or unsuccessful. This is the second method for assessing the effect of M&A activity, which is described in terms of synergies. Synergies are the long-term measure that represents the cost efficiency benefits and the competitive benefits of companies consolidating. In the short term,

an acquirer may risk devaluing a company to gain more revenue. These results can oppose the short-term outcomes, making the analysis complex.

With all these deals comes a vast amount of data, and the literature has shown the use of various sources. These sources range from hand-collected information of news articles and market reactions surrounding M&A activities to the use of standardised datasets. However, there is still no golden standard in the use of M&A information within the academic literature. The datasets most commonly used are Compustat, Orbis, CRSP, Zephyr & SDC Platinum. Comparative analysis, as seen in Bollaert and Delanghe (2015), reveals that SDC Platinum is the most accurate among news sources, with an edge in general research. Thomson Reuters (2010) states that SDC Platinum is the industry standard, with no other single dataset surpassing it in content. The content on M&A information consists of over 900,000 global M&A transactions, of which the most significant portion is over 280,000 US-related deals. Therefore, SDC Platinum is chosen for this research and will form the core of the proposed research pipeline.

With SDC Platinum being widely used in academic research, it has also been evaluated for its integrity and validity. Barnes et al. (2014) evaluated 20 years of data from SDC Platinum and provided practical guidelines on the pros and cons of using hand-collected information compared to a general dataset. The hand-collection method has the potential to produce more accurate data, but it is also more time-consuming and faces consistency issues due to human errors that can vary between researchers. Discrepancies within the SDC data were highlighted when focusing on smaller deals, but these claims lacked significance. For this exploratory research, these biases are taken into account but are not expected to pose significant risks to the results or integrity.

The SDC Platinum dataset includes 167 distinct variables that cover numerous important determinants of M&A activity. It includes the measures for market reaction and firm value, such as equity and enterprise value, stock prices, deal premiums, and valuation multiples. Also crucial for prediction purposes are the forecast and management expectation variables, which reflect expectations of future performance, including synergies and accretion effects. Variables for deal context are also included, ranging from financial, accounting and transactional variables that describe firm fundamentals through assets, debt, earnings, profitability and the percentage of shares acquired, deal type, value, termination fees, and payment structure to industry, geographic, and outcome variables that provide deal context and results, identifying sectors, locations, and completion status. Together, these variables support a structured dataset to analyse market behaviour, firm performance, and deal outcomes in M&A research.

In summary, M&A deals are a key topic of economic and financial research. However, extracting meaningful relationships from these deals remains challenging due to the complex and diverse data distribution associated with these deals and the absence of a standard data source for analysis. Therefore, this paper will present preliminary results using a diverse range of metrics from the standardised SDC dataset, providing a foundation for future work to build upon.

2.2 Sentiment Analysis in Financial Context

Sentiment analysis is a textual data analysis method that extracts quantitative features from qualitative sources, such as news articles, company announcements, and general public texts. The first form of sentiment analysis, or specifically Natural Language Processing (NLP), was done around the 1960s. The early works, such as those from Stone and Hunt (1963), pioneered the computer approaches to data and text analysis. In the early phases, rule-based and symbolic models focused on machine translation and text analysis, laying the groundwork for the statistical and machine learning approaches that are still in use today. Over the last 20 years, new methodological approaches have been introduced, helping to deepen our understanding of the economic and financial factors at play in market activities. These general approaches began with word lists and have evolved into recent advanced machine learning-based analysis models. Effective analysis of M&A deals requires variables that capture market behaviour, firm performance, and deal results, enabling the development of predictive methods to better understand and forecast deal outcomes.

Loughran and McDonald (2011) transformed financial sentiment research by demonstrating that many words considered negative in general language (e.g., 'liability') are neutral in economic contexts. By creating domain-specific dictionaries, they improved sentiment classification accuracy and established empirical links between disclosure tone and market reactions. Related evidence shows that market anticipation of M&A is associated with increases in stock market valuation (Bennett and Dam, 2019). Recent studies validate that sentiment derived from news articles provides evidence on the effect of M&A outcomes (Hajek and Henriques, 2024). Morgan (2018) highlight the predictive power of textual sentiment analysis, utilising datasets from the established SDC Platinum dataset and standardised public SEC.gov filings. Across these studies, a recurring pattern is observed: either the data is inconsistently standardised (e.g., the 298,134 news headlines in Hajek and Henriques (2024)), or the explanatory power is limited by low R-squared values because important variables influencing deal outcomes (managerial intent, market conditions, regulatory factors) remain unmodelled or omitted (Morgan, 2018).

To capture this sentiment, a corpus of M&A data related to M&A deals needs to be retrieved. There are two general types of filings studied in the literature, sourced from the SEC's EDGAR database. First are the report forms, such as the 10-K, 10-Q, and 8-K, which correspond to the annual report, quarterly report, and current report, respectively. These are used to study accounting transparency, financial performance metrics, disclosure behaviour and language, as well as event studies for the market reactions to the announced company results. The most commonly used form is the 10-K Annual Report. The second type is the proxy statements, such as DEFM14A, S-4, SC14D9, and PREM14A. These proxy statements include information on an upcoming merger and acquisition (M&A) transaction that requires the shareholders' vote. To cast an informed vote, a deal summary, background information, board recommendations, and financial insights are shared. Both categories of filings have demonstrated strong potential for textual analysis applications in corporate finance research. However, in combination with the SDC Platinum dataset, the DEFM14A filings are chosen due to the best data availability. This combination of datasets has also promised early results from Morgan (2018).

Despite the advances in sentiment analysis and prior research demonstrating the explanatory and predictive power of textual sentiment in economic and financial contexts, its potential within

M&A research remains underexplored. Existing studies provide valuable evidence regarding the relationship between sentiment from quantitative sources, such as news articles and proxy statements, and market reactions and deal outcomes. However, the current challenges in the literature stem from limited methodological consistency, fragmented data sources, and the absence of a streamlined analysis pipeline to capture the complex underlying dynamics of sentiment and M&A deal metrics. Bridging these gaps requires an integrated methodology that combines a generalised M&A data pipeline with NLP methods to assess how sentiment influences M&A deal outcomes.

2.3 Subgroup Discovery in Financial Context

Subgroup Discovery (SD) is a descriptive data analysis method that focuses on identifying meaningful subgroups, subsets of a larger dataset, that exhibit unique or unusual behaviour in relation to a target variable. Unlike traditional correlation or regression methods, which describe relationships across entire datasets with single features, SD targets these subgroups with the ability to use multiple features, revealing their distinctive characteristic patterns. This method combines techniques from machine learning, data mining, and statistics, proving especially useful for complex, heterogeneous, and non-linear economic datasets.

The concept of SD was first defined by Wrobel (1997), who extended earlier work by Klösgen on single-relation tasks into multi-relation SD. Foundational tooling and ideas for practical pattern discovery were already explored by Klösgen (1996). In 1999, Klösgen (1999) described the financial applications of SD, including market analysis. These scientific contributions formed the foundation for the subsequent expansion of knowledge discovery methods and descriptive artificial intelligence.

SD-related methods have been applied across various domains, including medicine and engineering. Yet, their use in economic and financial analysis remains scarce, largely due to the dominance of linear econometric data analysis methodologies and the lack of adaptable mining tools. This gap creates an opportunity for SD to generalise high-dimensional, heterogeneous M&A datasets and retrieve new insights by uncovering deeper, non-linear, context-dependent relations that traditional approaches fail to explain.

However, as datasets grow, the hypothesis space explodes, making exhaustive search infeasible and causing many highly similar subgroups to appear at the top of a ranked list. Subgroup set mining and heuristic selection strategies address these redundancy problems by explicitly trading off subgroup quality against diversity in the result set (Van Leeuwen and Knobbe, 2011). These more advanced SD strategies and methods are later implemented in an SD toolkit called SubDisc, which provides researchers with the freedom to configure generic SD algorithms to meet their local pattern discovery needs and handle a range of data types and multi-attribute targets (Knobbe et al., 2021). In SubDisc the redundancy problems are addressed through alternative beam selection strategies, such as description-based, cover-based and compression-based selection, which are designed to produce non-redundant subgroup sets rather than a purely quality-based top-k list. In this research, the standard configuration of SubDisc is used, that is, beam search combined with a quality measure and top-k selection. Concretely, the search proceeds as a breadth-limited beam search with Cortana Quality for nominal targets and explained variance for numeric targets, and at each depth the beam is filled by keeping the subgroups with the highest quality scores. This setting can be summarised as beam search plus quality measure (WRAcc or Cortana Quality) plus top-k selection.

Together with the quantitative core derived from the standardised SDC dataset and the quantitative features extracted with the sentiment analysis, the SubDisc subgroup discovery data mining tool will be added as the analytical core in this research. In concrete terms, each subgroup is described as a conjunction of simple conditions on deal or sentiment attributes (for example, "small cash-only deals with a friendly attitude"), and SD evaluates how strongly the corresponding subset of deals deviates from the global pattern for a chosen target variable, such as deal completion or target share price change. By combining these layers, an integrated analysis pipeline is created that can assess different perspectives on M&A deals, market behaviour, deal performance, and outcomes. The following chapter will describe the methodological approach of this research.

3 Methodology

This chapter will highlight the methodological choices for data collection, selection, and subgroup discovery to analyse both quantitative and qualitative data from 1,754 historical M&A deals sourced from the SDC Platinum dataset. Data collection involved retrieving relevant proxy statements (DEFM14A filings) linked to these deals. Feature engineering transformed raw data into appropriately formatted independent and dependent variables, with dependent variables serving as targets for subgroup analysis. Sentiment scores were quantitatively extracted from qualitative proxy statement texts using the NLP-based sentence-transformer model Paraphrase-mpnet-base-v2 (Transformers, 2024), which scored document embeddings to sentiment-related word groups via cosine similarity. The SubDisc tool was applied for subgroup discovery analysis, using the Cortana Quality measure with a beam search strategy for nominal data and explained variance with a beam-best strategy for numeric data. Across all targets, search parameters included a minimum significance threshold of 5%, refinement depths from 1 to 3, and coverage filters between 10% and 90%, ensuring the identification of meaningful patterns. These results are detailed for further discussion and visualised within a comprehensive research model diagram in Figure 1.

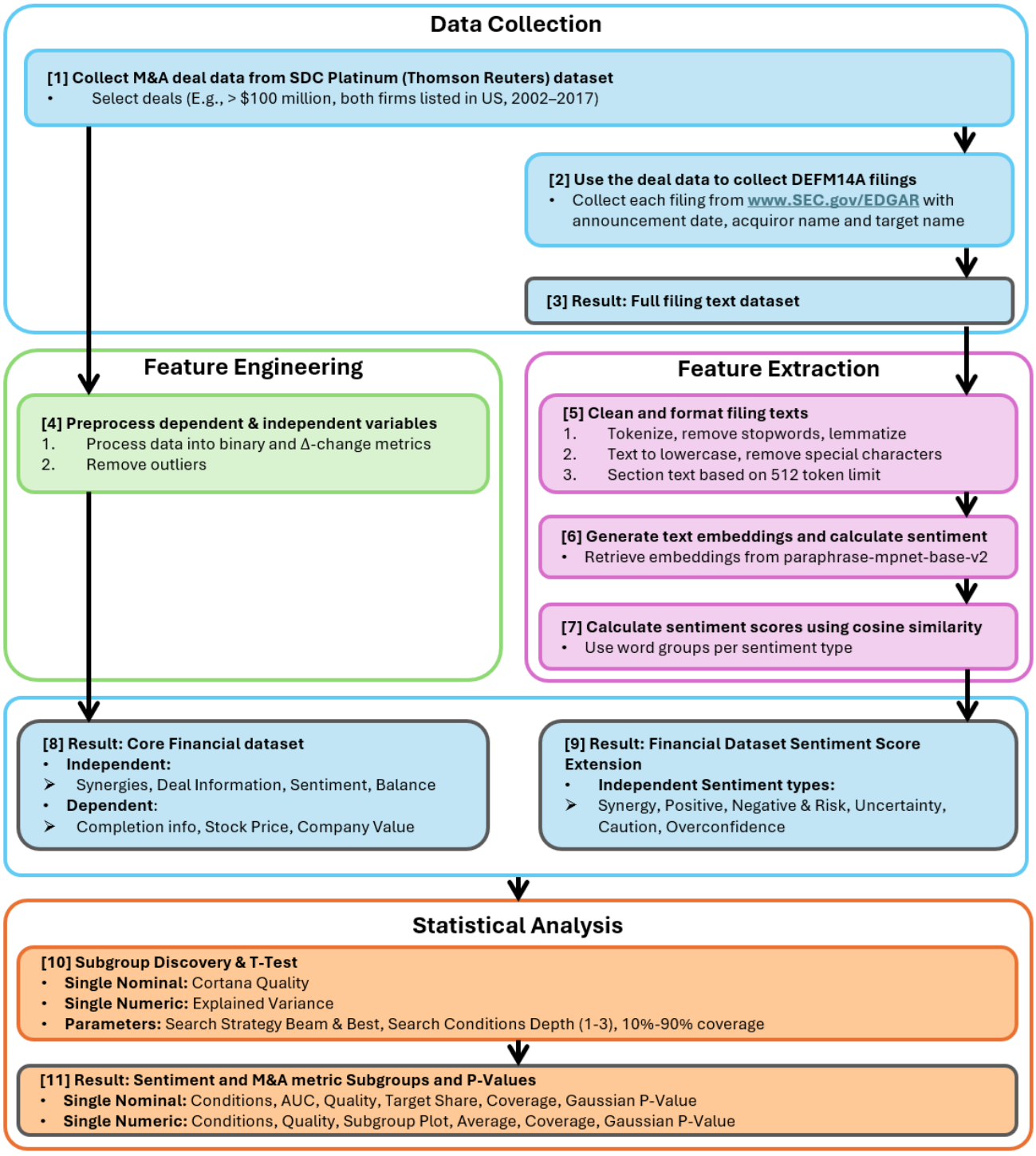

Figure 1: Research Model Diagram of the M&A Data Analysis Pipeline

3.1 Data Collection

For data analysis, any general or unique hand-picked data source will suffice; however, to ensure generalizability and openness to improvement in future work, the industry-standard and generally available SDC Platinum dataset is being used[1]. Via the LSEG workspace or academic portals, the dataset can be accessed and filtered. For this research, the filters in place are on M&A deals announced between 2002 and 2017 that are active and have a completed or withdrawn status. The companies involved in these deals are US-based and publicly traded, and the transaction value is in US Dollars, with a value exceeding $100 million. This has retrieved 1,754 M&A deals that include 167 variables, totalling almost 300,000 cells of raw input data.

The raw quantitative data can be used directly for the feature engineering steps described in Section 3.2. Given the information from the dataset, extra qualitative data surrounding the included deals will be hand-collected to provide a deeper analysis. For each unique M&A deal, retrieved as a row, in the SDC dataset, the corresponding DEFM14A filings will be retrieved using the target name, acquiror name and announcement date. The deal information is entered into the U.S. Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval System (EDGAR) and the related DEFM14A filing will be shown if available. Alongside the 1,754 deals, the corresponding 824 filings have been retrieved for deeper analysis. The lower sample size is due to either the company choosing another filing type, or the deal being withdrawn or transferred to another business entity. From here, feature extraction will be used to retrieve quantitative variables from the qualitative dataset, as described in Section 3.3, following the completion of feature engineering and feature extraction. The data are combined into a single, complete dataset, ready for Subgroup Discovery (SD) data analysis, as described in Section 3.4.

3.2 Feature Engineering

In this section, the feature engineering steps are described to retrieve correctly formatted and processed data for data analysis. This data will also be grouped as either independent variables or dependent variables. The dependent variables will be the targets during the subgroup discovery steps.

The SDC dataset encompasses a diverse range of quantitative deal characteristics, valuation measures, balance sheet metrics, and firm descriptors. The feature engineering process consists of three stages. The first stage is selecting the relevant and excluding the irrelevant data columns in the dataset. Each column contains M&A deal-related variables, which are split into independent variables, dependent variables, and excluded variables. Independent variables capture contractual, structural, and financial characteristics of the transaction, such as the percentage of shares acquired, deal type, leverage, and forecasted or historical financial ratios. Dependent variables represent outcome measures of deal performance or market reaction, such as acquiror and target stock prices, deal premiums, and valuation metrics at announcement and completion dates. Excluded variables were omitted since they provide limited analytical value or duplicate existing information. They were typically identifiers, such as company names, CUSIPs, static geographic fields or descriptive texts. In the second stage, the irregular columns, such as stock prices, are converted into percentage

[1]Access to the SDC Platinum dataset was facilitated by Assistant Professor Dr. A.H. Zohrehvand in March 2022.

change metrics by calculating the delta of the different time periods available in the dataset. Suitable nominal data is then converted into binary or Boolean metrics. This will standardise the inputs to create a more streamlined analysis and report process. In the third stage, the outliers were removed. When assessing the distribution of the variables, outliers that could skew the results were identified and removed. Ten outliers were identified, with their values being at least three times higher than the maximum value or three times lower than the minimum value, which followed them in the dataset. These outliers only occurred on the percentage change metrics, such as stock value, premium, and deal enterprise equity value. The complete list of raw variables which were included or excluded from the SDC Platinum dataset is shown in Appendix A.

3.3 Feature Extraction

In Section 3.3, the quantitative features will be extracted from the qualitative data of the hand-collected proxy statements using the NLP-based sentence-transformer model, paraphrase-mpnet-base-v2, to retrieve sentiment scores for the proxy statement texts. This is achieved by mapping the embeddings to word groups, such as positive and synergies, using cosine similarity.

From the SDC dataset, the announcement date of the deal and the names of the target and acquirer are retrieved and used in the public (SEC EDGAR) filing system. When filtering for DEFM14A filings, the related text files are being retrieved. These are stored in a data file for later use in the analysis.

Before the analysis, three extraction steps must be taken. First, the textual data is preprocessed and integrated into a dataset. Then, the texts and sentiment word groups are embedded using a transformer model. Finally, to retrieve the sentiment scores, the overlap between the sentiment groups and the transformed text embeddings is calculated using cosine similarity.

The DEFM14A filings are first cleaned by tokenising and lemmatising the text, converting it to lowercase, and removing all unique and non-alphabetic characters. Bird et al. (2009) also emphasise that these techniques help preserve the semantic core of words. An example for this processing output is the transformation of the text "Our Board considered the $50.00 per share in cash to be paid as merger consideration in relation to our Board's estimate" to "board considered per share cash paid merger consideration relation board estimate", where the sentiment sparse texts are removed. The conjugations of words are returned to their base form. After these preprocessing steps, the new dataset with transformed texts is retrieved. Table 1 shows the difference between the original form and the transformed form. It highlights the removal of approximately half of the words, resulting in an average of around 70,000 tokens used by the model per filing.

Table 1: Illustrative preprocessing: original vs transformed snippet

Original	**Transformed**
Our Board considered the $50.00 per share in cash to be paid as merger consideration in relation to our Board's estimate.	board considered per share cash paid merger consideration relation board estimate

For the second step of feature extraction, text embeddings are being retrieved from the filing

text and the dedicated sentiment themes. The paraphrase-mpnet-base-v2 sentence transformer model, developed by Reimers and Gurevych (2019), is chosen as it is one of the best ranking models based on comparative studies of different embedding models, such as the work from Issa et al. (2023), in terms of effectiveness (f1-score, recall, precision, MAP) and relative compute time. Another consideration in model choice is the token input limit. To meet the 512-token input limit of the transformer model paraphrase-mpnet-base-v, each filing has 70,000 words and is divided into roughly 140 sections of 512 tokens each. This division, shown in Table 2, ensures that the context of the text is preserved for analysis (Reimers and Gurevych, 2019; Li, 2010). Together with the embedded text, sentiment themes are used during the sentiment analysis. These themes are established with frequently occurring word groups within the texts. These word groups represent six sentiment dimensions, including positive, negative/risk, uncertainty, caution, overconfidence, and synergy. Each group contains representative terms frequently observed in DEFM14A filings, and the exact overview is shown in Appendix C.

The paraphrase-mpnet-base-v model transforms each text or word group into a 768-dimensional vector, where each dimension represents a unique feature of the word's meaning or context. The mathematical representation of a 768-dimensional $\mathrm{w}^{\rightarrow}$ vector is:

$$\mathbf{w} = [w_1, w_2, w_3, \ldots, w_{768}]$$

Then, the sentiment similarity between the filing texts and the word group sentiment is determined by calculating the cosine similarity, using scikit-learn (Pedregosa et al., 2011). Cosine similarity is a method that calculates the similarity between two vectors, ranging from -1 to 1. A score of 1 means the vectors are perfectly aligned, while -1 indicates they are opposites. The mathematical formula for cosine similarity is:

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \, \|\vec{B}\|}$$

Here, $\vec{A}$ and $\vec{B}$ are the vectors being compared, $\vec{A} \cdot \vec{B}$ is their dot product, and $\|\vec{A}\|$, $\|\vec{B}\|$ are the magnitudes.

To illustrate how a 768-dimensional vector space and cosine similarity work, consider the following sentence embedding and the word "positive" embedding:

"Projections indicate improved scaling and rising revenues" = [0.15, 0.40, . . . , 0.80]
positive = [0.12, 0.45, . . . , 0.78]

Given the two example vector representations, the scores are closely related. With cosine similarity, the average difference for each vector is calculated, and the more related the vectors are, the higher they score. Between these examples, as they are closely related, the cosine similarity would approach a similarity score of 1. When putting the embedding of the sentence in the context of the other word groups, the positive sentence would also empirically and hypothetically score high with the synergy word group and slightly positive on the overconfidence word group, where it would score negative on negative/risk, uncertainty and caution as the sentence doesn't contain any related words for that sentiment. The general approach of sentiment is used to capture the

semantic similarity between two different qualitative sources. This provides the ability to analyse the sentiment within the text.

The modified variables from the feature engineering and feature collection steps will be combined in a final dataset. An overview of these variables can be found in Appendix B. These variables will be used in the Subgroup Discovery data analysis as described in the next subsection.

3.4 Subgroup Discovery Data Analysis

In this section, the analysis of the full dataset collected during the previous steps is described. To assess the deeper relationships within the M&A data, the subgroup discovery (SD) data analysis method has been chosen due to its ability to discover interpretable patterns within complex data.

Conceptually, SD solves the following problem: given a dataset with descriptive attributes X and a target variable Y, find human-readable rules (subgroups) of the form *IF (conditions on X) THEN (the distribution of Y in this subgroup is noticeably different from the overall distribution).* Each candidate subgroup is evaluated by a quality measure that compares the target distribution inside the subgroup relative to the global target distribution, and the goal of the algorithm is to search the space of such rules and return the highest-quality ones. The found subgroups will help to assess patterns in the data.

In this research, SD is implemented using the SubDisc tool because it supports both nominal and numerical targets and allows the user to configure different quality measures and search strategies. SubDisc instantiates the general SD idea by generating simple one-condition subgroups, expanding them stepwise with additional conditions, and retaining only the best-scoring candidates at each step according to the chosen quality measure. The capabilities and workflow of SubDisc were initially proposed by Meeng and Knobbe (2011), which supported the analysis pipeline of this research greatly.

3.4.1 General data analysis steps

To identify these Subgroup Discovery patterns, several steps are required within the SubDisc tool. First, the dataset is imported, and a selection is made on which variables to include as input and which to use as targets. The input variables range from deal value, anticipation, and expected outcomes to deal type, context and sentiment. The targets will be split into two groups: nominal and numeric. These target variables share the same search strategy settings and search conditions but will need their own search quality measures. The Cortana quality measure used for nominal data is further described in subsection 3.4.2. The numeric quality measures and strategies, such as the use of explained variance and the numeric strategy, are described in subsection 3.4.3. The search strategies used for this research are the heuristic strategy type beam, with the search breadth-limited to 100 candidates, and with the numeric strategy set to 'best'. The search strategy type, beam search, is selected due to its ability to balance subgroup diversity and computational load. It will achieve this by efficiently pruning to the top-k and optimally forming candidates until the scores do not improve. In the context of M&A data, a diverse set of subgroup outcomes will be considered by the use of this strategy. Intuitively, the beam search algorithm starts from very

simple one-condition rules and iteratively adds extra conditions, but at each step it only keeps the best-scoring candidates in the "beam", so that the search remains computationally feasible while still exploring a rich set of potentially interesting subgroups. For the numeric attributes, the numeric strategy is set to 'best'. This will evaluate all possible splits for a numeric variable within a candidate subgroup. From here, the best-performing splits are further explored. This will have higher computational costs than the other options, bins or best-bins, but it is optimal for this data analysis.

Then, to retrieve relevant and interpretable results for both nominal and numerical target variables, the following search conditions are established. The coverage considered is set to cover a minimum subgroup of at least 176, which is 10% of the 1754 deals. Additionally, a maximum coverage of 0.9 is set, which will at most include a subgroup of 90% of the dataset. These thresholds are in place to ensure diverse subgroups that are both statistically significant and practically relevant, while also avoiding large subgroups that would mirror the overall dataset. A lower minimum coverage (for example 1% or 5%) would allow many very small subgroups that are numerically extreme but not robust and hard to interpret, whereas a higher maximum coverage (for example 95% or 100%) would mainly return almost global patterns with little contrast to the overall data. The 10% and 90% cut-offs therefore balance stability of the statistics and practical interpretability of the subgroups. In order to get diverse results for the following chapters, a refinement depth of 1 for single feature results and a depth of 2 and 3 for multi-feature results is set, and the corresponding results are registered. The depth will reveal combinations of conditions that yield an improved quality score. With higher depth comes more complex dimensions and higher amounts of output data to explore. A depth below the third level yields no better results and makes the analysis too complex.

When reviewing the SD results, a p-value is reported. This can be interpreted as the probability of observing a subgroup with at least the same difference in target share under the null hypothesis that the subgroup matches the full dataset's target distribution. In other words, the null hypothesis assumes the subgroup's target proportion equals the overall proportion, and the test assesses whether the observed difference arises from random variation, typically using a normal approximation based on subgroup size. A low p-value indicates the subgroup's deviation is unlikely due to chance. In the result tables, only subgroups with a p-value below 0.05 are reported, aligning with a conventional 5% significance threshold. This threshold draws from swap-randomisation using 100 candidates, serving as the minimum quality measure in the SD analysis. For p-value calculation, a Gaussian approximation compares the observed difference in means to its estimated standard error, signalling that a low p-value reflects a shift unlikely from random sampling alone.

3.4.2 Nominal data analysis steps

The nominal data indicate whether the deal is completed, indicated by $target = 1$, or withdrawn, indicated by $target = 0$. Out of the 1,754 deals, 1,551 (88.43%) were completed with $target = 1$. During SD, patterns are described by subgroup conditions. A hypothetical example is that small deals with high positive sentiment have a higher completion rate, expressed as follows: Deal Value $\leq 500 \wedge$ Positive Sentiment $\geq 0.1 \wedge$ Negative Sentiment ≤ 0.2. The conjunction of these conditions defines a subgroup within the data for which the completion rate can be compared to the global

baseline.

The relational validity and significance of these subgroups are evaluated using four metrics: target share, ROC curve area under the curve (AUC), p-value, and Cortana Quality. Among these, Cortana Quality is the optimization criterion that the SubDisc search procedure aims to maximise for nominal targets, where the target share, ROC AUC, and p-value are reported post hoc to interpret the resulting subgroups. The target share represents the proportion of completed deals within the subgroup compared to the overall completion rate in the dataset. The ROC curve summarises the ability of the subgroup rule to discriminate between completed and withdrawn deals by plotting the true positive rate (TPR) against the false positive rate (FPR) at different thresholds. The area under this curve (AUC) quantifies how well the subgroup distinguishes between positive and negative instances. In the AUC scoring, a random classifier would score 0.5, whereas a perfectly discriminative rule would score 1.

Finally, the Cortana Quality measure from SubDisc is considered. This measure is a refined version of the widely used quality measure Weighted Relative Accuracy (WRAcc), introduced in early SD literature by Lavrač et al. (1999). WRAcc evaluates the predictive value of a subgroup relative to a baseline average and is based on the relative success within the subgroup, the overall success rate in the data, and the subgroup size. Its mathematical definition is:

$$\text{WRAcc} = \underbrace{\frac{|\text{Pos} \cap \text{Cond}|}{N}}_{\text{relative positive rate inside subgroup}} - \underbrace{\frac{|\text{Pos}|}{N}}_{\text{overall positive rate}} \cdot \underbrace{\frac{|\text{Cond}|}{N}}_{\text{relative subgroup size}},$$

where N is the total number of instances in the data, $|\text{Pos} \cap \text{Cond}|$ is the number of positive instances inside the subgroup, $|\text{Pos}|$ is the number of positives in the full dataset, and $|\text{Cond}|$ is the subgroup size. Cortana Quality is a scaled variant of WRAcc, which means that the value range of Cortana Quality does not depend on the prior distribution of the target and is fixed between -1 and 1 in contrast to WRAcc. This scaling ensures comparability across subgroups regardless of class imbalance. In addition, Cortana Quality explicitly penalises subgroups that are very similar in coverage or quality to already discovered subgroups and rewards subgroups that provide strong contrast with the global distribution. As a result, it favours a smaller and more diverse set of subgroups than WRAcc alone, reducing redundancy in the final result list and supporting interpretation (Knobbe, 2022) .

3.4.3 Numerical data analysis steps

When working with numeric attributes in SD, the challenge arises in finding a suitable balance between search runtime, result quality, subgroup size, and redundancy. The framework proposed by Meeng and Knobbe (2021) in their systematic review, together with their suggestions for future work, is integrated into the numerical data analysis steps of this research to choose an appropriate numeric search strategy.

The numerical targets considered in the SD analysis include *time to completion*, *stock price change of the acquiror*, *stock price change of the target*, *the average of both*, and the changes in *equity value* and *enterprise value*. The relational validity and significance of the resulting subgroups are evaluated using four metrics: the average target value in the subgroup, the subgroup model plot,

the p-value, and the explained variance. Among these, explained variance is used as the quality measure that SubDisc maximises during the search; the other three metrics are used to interpret and assess the relevance of the discovered subgroups.

For the first metric, the average target value in the dataset is compared with the average within the subgroup. For example, when evaluating stock price changes, an increased average stock price in the subgroup relative to the global average suggests that the conditions defining the subgroup are associated with more positive stock market reactions. The second metric is the subgroup model plot. This is used to assess whether the distribution of the target within the subgroup differs meaningfully from the global distribution. If the subgroup curve closely follows the same shape and location as the base distribution, then the subgroup behaves very similarly to the overall data and has limited additional explanatory value; by contrast, a subgroup whose distribution is shifted or has a clearly different shape is more informative. Then the p-value is addressed, where the lower value is related to a more significant result.

The final metric, explained variance, often represented as R^2, is used as a measure in numerical data analysis to assess how well the subgroup describes or predicts the variation in the target numeric variable. It quantifies the proportion of the variance in the data explained by the subgroup conditions compared to the total variance of the target variable. Its mathematical representation is:

$$R^2 \;=\; 1 - \frac{\sum_i (t_i - f_i)^2}{\sum_i (t_i - \bar{t})^2},$$

Here, t_i denotes the observed target values, f_i the values predicted by the subgroup model (for instance, the subgroup mean), and $\bar{t}$ the global mean of the target. A value of $R^2 = 1$ indicates that the subgroup conditions perfectly explain the variation in the target, whereas $R^2 = 0$ implies that the subgroup does not improve upon simply predicting the global average. This metric is widely used in numerical data analysis to evaluate the effectiveness of models or subgroup conditions in explaining changes in target variables such as stock prices or deal completion times (Knobbe et al., 2017).

The steps considered for the different target variables will help to retrieve patterns with statistically significant insights into M&A deal outcomes. The p-values for significance, nominal and numeric scores and the distribution and ROC plots for pattern interpretation are stored. These outcomes are further explored in Chapter 4, where the results are presented, and evaluated in Chapter 5, along with the discussion.

4 Results

Chapter 4 presents the findings from the subgroup discovery analyses on both nominal and numeric targets. For these targets, single-feature and multi-feature combinations are evaluated, yielding statistically significant results. The single-feature results of the nominal deal completion outcome are dominated by sentiment variables that indicate a high relational value to the deal completion outcomes based on their high ROC curve AUC scores and Cortana Quality, and total target share being almost a perfect predictor, nearing the score of 1, with a high coverage. The multi-feature results reveal more diverse and complex patterns, achieving a very high AUC and quality score. These patterns were primarily observed in conditions that favoured smaller, friendly, non-hostile, and non-defensive deals, accompanied by a positive pre-deal anticipation. For numeric outcomes, both single and multi-feature analyses generally retrieved significant but modest explained variance quality scores. Despite these lower quality scores, the patterns of less complex deals, such as small, friendly, and cash-only, are also evident in the numeric results. However, the stock price change outcomes for the target presented meaningful findings.

4.1 Nominal Variable Outcomes

For the first set of results, the nominal data analysis for the deal completion variables is being considered. The implications of the patterns found will help create a deeper understanding of the influences on the success of a deal. The nominal targets considered are deal completion = 1 for completed deals and deal completion = 0 for withdrawn deals.

This section highlights the significant patterns of sentiment that dominate the single-feature results. For the multi-feature results, an AUC score of higher than 0.8 is found. Smaller deals with a positive run-up and friendly, non-hostile, and defensive attitudes yield the highest-scoring subgroups for deal completion.

4.1.1 Deal completion single feature outcomes

Table 2 shows the top 10 single-feature results for the nominal deal completion target. These results primarily show sentiment-related conditions, which are presented in ranges indicating whether they are higher or lower than a certain amount; these ranges correspond to a subgroup of the data. The strongest ranges are positive sentiment and uncertainty, being represented by the upper part of the sentiment range, and caution, being represented by the lower part of the range. The other sentiments, including negative and risk, synergy, and overconfidence, also retrieve relevant results metrics, but in this dataset, they don't show a distinction by covering the same subgroups.

Table 2: Deal Completion — Single-Feature Results (Top 10)

Coverage	Quality	Target Share	p-Value	Condition
820	0.37	0.96	< 0.001	`Sent_Positive` ≥ 0.036
801	0.37	0.97	< 0.001	`Sent_Uncertainty` ≥ 0.108
817	0.37	0.96	< 0.001	`Sent_Caution` ≤ 0.244
824	0.36	0.96	< 0.001	`Sent_Synergy` ≤ 0.288
824	0.36	0.96	< 0.001	`Sent_Synergy` ≥ 0.153
824	0.36	0.96	< 0.001	`Sent_Positive` ≤ 0.132
824	0.36	0.96	< 0.001	`Sent_Negative & Risk` ≤ 0.226
824	0.36	0.96	< 0.001	`Sent_Negative & Risk` ≥ 0.114
824	0.36	0.96	< 0.001	`Sent_Uncertainty` ≤ 0.205
824	0.36	0.96	< 0.001	`Sent_Caution` ≥ 0.128

Here, all results are significant, with p-value scores of less than 0.001 and follow the < 0.001 notation convention as stated by Zhu (2016). The target share of these subgroups is generally 8% higher, with 96% of the target subgroup data being positive, compared to an average target positive rate of 88.43%. The Cortana Quality for all sentiments is higher than 0.36. The AUC score from the ROC curve in Figure 2 shows that most of the sentiment-related points are located around comparable locations, indicated by the corresponding FPR and TPR scores.

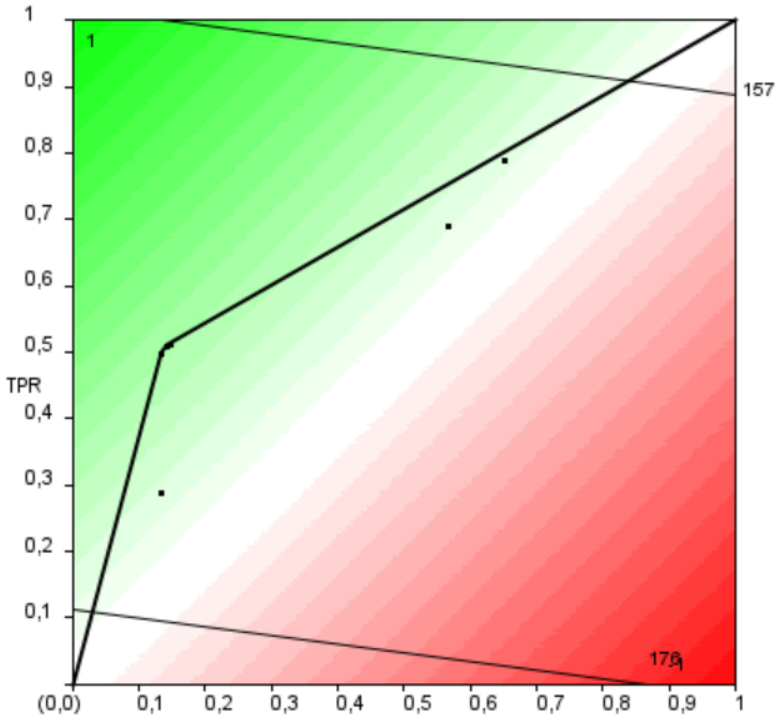

Figure 2: Single-feature deal completion results

The highest-scoring conditions are the positive sentiment subgroup (≥ 0.035) with an FPR of 14% and a TPR of 51%, and the uncertainty sentiment subgroup (≥ 0.108) with an FPR of 13% and a TPR of 50%. The other points in the ROC curve of Figure 2 yield a high target share but have a low Cortana Quality of below 16%; these conditions will be further explored during the multi-feature analysis. The single-feature analysis of incomplete deals yielded no strong results, as the Cortana Quality is also below 0.16, with a weaker target share difference of between 2% and 6% and a lower AUC score.

4.1.2 Deal completion multi-feature outcomes

When combining multiple conditions, more intricate patterns will arise, and more diverse and higher-scoring subgroups can be formed. The multiple feature analysis combines the results of depth 2 and depth 3. Any results at higher depth did not yield improved outcomes. Across these settings, the beam search produced a few dozen candidate multi-feature subgroups per depth, of which only the top-scoring subgroups are reported in the result tables.

During this analysis, new patterns were identified with Cortana Quality scores ranging from 0.37 to 0.49. These multi-feature patterns show a significantly higher coverage of over 1500 instances compared to the approximately 800 coverage seen in single-feature results. Although the target shares for multi-feature subgroups are lower, their increased coverage and greater diversity result in overall higher Cortana Quality. The best scoring patterns include a friendly attitude, smaller deal value and a positive run-up. The patterns of depth 2 show a mix of less complicated deals, ranging from non-hostile and non-defensive to non-rumoured, non-buyout or private, as well as

non-complex deals and those without financial sponsor involvement. The depth 3 patterns mainly focus on friendly, non-hostile and non-defensive deal attitudes in combination with deal size and run-up. In terms of diversity, these high-quality subgroups exhibit different logical descriptions (for instance, by adding or removing conditions on deal complexity, financial sponsor involvement or consideration structure), but they still overlap strongly in coverage: most of them pick out largely the same subset of small, friendly, cash-driven deals with favourable sentiment. As discussed later in the methodological limitations, this moderate diversity in covered transactions is consistent with the use of standard top-k beam search without explicit diversity constraints.

All reported multi-feature results are significant, with p-values lower than 0.001. The ROC curves and their AUC scores are 0.805 and 0.808 for depths 2 and 3, respectively.

Compared to the single-feature ROC curve in Figure 2, Figure 3 shows more relevant subgroups discovered in depth 2 and those that are more concentrated in depth 3. The sentiment-related subgroups are concentrated around a very low FPR of 10% and a TPR of 50%, and show slight improvement compared to the single-feature results in Figure 2. At these higher depths, new subgroups are identified and displayed near the top, close to the threshold limit, with very high TPR rates and slightly lower FPR rates. When experimentally exceeding the 90% maximum coverage threshold limit, no improved AUC score is found.

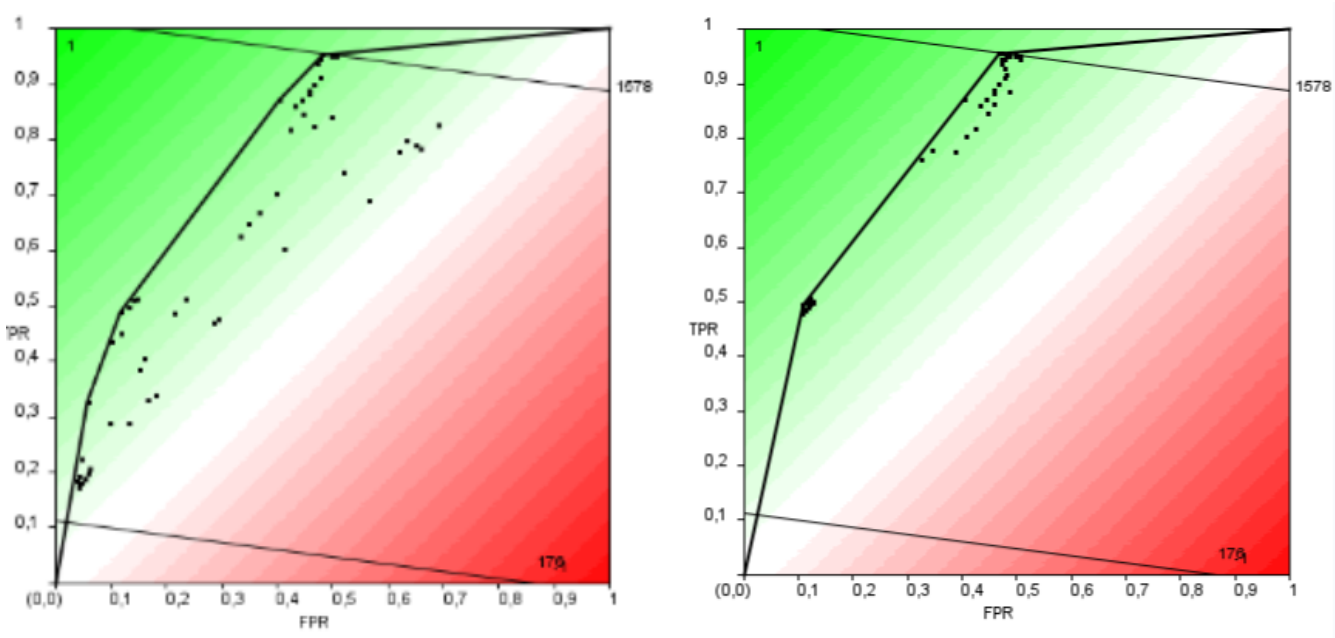

Figure 3: Multi-feature deal completion results with depth = 2 (left) and depth = 3 (right)

As shown in Figure 3 on the right with the depth 3 ROC curve plot, the best scoring conditions can be classified into two groups: the sentiment subgroups in the middle left of the plot and the run-up subgroups in the top middle of the plot. The conditions with the best AUC scores are for the sentiment subgroups, the smaller deals with lower caution sentiment and friendly attitude noted with `Friendly Attitude = '1'` $\wedge$ `Deal Value` $\leq$ `31594.557` $\wedge$ `Sent_Caution (number)` $\leq$ `0.262784` with a FPR of 10% and TPR of 50%. For the run-up subgroup at the top it shares smaller deals with a friendly attitude. However, for mostly positive run-ups, which are noted as `Friendly Attitude`

= '1' ∧ Deal Value ≤ 31594.557 ∧ Run up Premium d% 1dvs1w (%pp) ≥ -28.52. Depth 3 generally validates the patterns of depth 2 with some minor quality enhancements.

4.2 Numeric Variable Outcomes

For the second set of results, the numerical variables are being analysed. The implications of the patterns found will help create a deeper understanding of the factors influencing deal outcomes. The numerical targets considered are related to merger success and synergy metrics, such as time to completion, equity value, and enterprise value, as well as market reactions, including acquiror, target, and average stock price changes. These outcomes will be measured based on coverage, p-value, the explained variance quality measure and the difference from the average. This difference will highlight either a positive or negative influence on the mix of subgroups.

The analyses initially yield low explained variance scores, which aligns with findings from the existing literature on complex deal data. Nevertheless, clear patterns emerge in the results. In particular, both time to completion and target share price change show distinct differences between subgroups. The time to completion appears related to factors such as deal value, type of transaction, and consideration structure. Deals involving smaller, non-complex, or cash-only transactions typically have a shorter time-to-completion window. In addition, the results indicate positive market reactions for targets involved in smaller, non-complex, and friendly deals, which often show higher share price increases on the announcement day. These increases range up to 8.3% higher than the average stock price increase for the targets. Despite the relatively low explained variance, the results show that specific combinations of deal characteristics and sentiment can still highlight relevant deal outcomes.

4.2.1 Time to completion outcomes

The analysis of time to completion discovered more than a thousand significant subgroups. The coverage was also relatively high in the context of the non-null status of the variables. Table 3 presents the top 10 conditions and their corresponding scores for the single-feature analysis. It will show a steep drop-off in quality, as well as the differences in the averages. The quality starts at 0.12 and drops below 0.1 after the third found subgroup. This is something to consider when assessing the implications of the results. The differences in average do reveal relevant patterns across the various conditions, with an average of 138 days to completion for the deals in the dataset. First, there are patterns in the consideration structure and deal type, where complex and stock-related deals generally require a higher number of days to complete. This is noted in the cash-only consideration structure, which shows a lower average, along with `DT Stock Swap or Collar = 0` and `DT Merger or Tender = 1`. These patterns can be observed in the conditions outside the top 10, but they will have even lower quality. Additionally, the deal value showed a clear distinction, where smaller deals have a shorter time to completion compared to the average, and larger deals have a longer time to completion. The sentiment-related conditions have a low explained variance score; however, a clear pattern emerges, where lower positive sentiment and lower synergy sentiment scores are associated with a longer time to complete.

Table 3: Top 10 single-feature time to completion subgroups

Coverage	Quality	Average	Avg. Diff.	St. Dev.	p-Value	Conditions
711	0.12	98.10	-40.38	72.93	< 0.001	Consideration Structure = Cash Only
918	0.12	106.67	-31.81	74.64	< 0.001	DT Stock Swap or Collar = 0
836	0.12	172.62	34.14	95.10	< 0.001	DT Stock Swap or Collar = 1
1493	0.09	150.52	12.04	89.52	< 0.001	DT Merger or Tender = 0
261	0.09	71.68	-66.80	66.20	< 0.001	DT Merger or Tender = 1
1517	0.07	129.02	-9.46	82.35	< 0.001	Deal Value $\leq$ 4981.292
237	0.07	205.35	66.88	114.31	< 0.001	Deal Value $\geq$ 5072.965
193	0.06	202.60	64.13	128.10	< 0.001	Expected Synergies Before Taxes $\geq$ 75.0
185	0.04	191.57	53.09	131.42	< 0.001	Sent_Positive $\leq$ 0.07284776
290	0.04	178.18	39.70	113.89	< 0.001	Sent_Synergy $\leq$ 0.22806005

When reviewing the multi-feature results, it is found that most results that score better than the single-feature results will show a shorter time to completion, rather than a longer one. Most of the described single-feature patterns are reaffirmed by the multi-feature results, but new patterns can also be addressed. A positive run-up, friendly attitude and non-hostile or non-defensive deal types also relate to a shorter time to completion. This suggests that factors that simplify M&A deals, such as a friendlier attitude or a simpler cash transaction rather than a stock trade, will have a positive impact on the deal completion duration.

To summarise the results, subgroups with the conditions of being small, friendly, and non-complex merger or tender deals that are cash-only show a lower time to completion. Lower positive and synergy sentiment is associated with a longer time to complete, while having a low explained variance. Generally, stock swap or collar-related deals take longer to complete.

4.2.2 Stock price outcomes

For the stock price changes, single-feature and multi-feature subgroup discovery have been applied to the announcement-day stock price changes of the acquiror, target, and the average of both. Generally, the results that were significant with a p-value < 0.05 showed an explained variance of less than 0.03, which makes the results unable to explain the stock price changes fully. The single-feature results have yielded fewer than 10 results, whereas the multi-feature results have identified a large number of subgroups.

With the acquiror stock price changes, no directly relevant subgroups can be addressed. All significant results indicate a decrease in stock price, which aligns with the data showing a negative average acquiror stock price.

The target has an average stock price change of 6%, and the single feature results, with conditions related to smaller cash-only merger or tender deals, show a 2–4% increase in stock prices. For

multi-feature results, many high stock price increases ranging from 7% to 8.3% above the average are discovered. These are generally shown within conditions related to smaller non-complex deals with a friendly attitude. Noteworthy is that a smaller run-up of ≤ -3.55 is also associated with positive stock price changes. The best scoring conditions are on depth 3 for `Run up Premium d% 1dvs1w (%pp)` $\leq$ `-3.55` $\wedge$ `DT Complex deals = '0'` $\wedge$ `Deal Value` $\leq$ `625.445`, with an average stock price of 14.4% and, a quality of 0.03 and a significant p-value of below 0.001.

The average stock price change doesn't reveal any new patterns and generally provides a milder representation of the target and acquiror stock price changes.

To summarise, the explained variance is relatively low with a max of 0.03. This suggests that the observed patterns do not account for the stock price changes exclusively. However, the target stock price change variable results promise changes greater than 7% compared to the average stock price, which is a significant increase that has high implications for investors.

4.2.3 Equity value and enterprise value outcomes

Both equity value and enterprise value change outcomes show little significant results and have a high standard deviation, ranging from 10% to 30%. The explained variance is just like the stock price outcomes below a value of 0.02.

Overall, the results highlight the predictive power of certain target variables, such as deal completion and target stock price change, as well as the relevance of specific input variables, including sentiment and those related to deal complexity. This will enhance the scope of future research, where target stock price analysis is expected to yield more relevant and robust results than equity value or enterprise value. Also, the conditions are based on the independent input variables. Out of the 152 independent variables in the SDC Platinum dataset, only 23 were selected for the final analysis, and only those related to complexity and sentiment yielded relevant results.

5 Discussion

Chapter 5 evaluates the main findings of this research and interprets their implications. First, the results are discussed in the context of the research scope and the main research question. The contributions to theory and their practical implications are then described. Following this, the limitations of the research are highlighted, and a recommendation for future work, based on the findings is provided.

5.1 Summary

The primary objective of this research is to systematically identify which language features and deal metrics are most strongly associated with M&A deal completion and related market reactions, and to formalise a reproducible analysis pipeline that future studies can extend. This has been achieved by applying feature engineering to the SDC Platinum dataset in combination with an innovative method for feature extraction from DEFM14A filings from SEC.gov. This has created a quantitative final dataset to which modern methods, such as Subgroup Discovery, have been applied to convert complex patterns within the data into understandable and interpretable characteristics. The central research question driving these objectives is: Which language and deal characteristics are most strongly associated with the success of M&A transactions and related market reactions, such as changes in share prices or company value? The research is of an exploratory nature to be as relevant and applicable as possible for future work, and it helped gather new insights into the relevance of general deal characteristics, the impact of sentiment categories such as positivity and risk, and the possibilities of combining textual characteristics with traditional financial variables. The limitations and areas for improvement within the proposed research pipeline have also been identified, ranging from dataset choices to transformer model capabilities. Methodologically, this research uses a combination of a standardised, quantitative dataset (SDC Platinum) with additional qualitative textual data (DEFM14A proxy statements). The application of Subgroup Discovery enables the discovery of deeper, non-linear relationships that traditional methods often overlook (Atzmueller, 2015). The main findings show that sentiments, particularly positive and uncertainty-related language, show a substantial association with deal completion. Complexity in deals, such as deal size and the attitude of the parties involved, also appears to show significant trends: smaller, friendlier and less complex deals are more likely to be completed and lead to more favourable market reactions. In the case of numerical outcomes, such as time to completion and share price changes, the results show a less intense but still consistent pattern in line with the above findings and add the importance of the deal consideration structure, where cash-only deals are considered less complex than stock-related deals, which relates to improved time to completion and stock prices. Although the explained variance for numerical indicators remains limited, this actually underlines the complexity of M&A transactions and underlying dynamics. The proposed research methodology represents a significant step toward a better understanding and standardisation of analysis within this domain, which also invites refinement and expansion in future research.

5.2 Limitations

This section provides an in-depth reflection on the limitations and key considerations associated with the data and analysis methods used in this research. By zooming in on the data collection, feature extraction, and the datasets, insights are provided into the reliability and possible weaknesses of the basis on which the analyses are built. The SDC Platinum dataset is one of the most widely used sources for M&A research, offering an extensive collection of deal data. Nevertheless, this dataset has several limitations as inaccuracies may exist in the recording of deal statuses, time spans and financial characteristics. Firstly, its focus on American transactions limits the generalisability of the results to international or cross-border M&A, where other dynamics may come into play. The decision to analyse only US deals from the period 2002-2017 narrows the scope, leaving trends or insights outside these boundaries undescribed. Market conditions, regulations and the economic context change over time, which affects the comparability of deals across the years. Macroeconomic influences, which undoubtedly impact deal outcomes, can only be incorporated into the model to a limited extent, introducing potential confounders. For example, the 2008 financial crisis has been included in the dataset but not directly addressed in the deal data, resulting in a negative skew in the data. The size of the dataset with 1,754 deals is relatively limited, especially given the complexity of M&A transactions, which can limit the statistical power to discover intricate patterns. Furthermore, the data is influenced by the reliability of the sources, which may contain errors in timing, reporting and definition. The dates of ranking are included in the dataset for validation, while the downsides described in the evaluation by Barnes et al. (2014) are also taken into consideration. Finally, the choice of which variables to include influences the results; some potentially relevant variables may not be available or may have been deliberately excluded due to methodological choices. All variables should have been included to leverage the potential of the SDC Platinum dataset fully. However, this would exceed the research scope and compromise the general interpretability of the data analysis. The dataset does not contain complete information for all 1.754 deals, resulting in NULL values and missing data. This can lead to a bias in the dataset and limitations in its representativeness for the broader population of M&A transactions. Missing data can also complicate statistical analyses and affect the validity of conclusions. The DEFM14A filings text has only been collected for 824 deals, where most are for completed deals. The withdrawn deals generally have no filings, which will skew the deal completion results for this data. Important synergy-related variables, such as expected pro forma EPS after synergies and expected synergies before taxes, had only around 300 data value inputs. The data with the highest occurrence rate, such as deal types and attitudes or run-up, have been identified in the subgroups. Choices in data preprocessing, such as excluding outliers or implementing data cleaning steps, can also lead to bias or the loss of important information. More standardised rules for data cleaning would help. The manual collection of SEC proxy filings is labour-intensive and can lead to inconsistencies in the dataset. When there was doubt about the relevance of the DEFM14A filing to the deal in the SDC dataset, the data was excluded, reducing the sample size. The filing texts themselves often contain a mix of relevant and less relevant text passages, making it difficult to isolate the most informative pieces with absolute precision. The sentiment distribution has been addressed during analysis. For example, in the positive and negative sentiment groups, as shown in Figure 4 below, the trends of sentiment distribution over the filing are divided into 20 sections. A clear peak around grouped section 6, which represents sections 42 to 49 in the 140 chunked sections. These sections generally contain the board's rationale and considerations section of the filing. A

rise in sentiment is evident from the grouped section 12 in the plot, towards 20, as represented by the chunked section 84, which spans from 140 in the processed filing texts. This may be due to the change in language resulting from the addition of appendices to the filing. No general information can be retrieved from these standard sections compared to the grouped Section 12.

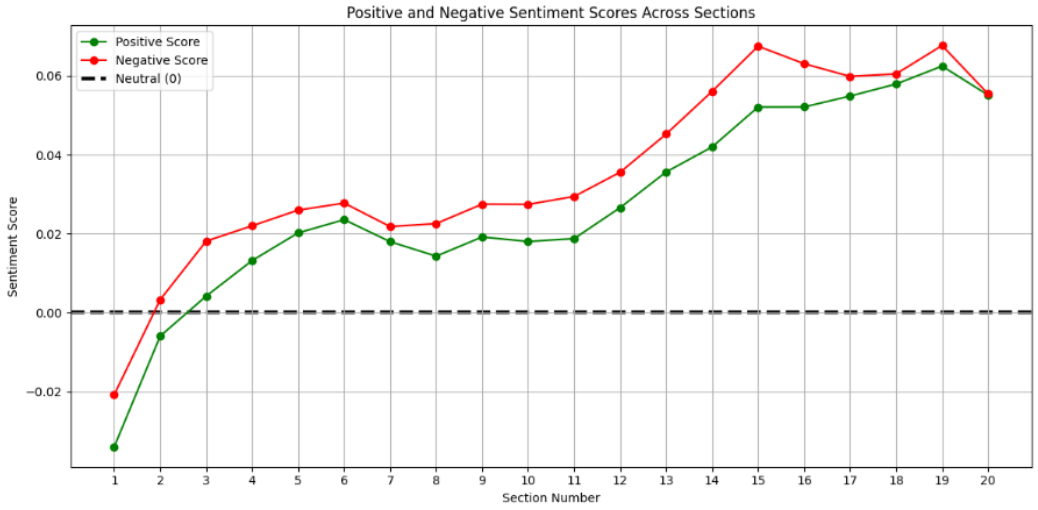

Figure 4: Positive and negative sentiment distribution across grouped sections of the filings

Also relevant is Figure 5, which explores the sentiment distribution for the synergy-themed embeddings. Here, the same peak around grouped section six can be found, as well as a peak at grouped section 2, which is representative of the filings in its summary chapter. The influence of the sentiment around the appendix is less meaningful, as lower synergy-related texts are expected.

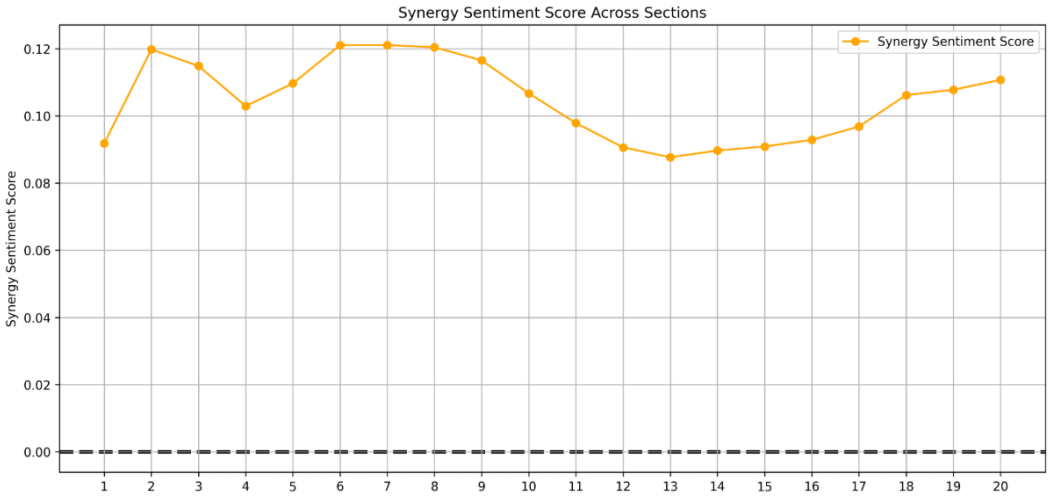

Figure 5: Synergy sentiment distribution across grouped sections of the filings

These plots represent the general distribution of all the different sentiment themes. However, as seen, the sentiment can vary, and this research uses the average sentiment score for all filings instead

of examining the intricacies within each filing. Therefore, analysing only the high-sentiment-related sections would benefit the results. Improvements could include the automated or semi-automated selection of document sections relevant for sentiment analysis. The choice of sentiment extraction model, in this case a transformer-based model such as paraphrase-mpnet-base-v2, has limitations. For example, there are restrictions on the token limit per input, which means that long texts must be split up, potentially causing a loss of context. Additionally, the selection of sentiment categories and word groups is crucial; an incorrect choice can lead to an overestimation or underestimation of the actual sentiment impact. As a result, interpretation remains cautious, and results should be considered indicative. In evaluating the results, overlap was observed between the scores for negative, risk, and uncertainty. In future research, either a better word group can be manually chosen or dynamically selected based on its relatedness score to the word group. Other adjustments include selecting a more optimised NLP model. This can be a model adjusted to financial language, such as Finance BERT, or a model that addresses context better with a larger context window, like the Bigbird model, which boasts a context window of up to 4.096 tokens compared to the current 512 tokens in this research. The current LLM developments in 2025 will also provide the ability to analyse context windows larger than 1 million tokens; however, the downsides of LLM analysis, such as hallucinations and the black-box origin of the models, need to be considered.

A further methodological limitation concerns the specific subgroup discovery configuration used in this research. Although Section 2.3 briefly introduced subgroup set mining as a way to obtain non-redundant sets of patterns, the currently applied subgroup discovery methodology is based on standard top-k mining in SubDisc. The search is performed with a beam search, a quality measure (Cortana Quality for nominal targets and explained variance for numeric targets), and a beam that is filled at each depth by keeping the subgroups with the highest quality scores. This beam search plus quality measure plus top-k selection setup is effective for finding high-scoring subgroups, but it can lead to relatively low diversity in the final result list, as can be seen in Table 2, where several conditions describe highly overlapping subgroups. The subgroup set mining literature proposes several concrete strategies to improve this situation: description-based beam selection avoids adding subgroups whose descriptions differ only marginally from patterns already in the beam, cover-based beam selection actively penalises overlap in the tuple cover of subgroups by weighting quality with a measure of how many new transactions a candidate subgroup contributes, and compression-based strategies use model-based similarity to favour subgroups that add genuinely new information about the data (Van Leeuwen and Knobbe, 2011). In addition, one can mine a large pool of candidate subgroups and then apply a post-selection step that optimises a diversity criterion such as cover redundancy on top of quality. These approaches would likely increase the diversity of the subgroups found for deal completion and stock price reactions, and they directly address the observation that some of the current result tables still contain quite homogeneous patterns. Such subgroup set mining strategies were not implemented within the scope of this research for two main reasons: first, making principled choices for the diversity parameters, such as the cover weighting factor and the desired trade off between quality and diversity, would require an additional round of systematic experiments on the relatively large and diverse M&A dataset. Second, introducing diversity constraints changes the behaviour of the search algorithm in nontrivial ways, which would have required additional space to analyse and explain the differences between the standard top-k results and the diversified subgroup sets. For similar reasons, this research also chose to analyse one nominal target and one numeric target at a time instead of modelling

multiple targets jointly as a single complex target: treating deal completion, stock price reactions and valuation changes simultaneously would require an Exceptional Model Mining setup with multi-target quality measures and more intricate model classes, as well as substantially more experimentation to interpret the resulting joint patterns. Given the scope of the bachelor project, the focus was therefore on a clear and interpretable baseline configuration with separate nominal and numeric targets; a natural extension of this work is to rerun the analysis with description based or cover based beam selection and with multi-target quality measures, and to compare whether these subgroup set mining variants over complex targets yield more diverse but still interpretable patterns that add value beyond the current results.

To summarise, while limitations are present, many of these can be addressed or mitigated by the proposed methodological pipeline from this research. The goal of this research was to explore the potential associations within the mix of M&A related variables, given the scarcity of available literature.

5.3 Contributions

5.3.1 Relation to existing literature

This research contributes to the existing literature on mergers and acquisitions by offering new insights into the predictive value of language and deal characteristics for the success of M&A transactions and related market reactions. Whereas the work of Andrade et al. (2001) lays an essential foundation for understanding the economic drivers and outcomes of M&A, this research builds on this foundation by providing an integrated analysis of both qualitative and quantitative data. By combining SDC Platinum data with textual information from DEFM14A filings and analysing it using modern methods such as Subgroup Discovery, an analysis pipeline is created that can better capture and explain the complexity of deal outcomes. The methodological foundations of sentiment analysis, as laid down in the work of Loughran and McDonald (2011), are expanded by applying NLP techniques to qualitative M&A data. This research examines how sentiment categories, such as positivity, risk, and overconfidence, in proxy statements play a crucial role in describing and explaining deal completion and market reactions. By further developing previous studies, such as those on the qualitative implications of text on deal outcomes by Morgan (2018) on filing texts and Hajek and Henriques (2024) on news sentiments, it becomes possible to establish stronger links between qualitative data and quantitative outcomes than was previously possible. The use of transformer-based models, such as Paraphrase-mpnet-base-v, represents a significant technological advancement in processing complex text data within financial contexts. This research demonstrates that the natural language processing of qualitative sources is not only methodologically applicable but also contributes significantly to predicting economic outcomes in M&A. This approach provides a robust and reproducible foundation for future studies that combine qualitative and quantitative data in the context of M&A research.

5.3.2 Practical implications and future work

The findings of this study underscore the importance of accurate and conscious communication in M&A processes. The boards of companies can better tailor their communication to market-relevant sentiments, thereby providing greater insight into risks and increasing confidence in the transactions. The research outcomes also provide investors with relevant tools to assess deal completion, which in turn correlates with higher deal returns, and a significant 7% target stock price increase under the identified conditions. This provides investors and researchers with tools to evaluate language use and deal characteristics as additional indicators in investment decisions and risk analyses. The development of an integrated and standardised data analysis methodology opens up new perspectives for refinements and follow-up research. Future studies could extend this methodology to different markets, time periods and additional data sources to increase generalisability. Future research utilising more advanced machine learning techniques and improved datasets could further enhance the predictive power and interpretation of the qualitative and quantitative data pipeline integration, as well as explore new relationships.

6 Conclusion

This research explored the relationships between quantitative and qualitative M&A deal data and their outcomes, introducing an integrated data analysis pipeline for future investigations. Motivated by the increasing complexity and volume of M&A data, as well as advances in NLP and data mining techniques, the study aimed to identify language features and deal metrics that are most strongly associated with deal completion and market reactions, such as changes in stock price and company value. Key findings revealed that sentiment, particularly positive and uncertainty-related language, as well as smaller, non-complex, friendly, and cash-based deals, are strongly associated with successful M&A outcomes. Additionally, the current findings reveal an 8% significant increase in the target stock price around the announcement.

The primary research question, which language features and deal metrics are most strongly associated with M&A deal completion and market reaction, was addressed through analyses of general deal metrics, sentiment categories (positivity, risk, overconfidence, etc.), and the integration of textual features with traditional financial variables. Results demonstrated that sentiment metrics dominated single-feature predictions of deal completion, and combined multi-feature patterns revealed that smaller, friendlier, and less complex deals have higher likelihoods of completion and more favourable market responses. Numerical targets such as time to completion and stock price changes showed consistent but lower explanatory power, highlighting the intrinsic complexity of M&A transactions.

Building on foundational prior work such as Andrade et al. (2001) on economic drivers of M&A and Loughran and McDonald (2011) on domain-specific sentiment analysis, this research contributes a novel, reproducible research methodology that merges standardised SDC Platinum deal data with qualitative textual insights extracted from SEC DEFM14A filings using transformer-based NLP models. The use of Subgroup Discovery further allowed the discovery of nuanced, non-linear patterns that traditional linear methods often miss. These methodological advancements expand the theoretical understanding of how communication tone and deal characteristics jointly influence M&A success and market behaviour.

Limitations of this study include data availability constraints, notably the incomplete proxy filings for withdrawn deals, potential biases from focusing on U.S.-based deals between 2002 and 2017, and a relatively low explained variance for numerical outcomes. Methodological challenges, such as token limits in transformer models, imperfect sentiment category selection, and the exclusion of some potentially relevant variables, also inform a cautious interpretation. Nonetheless, the flexibility of the analysis enables mitigation of these limitations and iterative refinement.

In final remarks, this research establishes a significant stepping stone for integrating qualitative and quantitative analyses in M&A studies, providing a reproducible methodology that enhances the prediction and understanding of deal outcomes. It offers practical implications for corporate communication strategies, investor decision-making, and policy considerations. Future work can expand the created data analysis pipeline by incorporating larger datasets, additional qualitative and quantitative sources, enhanced NLP models with broader context windows, and more sophisticated machine learning methods, thereby further unlocking the complex dynamics that shape mergers and acquisitions.

References

Andrade, G., Mitchell, M., and Stafford, E. (2001). New evidence and perspectives on mergers. *Journal of Economic Perspectives*, 15(2):103–120.

Atzmueller, M. (2015). Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49.

Bachelier, L. (1900). Théorie de la spéculation. *Annales scientifiques de l'Ecole normale supérieure*, 17:21–86.

Barnes, B. G., Harp, N. L., and Oler, D. K. (2014). Evaluating the sdc mergers and acquisitions database. *Financial Review*, 49(4):793–822.

Bennett, B. and Dam, R. A. (2019). Merger activity, stock prices, and measuring gains from m&a. *Stock Prices, and Measuring Gains from M&A*.

Berens, P., Cranmer, K., Lawrence, N. D., von Luxburg, U., and Montgomery, J. (2023). Ai for science: an emerging agenda. *arXiv preprint*.

Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with python. *O'Reilly Media*.

Bollaert, H. and Delanghe, M. (2015). Securities data company and zephyr, data sources for m&a research. *Journal of Corporate Finance*, 33:85–100.

Feng, S. and Wang, Z. (2020). Mutual influences and trends of global economic growth and global mergers and acquisitions. 214:01032.

Gaughan, P. A. (2010). *Mergers, Acquisitions, and Corporate Restructurings*. John Wiley & Sons.

Hajek, P. and Henriques, R. (2024). Predicting m&a targets using news sentiment and topic detection. *Technological Forecasting and Social Change*, 201:123270.

Institute for Mergers, Acquisitions and Alliances (IMAA) (2025). Number & value of m&a worldwide. https://imaa-institute.org/mergers-and-acquisitions-statistics/#Worldwide. Accessed: 18 Oct 2025.

Issa, B., Jasser, M. B., Chua, H. N., and Hamzah, M. (2023). A comparative study on embedding models for keyword extraction using keybert. In *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pages 40–45. IEEE.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant.

Klösgen, W. (1999). Applications and research problems of subgroup mining. In *International Symposium on Methodologies for Intelligent Systems*, pages 1–15. Springer.

Knobbe, A. (2022). Quality measure. https://github.com/SubDisc/SubDisc/wiki/Quality-Measure. Accessed: 27 November 2025.

Knobbe, A., Gobeil, J., van Dijk, R., and Palenstijn, W. J. (2021). Subdisc: A subgroup discovery toolkit.

Knobbe, A., Orie, J., Hofman, N., van der Burgh, B., and Cachucho, R. (2017). Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, 31(6):1872–1902.

Lavrač, N., Flach, P., and Zupan, B. (1999). Rule evaluation measures: A unifying view. In *International Conference on Inductive Logic Programming*, pages 174–185. Springer.

Li, F. (2010). The information content of forward-looking statements in corporate filings: A naive bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*.

Martynova, M. and Renneboog, L. (2008). A century of corporate takeovers: What have we learned and where do we stand? *Journal of Banking & Finance*, 32(10):2148–2177.

Meeng, M. and Knobbe, A. (2011). Flexible enrichment with cortana – software demo. In *Proceedings of BeneLearn 2011: The Twenty Annual Belgian-Dutch Conference on Machine Learning*, pages 117–119. BeneLearn.

Meeng, M. and Knobbe, A. (2021). For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, 35(1):158–212.

Morgan, P. E. (2018). Predictive power? textual analysis in mergers & acquisitions. *Marriott Student Review*, 2(2).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP*, pages 3982–3992.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21–23, 1963, Spring Joint Computer Conference*, pages 241–256.

Thomson Reuters (2010). Sdc platinum factsheet.

Transformers, S. (2024). paraphrase-mpnet-base-v2 (revision e6981e5).

Van Leeuwen, M. and Knobbe, A. (2011). Non-redundant subgroup discovery in large and complex data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer.

Zhu, W. (2016). p¡ 0.05,¡ 0.01,¡ 0.001,¡ 0.0001,¡ 0.00001,¡ 0.000001, or¡ 0.0000001.... *Journal of sport and health science*, 5(1):77–79.

A Appendix A: Variable overview

The variable overview in Appendix A shows that only a subset of the 167 raw SDC Platinum fields has been used in the subgroup discovery analysis. Conceptually, the included variables are those that are central to the research questions and well suited for pattern mining: core deal structure and complexity indicators (deal value, deal type, consideration structure, attitudes, financial sponsor involvement, run-up premiums, and industry group similarity), synergy-related expectations, and the outcome variables that capture deal success and market reactions (deal completion, time to completion, stock price changes, and equity and enterprise value changes). In contrast, a large group of variables has been excluded because they are non-informative identifiers, extremely sparse or minimally distinct in this sample, near duplicates of already included fields, or belong to detailed financial ratio and forecasting families that would substantially increase dimensionality without being directly needed for this exploratory baseline. This pruning was necessary to keep the hypothesis space manageable, to reduce redundancy, and to maintain interpretability of the resulting subgroups.

Variable	**Variable type**	**Used in analysis**	**Reason for exclusion**
M&A Deal Number	Independent	No	Non-informative identifier
Target Name	Independent	No	Non-informative identifier
Acquiror Name	Independent	No	Non-informative identifier
Acquiror Nation	Independent	No	Minimally distinct data
Target Nation	Independent	No	Minimally distinct data
Announcement Date	Independent	Yes	Used in data collection
Deal Value	Independent	Yes	Used as SD attribute
Target Public Status	Independent	No	Minimally distinct data
Acquiror Public Status	Independent	No	Minimally distinct data
Form of the Deal	Independent	No	Minimally distinct data
Deal Status	Dependent	Yes	Processed to Boolean and used as "Deal Completed"
Deal Status	Dependent	No	Double entry
Form of the Deal	Independent	No	Double entry
Deal Type	Independent	Yes	Processed into separate Boolean deal type attributes
Consideration Structure	Independent	Yes	
Deal Attitude	Independent	Yes	
Deal Purpose	Independent	No	Qualitative entry made by SDC
Deal Synopsis	Independent	No	Qualitative entry made by SDC
Rank Date	Independent	No	Only relevant for data validation
Completion Date	Dependent	Yes	

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Effective Date	Dependent	No	Out of scope, relevant for future research
Deal Value	Independent	No	Double entry
Deal Value inc. Net Debt of Target	Independent	No	Out of scope, relevant for future research
Deal Value exc. Assumed Liabilities	Independent	No	Out of scope, relevant for future research
Deal Value as-of Date	Independent	No	Out of scope, relevant for future research
Percent of Shares Acquired	Dependent	No	Out of scope, relevant for future research
Percent of Shares Acquiror is Seeking to Purchase	Independent	No	Out of scope, relevant for future research
Percent of Shares Held by Acquiror at Announcement	Independent	No	Out of scope, relevant for future research
Percent of Shares Owned after Transaction	Dependent	No	Out of scope, relevant for future research
Share Price Paid by Acquiror	Independent	No	Out of scope, relevant for future research
Initial Offer Price	Independent	No	Out of scope, relevant for future research
Financial Sponsor Involvement (Y/N)	Independent	Yes	
Financial Sponsor SDC Cusip	Independent	No	Non-informative identifier
Financial Sponsor Name	Independent	No	Non-informative identifier
Financial Sponsor Role	Independent	No	Qualitative entry made by SDC
Cross Border Deal Flag (Y/N)	Independent	No	Minimally distinct data
Currency of Deal	Independent	No	Minimally distinct data
Expected Accretive/ Dilutive Type	Independent	No	Out of scope, relevant for future research
Expected Accretive Date	Independent	No	Out of scope, relevant for future research
Expected Pro Forma EPS After Synergies	Independent	Yes	
Expected Synergies Before Taxes	Independent	Yes	
Expected Synergy Date	Independent	Yes	
Status of MMC Ruling	Dependent	No	Minimally distinct data

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Target Termination Fee	Independent	No	Out of scope, relevant for future research
Acquiror Termination Fee	Independent	No	Out of scope, relevant for future research
Deal Market Value to EBITDA	Independent	No	Out of scope, relevant for future research
Deal Market Value to EBIT	Independent	No	Out of scope, relevant for future research
Deal Market Value to Pre-tax Income	Independent	No	Out of scope, relevant for future research
Deal Market Value to Net Income	Independent	No	Out of scope, relevant for future research
Deal Enterprise Value to Sales	Independent	No	Out of scope, relevant for future research
Deal Enterprise Value to EBITDA	Independent	No	Out of scope, relevant for future research
Deal Enterprise Value to EBIT	Independent	No	Out of scope, relevant for future research
Deal Enterprise Value to Net Income	Independent	No	Out of scope, relevant for future research
Deal Enterprise Value to Net Assets	Independent	No	Out of scope, relevant for future research
Offer Price to EPS	Independent	No	Out of scope, relevant for future research
Premium % 1 Day Prior to Announcement	Independent	Yes	
Premium % 1 Week Prior to Announcement	Independent	Yes	
Premium % 4 Weeks Prior to Announcement	Independent	Yes	
Related SDC Deal Number	Independent	No	Non-informative identifier
Related Deal Type	Independent	No	Out of scope, relevant for future research
Acquiror SDC Cusip	Independent	No	Non-informative identifier
Acquiror CIDGEN	Independent	No	Non-informative identifier
Acquiror Nation	Independent	No	Double entry
Acquiror Subregion	Independent	No	Minimally distinct data
Acquiror Region	Independent	No	Minimally distinct data

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Acquiror Macro Industry	Independent	No	Out of scope, relevant for future research
Acquiror Mid Industry	Independent	No	Out of scope, relevant for future research
Acquiror TRBC Industry	Independent	No	Out of scope, relevant for future research
Acquiror TRBC Industry Group	Independent	Yes	
Acquiror TRBC Business Sector	Independent	No	Out of scope, relevant for future research
Acquiror TRBC Economic Sector	Independent	No	Out of scope, relevant for future research
Acquiror Primary SIC	Independent	No	Out of scope, relevant for future research
Acquiror Public Status	Independent	No	Double entry
Acquiror State	Independent	No	Out of scope, relevant for future research
Acquiror Stock Exchange	Independent	No	Out of scope, relevant for future research
Acquiror Ticker Symbol	Independent	No	Non-informative identifier
Acquiror Business Description	Independent	No	Qualitative entry made by SDC
Acquiror Ultimate Parent Name	Independent	No	Non-informative identifier
Acquiror Ultimate Parent Nation	Independent	No	Minimally distinct data
Acquiror Ultimate Parent Region	Independent	No	Minimally distinct data
Acquiror Ultimate Parent Subregion	Independent	No	Minimally distinct data
Acquiror Stock Price on Announcement Day	Dependent	Yes	
Acquiror Stock Price on 1 Day After Announcement	Dependent	Yes	
Investor SDC Cusip	Independent	No	Non-informative identifier
Investor Names	Independent	No	Non-informative identifier
Investor Group Flag (Y/N)	Independent	No	Minimally distinct data
Investor Public Status	Independent	No	Minimally distinct data

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Buyside: Sovereign Wealth Fund Involvement	Independent	No	Minimally distinct data
Acquiror Ultimate Parent Primary SIC	Independent	No	Out of scope, relevant for future research
Acquiror Ultimate Parent Macro Industry	Independent	No	Out of scope, relevant for future research
Acquiror Ultimate Parent Mid Industry	Independent	No	Out of scope, relevant for future research
Investor Ultimate Parent SDC Cusip	Independent	No	Non-informative identifier
Investor Ultimate Parent Names	Independent	No	Non-informative identifier
Investor Ultimate Parent Public Status	Independent	No	Minimally distinct data
Acquiror Net Sales LTM	Independent	No	Out of scope, relevant for future research
Acquiror EBIT LTM	Independent	No	Out of scope, relevant for future research
Acquiror EBITDA LTM	Independent	No	Out of scope, relevant for future research
Acquiror Pre-tax Income LTM	Independent	No	Out of scope, relevant for future research
Acquiror Net Income LTM	Independent	No	Out of scope, relevant for future research
Acquiror Earnings Per Share LTM	Independent	No	Out of scope, relevant for future research
Currency of Acquiror Financials	Independent	No	Minimally distinct data
Date of Acquiror Financials	Independent	No	Only relevant for data validation
Target Name	Independent	No	Double entry
Target SDC Cusip	Independent	No	Non-informative identifier
Target CIDGEN	Independent	No	Non-informative identifier
Target Nation	Independent	No	Double entry
Target Subregion	Independent	No	Minimally distinct data
Target Region	Independent	No	Minimally distinct data
Target Macro Industry	Independent	No	Out of scope, relevant for future research

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Target Mid Industry	Independent	No	Out of scope, relevant for future research
Target TRBC Industry	Independent	No	Out of scope, relevant for future research
Target TRBC Industry Group	Independent	Yes	
Target TRBC Business Sector	Independent	No	Out of scope, relevant for future research
Target TRBC Economic Sector	Independent	No	Out of scope, relevant for future research
Target Primary SIC	Independent	No	Out of scope, relevant for future research
Target State	Independent	No	Out of scope, relevant for future research
Target Stock Exchange	Independent	No	Out of scope, relevant for future research
Target Ticker Symbol	Independent	No	Non-informative identifier
Target Business Description	Independent	No	Qualitative entry made by SDC
Target Number of Employees	Independent	No	Out of scope, relevant for future research
Target Ultimate Parent Name	Independent	No	Non-informative identifier
Target Ultimate Parent Nation	Independent	No	Minimally distinct data
Target Ultimate Parent Subregion	Independent	No	Minimally distinct data
Target Ultimate Parent Region	Independent	No	Minimally distinct data
Target Stock Price on Announcement Day	Dependent	Yes	
Target Stock Price on 1 Day After Announcement	Dependent	Yes	
Sellside: Financial Sponsor Activity Flag (Y/N)	Independent	No	Out of scope, relevant for future research
Sellside: Sovereign Wealth Fund Involvement	Independent	No	Minimally distinct data
Target Ultimate Parent Primary SIC	Independent	No	Out of scope, relevant for future research
Target Ultimate Parent Macro Industry	Independent	No	Out of scope, relevant for future research
Target Ultimate Parent Mid Industry	Independent	No	Out of scope, relevant for future research

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Net Sales LTM	Independent	No	Out of scope, relevant for future research
EBIT LTM	Independent	No	Out of scope, relevant for future research
EBIT 3 Year Growth (%)	Independent	No	Out of scope, relevant for future research
EBIT 5 Year Growth (%)	Independent	No	Out of scope, relevant for future research
EBITDA LTM	Independent	No	Out of scope, relevant for future research
EBITDA Margin LTM	Independent	No	Out of scope, relevant for future research
Pre-tax Income LTM	Independent	No	Out of scope, relevant for future research
Net Income LTM	Independent	No	Out of scope, relevant for future research
Earnings Per Share LTM	Independent	No	Out of scope, relevant for future research
Cash and Short Term	Independent	No	Out of scope, relevant for future research
Total Assets	Independent	No	Out of scope, relevant for future research
Short Term Debt	Independent	No	Out of scope, relevant for future research
Net Debt	Independent	No	Out of scope, relevant for future research
Total Liabilities	Independent	No	Out of scope, relevant for future research
Total Debt	Independent	No	Out of scope, relevant for future research
Common Equity	Independent	No	Out of scope, relevant for future research
Equity Value at Announcement	Dependent	Yes	
Equity Value at Effective Date	Dependent	Yes	
Enterprise Value at Announcement	Dependent	Yes	

Continued on next page

Variable	Variable type	Used in analysis	Reason for exclusion
Enterprise Value at Effective Date	Dependent	Yes	
Book Value per Share	Independent	No	Out of scope, relevant for future research
Currency of Target Financials	Independent	No	Minimally distinct data
Date of Target Financials	Independent	No	Only relevant for data validation
Forecasted Net Sales Current Fiscal Year	Independent	No	Out of scope, relevant for future research
Forecasted Net Sales Next 12 Months	Independent	No	Out of scope, relevant for future research
Forecasted EBIT Current Fiscal Year	Independent	No	Out of scope, relevant for future research
Forecasted EBIT Next 12 Months	Independent	No	Out of scope, relevant for future research
Forecasted EBITDA Next 12 Months	Independent	No	Out of scope, relevant for future research
Forecasted EBITDA Current Fiscal Year	Independent	No	Out of scope, relevant for future research
Forecasted EBITDA Year Two	Independent	No	Out of scope, relevant for future research
Forecasted Pre-tax Income Current Fiscal Year	Independent	No	Out of scope, relevant for future research
Forecasted Pre-tax Income Next 12 Months	Independent	No	Out of scope, relevant for future research
Forecasted Pre-tax Income Year Two	Independent	No	Out of scope, relevant for future research
Forecasted Net Income Current Fiscal Year	Independent	No	Out of scope, relevant for future research
Forecasted Net Income Next 12 Months	Independent	No	Out of scope, relevant for future research
Forecasted Net Income Year Two	Independent	No	Out of scope, relevant for future research

B Appendix B: Used variables in Subgroup Discovery

Appendix B summarises the final set of derived variables that enter the subgroup discovery step, including whether they are treated as independent or dependent, and whether they originate from the raw SDC Platinum fields, from feature engineering (such as percentage changes and time intervals), or from the sentiment feature extraction on DEFM14A filings.

Variable name	**Sample size**	**Cardinality**	**Variable type**	**Dependency**	**Source**
Deal Value	1754	1752	numeric	Independent	Data collection: Deal Value
Expected Pro Forma EPS After Synergies (number)	304	288	numeric	Independent	Data collection: Expected Pro Forma EPS After Synergies
Expected Synergies Before Taxes (number)	349	98	numeric	Independent	Data collection: Expected Synergies Before Taxes
Run up Premium d% 1dvs1w (%pp)	1752	1365	numeric	Independent	Feature engineering: Premium % 1 Day Prior to Announcement, Premium % 1 Week Prior to Announcement, Premium % 4 Weeks Prior to Announcement
Time to expected synergy (days)	317	290	numeric	Independent	Feature engineering: Announcement Date, Expected Synergy Date
Financial Sponsor Involvement (TRUE/FALSE)	1754	2	binary	Independent	Feature engineering: Financial Sponsor Involvement
Friendly Attitude	1754	2	binary	Independent	Feature engineering: Deal Attitude
DT Stock Swap or Collar	1754	2	binary	Independent	Feature engineering: Deal Type
DT Not Applicable	1754	2	binary	Independent	Feature engineering: Deal Type
DT Merger or Tender	1754	2	binary	Independent	Feature engineering: Deal Type
DT Rumored Deal	1754	2	binary	Independent	Feature engineering: Deal Type
DT Hostile or defensive	1754	2	binary	Independent	Feature engineering: Deal Type
DT Buyout or private	1754	2	binary	Independent	Feature engineering: Deal Type
DT Complex deals	1754	2	binary	Independent	Feature engineering: Deal Type
Consideration Structure (category)	1754	8	nominal	Independent	Feature engineering: Consideration Structure
Industry Group Same (TRUE/FALSE)	1754	2	binary	Independent	Feature engineering: Acquiror TRBC Industry Group, Target TRBC Industry Group
Sent_Synergy (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date
Sent_Positive (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date

Continued on next page

Variable name	Sample size	Cardinality	Variable type	Dependency	Source
Sent_NegativeRisk (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date
Sent_Uncertainty (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date
Sent_Caution (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date
Sent_Overconfidence (number)	824	823	numeric	Independent	Feature extraction: Target Name, Acquiror Name, Announcement Date
Deal Completed (TRUE/FALSE)	1754	2	binary	Dependent	Feature engineering: Deal Status
Time to Completion (days)	1551	328	numeric	Dependent	Feature engineering: Announcement Date, Completion Date
Acquiror Stock Price Change 1 Day After Announcement (%)	1733	1584	numeric	Dependent	Feature engineering: Acquiror Stock Price on Announcement Day, Acquiror Stock Price on 1 Day After Announcement
Target Stock Price Change 1 Day After Announcement (%)	1710	1496	numeric	Dependent	Feature engineering: Target Stock Price on Announcement Day, Target Stock Price on 1 Day After Announcement
Average Stock Price Change (%)	1693	1609	numeric	Dependent	Feature engineering: Acquiror Stock Price Change 1 Day After Announcement (%), Target Stock Price Change 1 Day After Announcement (%)
Equity Value change at effective date (%)	1547	1315	numeric	Dependent	Feature engineering: Equity Value at Announcement, Equity Value at Effective Date
Enterprise Value change at effective date (%)	1506	1320	numeric	Dependent	Feature engineering: Enterprise Value at Announcement, Enterprise Value at Effective Date

C Appendix C: Sentiment word groups

Sentiment Group	Keywords
Sent_Positive	growth, expansion, leadership, innovation, superior, unmatched, best-in-class, leading, record, breakthrough, unprecedented, extraordinary
Sent_NegativeRisk	risk, challenge, difficulty, weakness, obstacle, exposure, loss, liability, downturn, headwind, volatility, slowdown
Sent_Uncertainty	uncertainty, unpredictability, ambiguity, fluctuation, instability, doubt, unforeseeable, unclear
Sent_Caution	caution, prudent, measured, carefully, conservatively, mitigate, safeguard, contingency, reserve
Sent_Overconfidence	confident, assured, guaranteed, inevitable, certain, committed, definitely, no doubt, without risk
Sent_Synergy	synergy, integration, combined, consolidation, cost savings, efficiency, scale, accretive, EBITDA, margin

Table 6: Sentiment word groups used in the analysis