Decoding Corporate Culture: Harnessing Natural Language Processing to Unlock Organizational Insights at Abcam's Life Sciences Enterprise

UCL Business Analytics

11 August 2024

Word Count: 13,048

Abstract

This dissertation project addresses the challenge of analysing organizational culture through employee feedback at Abcam, a leading life sciences company. Despite the wealth of information in employee surveys, extracting meaningful insights from unstructured text data presents significant challenges. This study offers an innovative solution aimed at automating the analysis of employee comments and enhancing human resource management decision-making.

The primary objective is to develop a multi-label classification model specifically tailored for categorizing employee feedback based on Abcam's six cultural dimensions: Purpose, Customer Focus, Innovation, One Team, Growth and Development, and Diversity & Well-being. This approach leverages natural language processing (NLP) techniques and incorporates various word embedding models including Bag of Words (BoW), BERT, as well as machine learning algorithms such as Random Forests, Support Vector Machines (SVM), Neural Networks, and RoBERTa to effectively analyse employee comments and provide accurate classifications.

Findings indicate that a cross-encoder architecture using RoBERTa, when fine-tuned with weight adjustment for class imbalance and hyperparameter optimization, performs the best. It achieves an accuracy of 0.5476, demonstrating a 10% improvement across all metrics compared to the base model. However, these results are limited to the specific dataset from Abcam, as the project scope focused exclusively on this company's engagement survey data.

This dissertation demonstrates the potential of using NLP and machine learning to enhance organizational culture analysis. It also suggests future areas for research, including expanding the labelled dataset, exploring larger language models, and integrating the tool into existing HR systems. Although the implemented model is not fully integrated into Abcam's HR platform and currently exists as a standalone solution, it showcases a practical approach that could be adapted for broader commercial use in the life sciences industry and beyond.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Abcam

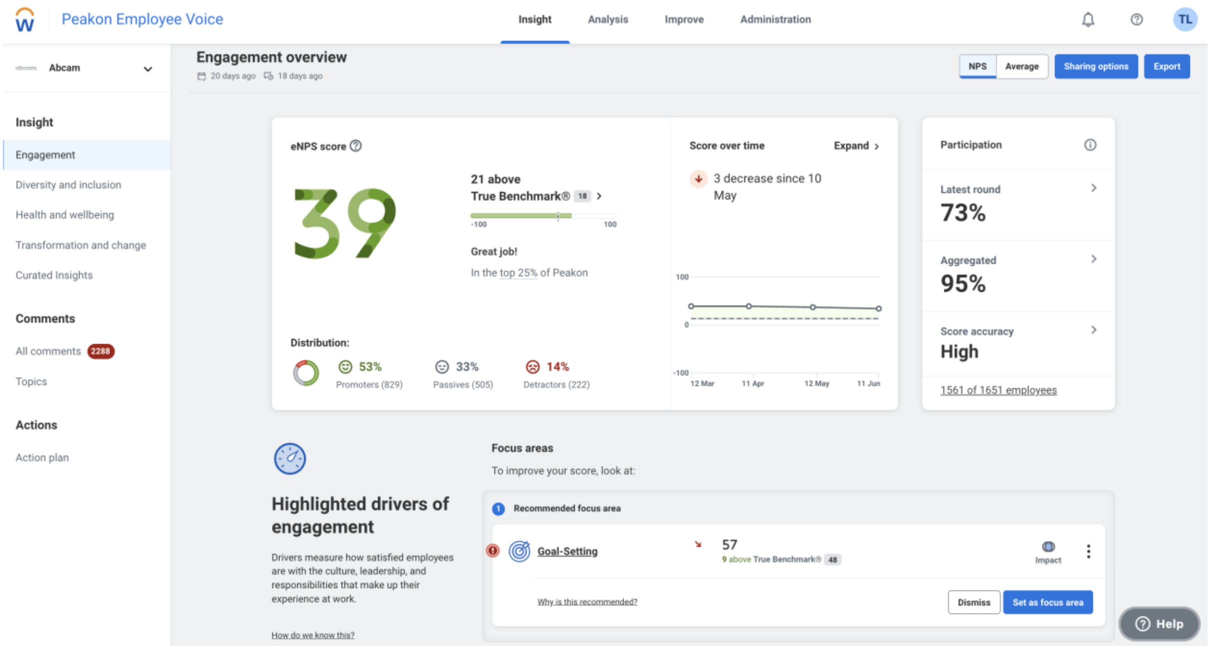
Abcam is a globally recognized life sciences company that has established itself as a leader in the research tools and reagents market. Founded in 1998 by academics from Cambridge University, Abcam has expanded its offerings beyond antibody supply to become a versatile provider of essential research tools (Abcam, 2022).

With its headquarters in Cambridge, UK, Abcam strategically expanded its global presence by establishing offices in key locations such as the United States, China, Australia, the Netherlands, and Singapore. This extensive international network enables Abcam to effectively cater to a diverse customer base across academia, industry, and clinical settings worldwide.

Abcam's main goal, as stated by the company itself, is to help life scientists achieve their mission more efficiently and enable scientific breakthroughs. They offer a vast range of products including primary and secondary antibodies, ELISA kits and matched antibody pairs, multiplex immunoassays, cell and tissue imaging tools, cellular and biochemical assays, as well as lysates. With an extensive Catalog of approximately 90,000 products available on the market today, Abcam has made a significant impact on scientific research with their products being cited in nearly half of all life science publications.

Abcam's workforce comprises around 1,500 employees globally, with 750 located at their Cambridge headquarters (Abcam, 2022). This team of scientists, technicians, and business professionals work together to drive innovation and maintain Abcam's position at the forefront of the life sciences industry.

In order to manage such a huge number of employees, Abcam utilized Peakon as one of their HRM tools; Peakon Employee Voice (Figure 1) from Workday is a cutting-edge platform designed to boost employee engagement and organizational performance using intelligent listening technology. Peakon enables employees to confidentially share their feedback regarding their experiences, expectations, and overall well-being by collecting this feedback through surveys and comments. (Workday, 2023).



Engagement overview
20 days ago 18 days ago

eNPS score
39
21 above True Benchmark® 18
-100 100
Great job!
In the top 25% of Peakon
Distribution:
53% Promoters (829)
33% Passives (505)
14% Detractors (222)

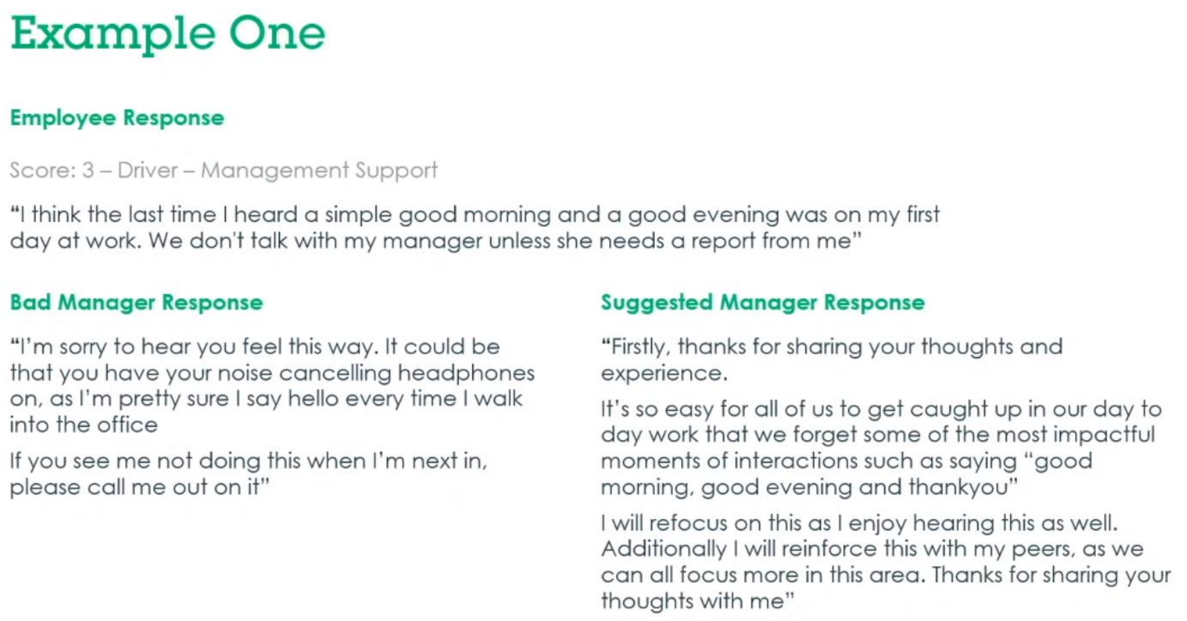
Score over time
Expand
3 decrease since 10 May

Participation
Latest round 73%
Aggregated 95%
Score accuracy High
1561 of 1651 employees

Highlighted drivers of engagement
Drivers measure how satisfied employees are with the culture, leadership, and responsibilities that make up their experience at work.
How do we know this?

Focus areas
To improve your score, look at:
Recommended focus area
Goal-Setting 57
9 above True Benchmark® 48
Impact
Why is this recommended?
Dismiss
Set as focus area

Figure 1 Overview of Peakon



Example One

Employee Response

Score: 3 – Driver – Management Support

"I think the last time I heard a simple good morning and a good evening was on my first day at work. We don't talk with my manager unless she needs a report from me"

Bad Manager Response

"I'm sorry to hear you feel this way. It could be that you have your noise cancelling headphones on, as I'm pretty sure I say hello every time I walk into the office

If you see me not doing this when I'm next in, please call me out on it"

Suggested Manager Response

"Firstly, thanks for sharing your thoughts and experience.

It's so easy for all of us to get caught up in our day to day work that we forget some of the most impactful moments of interactions such as saying "good morning, good evening and thankyou"

I will refocus on this as I enjoy hearing this as well. Additionally I will reinforce this with my peers, as we can all focus more in this area. Thanks for sharing your thoughts with me"

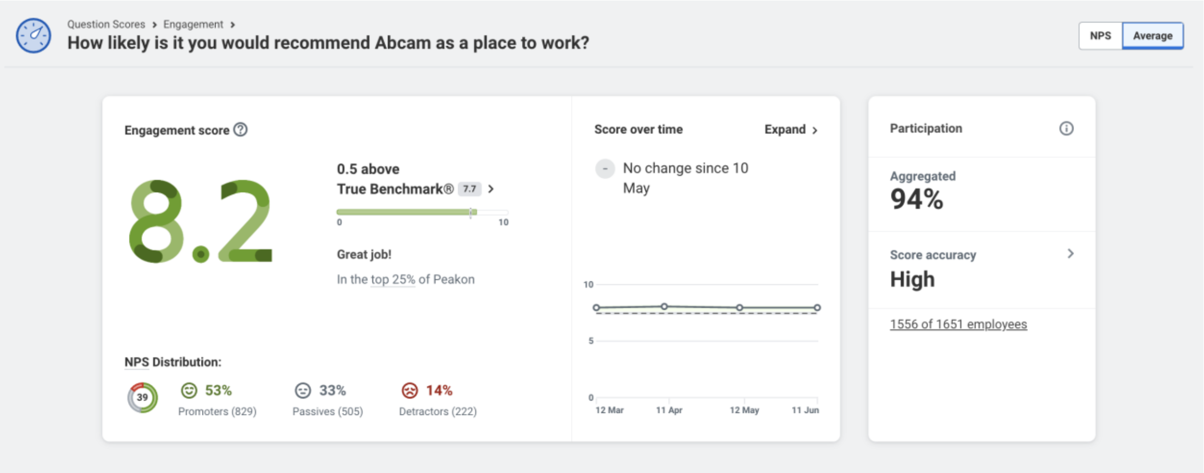
Figure 2 Example of Comment Interaction

Workday Peakon Employee Voice features real-time feedback, enabling continuous employee input, which is analysed to provide leaders with a comprehensive view of organizational. This helps identify improvement areas and track change impacts. Personalized dashboards for employees and managers display relevant metrics, while administrators use proxy views for better support. The platform promotes transparency and trust by showing employees their managers' responses, fostering open communication and collaboration. Advanced reporting and analytics offer detailed metrics, such as comment interactions by segment, giving administrators deeper insights into engagement patterns.

1.2 Motivation

In the contemporary business environment, organizational culture has emerged as a critical factor in company success, employee satisfaction, and overall performance (Schein, 2010). Recognizing this, Abcam has implemented a robust employee engagement strategy cantered around the use of Peakon.

For over three years, Abcam has been conducting monthly employee engagement surveys through Peakon. These surveys are designed to capture both quantitative and qualitative data (Figure 3&4).



Question Scores › Engagement ›

How likely is it you would recommend Abcam as a place to work?

Engagement score

8.2

0.5 above True Benchmark® 7.7

Great job!

In the top 25% of Peakon

NPS Distribution:

39 · 53% Promoters (829) · 33% Passives (505) · 14% Detractors (222)

Score over time

No change since 10 May

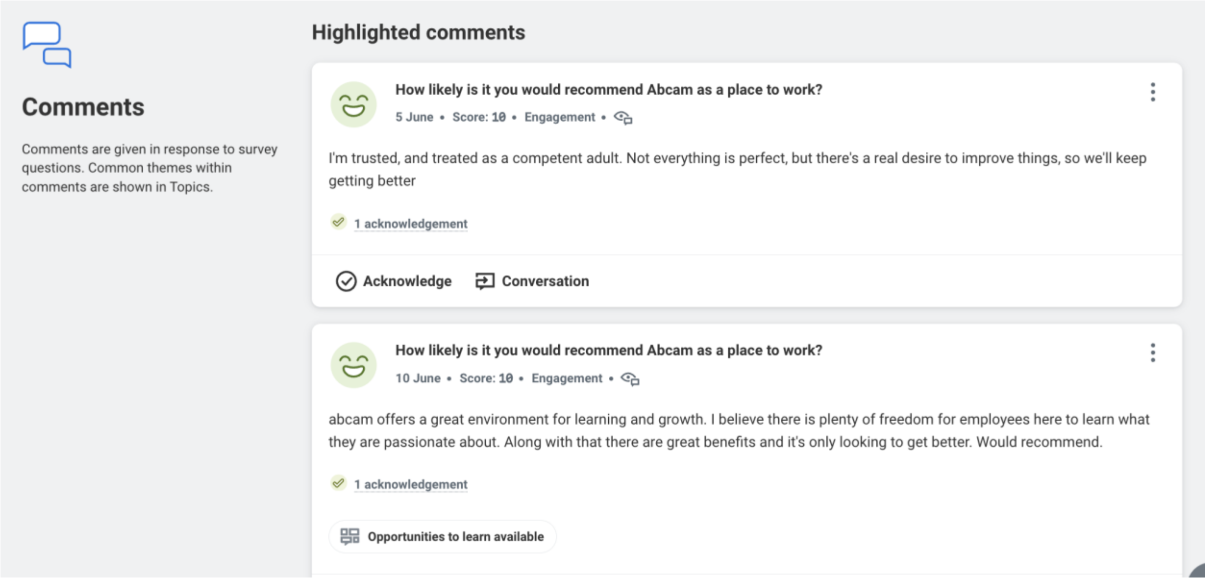
Participation

Aggregated 94%

Score accuracy High

1556 of 1651 employees

Figure 3 Example of Quantitative Data Collected through Net Promoter Score (NPS) Style Questions



Comments

Comments are given in response to survey questions. Common themes within comments are shown in Topics.

Highlighted comments

How likely is it you would recommend Abcam as a place to work?

5 June • Score: 10 • Engagement •

I'm trusted, and treated as a competent adult. Not everything is perfect, but there's a real desire to improve things, so we'll keep getting better

1 acknowledgement

How likely is it you would recommend Abcam as a place to work?

10 June • Score: 10 • Engagement •

abcam offers a great environment for learning and growth. I believe there is plenty of freedom for employees here to learn what they are passionate about. Along with that there are great benefits and it's only looking to get better. Would recommend.

1 acknowledgement

Opportunities to learn available

Figure 4 Example of Qualitative Data Gathered Via Each Question

The approach has led to a valuable dataset of employee feedback, consisting of approximately 1.5 million words. Typically, the survey includes 5-8 questions per month selected from a pool of about 40 different questions. Each question allows employees to provide detailed comments, resulting in a comprehensive understanding of employee on how they feel and what they've experienced throughout Abcam.

The abundance of text data offers a great opportunity for gaining insights, but it also brings significant challenges when it comes to analysis and interpretation. Firstly, analysing such a vast amount of textual data manually is not only time-consuming but also subjective and inconsistent. Moreover, the diverse workforce at Abcam contributes responses in Chinese and Japanese, which adds another layer of complexity to the analysis. Thus, these challenges we're facing right now have made it really necessary to have a highly advanced, automated solution that can efficiently process, analyse, and obtain valuable insights from this enormous amount of employee feedback.

1.3 Aims and Objectives

The main goal of this project is to develop an advanced Natural Language Processing (NLP) solution that can automatically classify and detect sentiment, enabling us to analyse employee comments from Abcam's engagement surveys. This aligns with recent advancements in applying NLP to HR analytics, as discussed by Pandey and Pandey (2019) & Laumer and Morana (2022). The project encompasses several specific objectives that will drive our work forward.

Firstly, we aim to build and train a machine learning model that can effectively classify employee comments according to Abcam's values framework. This framework, which includes the six key behaviours of:

Purpose:

Abcam emphasizes the importance of purpose-driven work and passion in fuelling strategy. Our commitment to meaningful initiatives aims at enabling scientific breakthroughs, aligning actions with our purpose, and continuously striving for higher standards of excellence.

Customer:

A customer-centric approach is essential for Abcam. We focus on exceeding customer expectations and driving sustainable growth by anticipating and addressing customer needs with agility and dedication.

Innovation:

Innovation is at the core of Abcam's approach. We prioritize new ideas and high-quality solutions, embracing a mindset that values rapid experimentation and calculated risks. This allows Abcam to maintain a competitive edge and achieve ambitious goals.

One Team:

Collaboration and teamwork are highly valued. Abcam breaks down silos to ensure seamless teamwork and knowledge sharing, promoting constructive debate as a foundation for innovation.

Growth & Development:

Continuous learning and development are crucial at Abcam. We provide the tools and resources needed for associates and teams to maximize their potential, committing to a culture of high performance and integrity.

Diversity & Wellbeing:

Abcam fosters a culture of belonging and empowerment. Inclusivity is seen as a strength and a source of innovation, helping them break down barriers and create opportunities for all.

The six dimensions of culture form the foundation of Abcam's high-performance company culture. The model, therefore, should be able to accurately categorize each comment by handling nuanced language and context-dependent expressions, building on recent advancements in multi-label text classification (Yang et al., 2018).

1.4 Methodology Overview

To achieve the project's objectives, a comprehensive methodology leveraging advanced NLP and machine learning techniques will be employed. This approach, structured into key phases, draws on state-of-the-art techniques in NLP and machine learning, particularly those applicable to HR analytics (Momin and Mishra, 2022). The project begins with data collection and exploration, involving data extraction and consolidation from the Peakon platform, followed by data labelling and exploratory analysis. Potential challenges such as class imbalance or bias that may impact modelling efforts will be identified, following best practices in HR data management (Angrave et al., 2016). The preprocessing stage involves cleaning and normalizing text data, handling multilingual content, and applying standard NLP techniques like tokenization, stemming, and lemmatization. Data quality and completeness issues will be addressed to ensure a solid foundation for modeling, utilizing advanced NLP techniques as outlined by Kowsari et al. (2019). Feature engineering will focus on creating relevant features from text data, exploring techniques including TF-IDF, word embeddings, and contextual embeddings like BERT to capture language nuances in employee comments.

Model development will experiment with state-of-the-art NLP models and frameworks, including BERT (Devlin et al., 2019), RoBERTa, XLNet (Yang et al., 2019), and FastText. Different architectures such as Neural Networks and Transformer-based models will be implemented and compared to determine the most effective approach. Training and validation will use rigorous techniques like cross-encoder and weight adjustment to ensure robust performance. Performance evaluation will involve defining and calculating relevant metrics such as accuracy and F1-score.

1.5 Structure of Report

This dissertation is structured into six comprehensive chapters, each designed to provide a thorough exploration of the project's various aspects.

Chapter 1 serves as an introduction, providing essential background information about Abcam, explaining the motivation behind the project, and outlining its relevance to the company's goals. It details the project's aims and objectives, offers an overview of the methodology, and outlines the structure of the report.

Chapter 2 presents a comprehensive literature review, encompassing the theoretical underpinnings and current research relevant to our project. It dives into the extensive historical background of organizational culture, explores methodologies for measuring it, examines various NLP techniques employed in text classification, and scrutinizes the application of machine learning in HR analytics.

The methodology is elaborated in Chapter 3, providing a detailed account of our approach. This includes an in-depth explanation of the data collection process, preprocessing techniques, and feature engineering approaches. The chapter also covers model selection and architecture, training and validation procedures, and our chosen evaluation metrics. Additionally, it discusses the methods used for ensuring model interpretability and explainability, as well as our strategies for deployment and integration.

Chapter 4 focuses on analysing and presenting the results of our work, comparing different model architectures such as traditional machine learning approaches (SVM, Random Forest) and advanced neural networks like RoBERTa. The chapter discusses the performance metrics of these models and explains how we fine-tuned the RoBERTa model using a cross-encoder architecture.

Chapter 5 presents a discussion of our results within the context of Abcam's organizational goals. It provides a methodological review, addressing the limitations of our current approach and suggesting potential improvements. The chapter also explores factors influencing model performance and proposes enhancements for future research. Additionally, it addresses ethical considerations, potential biases in our model, and offers recommendations for implementing and utilizing the tool for further research.

Finally, Chapter 6 concludes the dissertation by summarizing the key findings and contributions of our project. It reflects on how our work has impacted Abcam's cultural measurement capabilities and suggests directions for future research and potential improvements.

Chapter 2: Literature Review

2.1: Culture

2.1.1: Organisational Culture:

According to Schein (2010) in his seminal work Organizational Culture and Leadership, a robust and all-encompassing definition of organizational culture can be found:

"The culture of a group can now be defined as a pattern of shared basic assumptions that are learned by a group while solving its problems of external adaptation and internal integration. This pattern has proven to be effective enough to be considered valid and, therefore, is taught to new members as the correct way to perceive, think, and feel in relation to those problems" (Schein, 2010, pp.18).

From the definition, it becomes evident that culture emerges as a collective response to problem-solving within a group. In an organizational context, culture signifies a distinct manner in which individuals collectively react to various stimuli. Hence, it is intricately intertwined with social science, psychology, and behavioural science (Serpa, 2016). By delving into its origins and developmental history, we can acquire a more comprehensive understanding of the concept of organizational culture.

Historically, the study of organizational behaviour has employed concepts such as "group norms" and "climate," primarily focusing on observable behaviours within organizations. These concepts were further developed by Lewin, Lippitt, and White (1939), who introduced a more in-depth analysis of group dynamics and behaviours. However, researchers eventually recognized the limitations of these concepts in explaining the underlying structures that govern organizational behaviour at a deeper level, leading to the emergence of the concept of "organizational culture (Denison, Nieminen and Kotrba, 2012).

In the mid-20th century, organization culture gained more attention as scholars sought to understand the enduring patterns of behaviour and the logic underlying organizations beyond mere surface-level phenomena (Schein, 1965; Katz & Kahn, 1966). In addition, the need to differentiate organizational psychology from industrial psychology led to a focus on larger organizational units and the systems that govern their operations, with insights from sociology and anthropology enriching the understanding of organizational culture (Schein, 1965). This enhanced framework enabled researchers to explore variations in organizational behaviour and the levels of stability within organizational functions (Ouchi, 1981; Pascale & Athos, 1981).

Recognizing the performance disparities between U.S. companies and their international counterparts, particularly Japanese firms, served as a pivotal moment and crucial catalyst for a focused examination of organizational culture (Lincoln, Olson and Hanada, 1978). Researchers aimed to develop a framework that could explain these differences beyond attributing them solely to national cultural traits. For instance, Denison (1990) conducted comprehensive research on multinational culture and company performances, concluding that there might be a common method for organizational leaders to manage and measure culture rather than relying solely on national cultural backgrounds. By providing an independent framework for organizational culture, researchers believe it will lead to a more nuanced understanding of how organizational culture can impact effectiveness and competitive advantage (Wilkins and Ouchi, 1983)

Lots of early research on organizational culture employs various lenses to view how culture is formed, maintained, and altered within organizations. Some researchers emphasize a descriptive and clinical approach, advocating for an ethnographic understanding of culture that avoids premature managerial interventions (Martin, 1980; Van Maanen & Barley, 1984). This methodological diversity highlights the need for a careful investigation of cultural dynamics within organizations before suggesting changes.

2.1.2: Measurement of Organizational Culture:

Organizational culture, which includes shared values, beliefs, and practices, is important for how well a company performs and gets things done. Understanding the impact of organizational culture on stuff like employee engagement, innovation, and overall success requires measuring it. A strong organizational culture plays a huge role in boosting productivity and making sure a company succeeds in the long run (James and Collins, 1994).

Measuring culture within an organization can be challenging due to its abstract nature rather than being easily quantifiable (Chatman and Choi, 2019). However, despite the difficulties in measuring culture, several papers have made commendable attempts by conducting extensive literature research and identifying various approaches to measure it. The following part of literature review examines these approaches, models, and methodologies used to measure organizational culture while highlighting key studies and theoretical frameworks.

2.1.3: Approaches to Measuring Organizational Culture

The measurement of organizational culture employs both qualitative and quantitative methodologies. Qualitative methods, such as ethnographic studies and case studies, provide deep, context-rich insights into the cultural dynamics within organizations (Martin, 2002; Yin, 2018). Ethnographic studies involve immersive observation and detailed interviews, capturing the subtleties of cultural expressions and interactions. Case studies, on the other hand, focus on specific organizational contexts, allowing for a comprehensive analysis of cultural factors in situ (Lusia, Nurani and Mei, n.d.; Hopkins, 2006).

In addition, Heskett (2011) proposes a different theory for qualitatively measuring culture in his book *The Culture Cycle: How to Shape the Unseen Force That Transforms Performance*. According to Heskett, the effectiveness of organizational culture can be measured by evaluating its strength and health. Strength is determined by clearly articulating the mission, assumptions, values, beliefs, and behaviours and effectively communicating them to all members of the organization. Health is indicated by levels of trust, engagement, and ownership within the organization. This includes policies and practices that promote self-direction, accountability, transparency, risk-taking, collaboration, inclusion prioritization of the organization over self-interests and boundarylessness. Additionally fostering organizational learning and innovation requires continuous improvement adaptability speed agility.

Quantitative methods are primarily represented by surveys and questionnaires, which offer a more structured and scalable approach to measuring culture (Manetje and Martins, 2009). Notable among these is the Organizational Culture Assessment Instrument (OCAI) developed by Cameron and Quinn (2011), which categorizes organizational culture into four types: Clan, Adhocracy, Market, and Hierarchy. Similarly, Denison's Organizational Culture Survey (Denison, 1990) assesses culture based on mission, adaptability, involvement, and consistency. These instruments facilitate the comparison of cultural attributes across different organizations and over time.

2.1.4: Models of Organizational Culture Measurement

Several models have been proposed to conceptualize and measure organizational culture. The Competing Values Framework (CVF), proposed by Cameron and Quinn (2011), assesses culture based on dimensions of flexibility versus stability and internal versus external focus, resulting in four distinct culture types. This model is lauded for its comprehensive nature and widespread validation, though it may oversimplify the complexity of organizational cultures.

The Denison Model, which focuses on four key traits—involvement, consistency, adaptability, and mission—links these cultural traits directly to organizational performance (Denison, 1990). This model's strength lies in its empirical validation and its ability to demonstrate the impact of culture on organizational outcomes.

Edgar Schein's model delves deeper into the layers of organizational culture, proposing three levels: artifacts, espoused values, and basic underlying assumptions (Schein, 2010). While this model offers a profound depth of analysis, capturing the underlying assumptions poses significant measurement challenges.

2.1.5: Methodological Considerations

Ensuring reliability and validity is crucial when measuring organizational culture. Instruments such as the OCAI and Denison's survey undergo rigorous testing to ensure they accurately capture cultural dimensions (Cooke & Rousseau, 1988). Triangulation, which involves combining qualitative and quantitative methods, enhances the robustness of cultural assessments (Creswell & Plano Clark, 2017). Contextual factors, such as national culture and industry norms, also play a significant role in shaping organizational culture and must be considered in any measurement effort (Hofstede, 1991).

2.1.6: Challenges in Measuring Organizational Culture

Despite the abstract character of organisation culture, measuring organizational culture presents several challenges. The dynamic nature of culture means it evolves over time, complicating efforts to capture it consistently (Kotter, 1996). Moreover, culture is inherently subjective, with different members of the organization potentially perceiving it differently (Smircich, 1983). The complexity of culture, with its multifaceted and interwoven aspects, further adds to the difficulty of measurement (Schein, 2010).

2.1.7: Culture in Biotechnology Company:

In addition to the general examination of organizational culture, it would be advantageous to have literature specifically focused on the organizational culture within the biotech industry, as this is a crucial determinant of success in life-science companies, impacting innovation, employee performance and overall organizational efficiency (Lin and McDonough, 2011; Jaakson et al., 2011) - particularly relevant given Abcam's status as a biotechnology-based company. Fortunately, following extensive research several papers were identified and reviewed.

As mentioned in 2.1.1, organization culture refers to the shared values, beliefs, and norms that influence how employees interact and work towards company goals. Research indicates that in life-science companies, where technological advancement is rapid, a supportive culture plays a crucial role in achieving sustained success (Niosi & McKelvey, 2018). Similarly, Schein (2010) and Hellriegel & Slocum (2011) also emphasize the importance of a strong culture in fostering innovation, collaboration, and adaptability - all essential for thriving in an ever-evolving industry.

Based on literature research, Life-science firms must prioritize an innovative culture to remain competitive (Mittra, 2007). This involves not only encouraging risk-taking but also actively supporting new ideas through structured programs and initiatives that promote creativity(Mroczkowski, 2011). Providing the necessary resources for research and development is crucial, as it ensures that innovative ideas can be explored and developed into practical solutions(Tait, 2007). Firms need to invest in state-of-the-art laboratories, advanced technology, and continuous training for their employees to keep them at the forefront of scientific discovery and technological advancements (Isensee et al., 2020).

In the study of Baye and Yusuf (2023), who employs the Balanced Scorecard (BSC) model to assess organizational performance within the biotechnology sector, identifying Top Management Support (TMS) and Business Model Innovation (BMI) as key internal determinants. TMS is shown to enhance firm performance through effective resource allocation, strategic planning, and employee motivation, which are essential for organizational learning and resilience in biotech firms. Additionally, the study underscores the mediating role of company culture, which shapes employee behaviour and decision-making, fostering an environment conducive to sustainable performance. While BMI theoretically contributes to value creation and competitive advantage, its direct moderating effect on the TMS-venture performance (VP) relationship was found to be insignificant. These findings suggest that while TMS and company culture are critical for biotech venture (innovation) business success, the role of BMI requires further exploration, particularly through longitudinal studies to better understand its impact over time.

Apart from the importance of innovation inside life-science company, an effective leadership and management are pivotal in shaping and sustaining a positive organizational culture. Leaders must not only espouse the cultural values of the organization but also embody them in their daily actions, serving as role models for their teams. This commitment to cultural values should be evident in their decision-making processes, communication, and behaviour. Leaders need to be transparent, supportive, and approachable, fostering an environment where employees feel comfortable voicing their ideas and concerns. Additionally, leadership development programs can help cultivate these qualities in future leaders, ensuring the continuity of a positive culture (Nuscheler et al., 2019; Wijethilake & Lama, 2019).

High levels of employee engagement and job satisfaction are directly linked to a positive organizational culture (Jain et al., 2023). Life-science companies must create an environment where employees feel valued, respected, and motivated. This can be achieved through various means such as recognizing and rewarding employee achievements, providing opportunities for professional growth, and ensuring a healthy work-life balance. Regular feedback and open communication channels are also essential in understanding and addressing employee needs and concerns. A supportive workplace culture not only boosts morale but also enhances productivity and reduces turnover rates (Das & Tripathy, 2022; Lee, 2021).

Collaboration and teamwork are vital components of a successful organizational culture (Radu, 2023). In the life-science sector, this means fostering collaboration not just within departments but also across different departments and with external partners. Interdisciplinary teamwork can lead to innovative solutions that might not emerge within siloed groups (Jain et al., 2023). Encouraging a culture of teamwork involves promoting a shared vision, establishing common goals, and facilitating effective communication channels. Regular team-building activities and collaborative projects can strengthen these bonds. Moreover, partnerships with external organizations, such as universities, research institutions, and other companies, can bring in fresh perspectives and additional expertise, driving further innovation (Renko et al., 2022; Seo et al., 2020).

2.2. NLP

2.2.1 What is NLP

Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on the interaction between computers and human languages (Brockopp, 1983). It encompasses the development of algorithms and systems to understand, interpret, and generate human language. The history of NLP is rich and complex, marked by significant milestones and developments that have paved the way for its current state (O'Connor and McDermott, 2013).

2.2.2 History of NLP

The history of NLP dates back to the 1950s when the concept of machine translation first emerged. One of the earliest projects was the Georgetown-IBM experiment in 1954, which demonstrated the feasibility of machine translation. During the 1960s and 1970s, the focus was on rule-based systems. Researchers developed systems such as ELIZA, which could mimic human conversation, and SHRDLU, which could understand and manipulate blocks in a virtual environment (Green et al., 1961; Harris, 1984).

The evolution of Natural Language Processing (NLP) from the late 1970s to the present day has been characterized by significant paradigm shifts. Statistical models, leveraging large text corpora, emerged in the late 1970s and 1980s, advancing part-of-speech tagging and syntactic parsing (Hays, 1967; Hutchins & Somers, 1992). This approach gained momentum with the internet's rise in the 1990s, providing vast amounts of text data. The early 2000s saw the integration of machine learning techniques, with algorithms like Hidden Markov Models and

Conditional Random Fields enhancing tasks such as named entity recognition (Manning & Schuetze, 1999). The field was revolutionized in the 2010s by deep learning, particularly through recurrent and convolutional neural networks (Ruder, 2018). This progression highlighted the limitations of hand-crafted rules and the superiority of statistical methods in addressing natural language's complexity (Nadkarni et al., 2011). Karen Sparck Jones (1994) aptly summarized this development in four phases: machine translation focus, artificial intelligence influence, logico-grammatical styles, and data-driven approaches, illustrating the iterative refinement of NLP techniques over time.

2.2.3 Development Path of NLP

The evolution of Natural Language Processing (NLP) has been marked by distinct phases reflecting technological advancements. From the 1950s to the 1980s, rule-based systems like ELIZA and SHRDLU dominated, relying on handcrafted linguistic rules. While innovative, these systems were limited by their inflexibility and inability to handle natural language variability (Harris, 1984). The 1980s saw a paradigm shift towards statistical methods, leveraging large text corpora to learn language patterns. This era improved part-of-speech tagging and probabilistic parsing, highlighting the limitations of symbolic methods and introducing more robust, scalable solutions (Hutchins & Somers, 1992).

The early 2000s integrated machine learning techniques into NLP, with algorithms like Hidden Markov Models and Conditional Random Fields enhancing tasks such as named entity recognition. This phase enabled systems to learn from examples and improve over time, also introducing hybrid approaches combining rule-based and statistical methods (Manning & Schuetze, 1999). The 2010s witnessed a revolution with deep learning, as neural networks, particularly RNNs and CNNs, became central to NLP tasks. These models, capable of learning complex patterns from vast datasets, facilitated breakthroughs in translation, text generation, and sentiment analysis. Transformer models like BERT and GPT further extended NLP capabilities, enabling more accurate, context-aware language understanding (Ruder, 2018; Vaswani et al., 2017).

Recent research continues to build upon these advancements, addressing the limitations of earlier approaches and pushing towards more robust and scalable solutions. For instance, Mihalcea, Liu, and Lieberman (2006) explored the potential of Natural Language Programming to enhance programming accessibility, exemplifying the expanding intersections of NLP with other domains. The overall trajectory of NLP development reflects a continuous evolution from rule-based systems to data-driven approaches, incorporating increasingly sophisticated machine learning and deep learning techniques. This progression has significantly enhanced the accuracy and capability of language processing systems, rendering NLP integral to a wide array of applications across various fields.

2.2.4 Main Applications of NLP

As NLP continues to advance with the development of various models, its capability in text processing is greatly enhanced, leading to a wide range of applications. Neural machine translation (NMT), as a branch of NLP applications, has made significant progress in this field. For instance, Google Translate now supports over 100 languages for both text and speech translation (Wu et al., 2016). Machine translation plays a crucial role in localization by adapting products to different linguistic and cultural norms (Hutchins, 1986), while also enabling cross-lingual information retrieval and facilitating global access to knowledge (Dorr, 1993). Moreover, Speech recognition powers virtual assistants like Siri, Google Assistant, and Alexa, integral to modern smart devices (Jurafsky & Martin, 2018). It enables automated transcription services, aiding accessibility and documentation in educational and professional settings (Hirschberg & Manning, 2015), and is used in call centers to route calls and provide automated responses, improving customer service efficiency (Gonzalez et al., 2008).

In addition, Sentiment analysis is widely used in market research to analyze social media and customer reviews, guiding marketing strategies and product development. Political analysts use it to monitor public opinion on issues and candidates, while in financial markets, it's used to predict trends based on news and social media sentiment, aiding investment decisions. Information retrieval and extraction enhance search engines like Google and Bing, improving search experience and efficiency (Manning et al., 2008). This technology enables document summarization tools, valuable in law and academia (Nenkova & McKeown, 2011). In legal and medical fields, NLP systems extract information from documents and records, aiding decision-making processes (Demner-Fushman et al., 2009).

In healthcare industry, NLP processes clinical notes and electronic health records (EHRs) to enhance patient care and research (Nadkarni et al., 2011). Disease surveillance systems monitor healthcare data to detect outbreaks and track public health trends (Alemzadeh & Devarakonda, 2017). Chatbots and virtual health assistants provide patients with information and support, improving healthcare accessibility (Finch et al., 2019). In customer service, NLP powers chatbots that handle customer inquiries, reducing the need for human agents and improving response times (Jurafsky & Martin, 2018). Personalized recommendation systems analyze customer data to offer tailored product suggestions, enhancing satisfaction and driving sales (Ghahramani, 2015). Sentiment analysis tools analyze customer feedback to identify areas for improvement, allowing businesses to respond proactively to concerns (Wilson et al., 2005).

2.2.5 Recent Developments and Applications in Management Research

Kang et al. (2020) reviewed the application of NLP in management research, highlighting its growing importance in analyzing textual data to advance management theories across various disciplines. They noted the use of NLP in marketing, finance, and operations management, showcasing its versatility and impact in extracting valuable insights from large textual datasets. This literature review also emphasized the importance of adopting advanced NLP techniques, such as deep learning, to improve the accuracy and effectiveness of textual data analysis in management research.

2.2.6 Summarization

NLP has evolved from simple rule-based systems to sophisticated deep learning models capable of understanding and generating human language with remarkable accuracy. Its applications are diverse and impactful, ranging from machine translation to healthcare. As research in this field continues to advance, the potential for NLP to revolutionize the way we interact with technology is immense.

2.3 NLP & Organizational Culture

2.3.1 NLP Applications in Organizational Culture Analysis

Traditional approaches to measuring organizational culture, as discussed in section 2.1, have predominantly relied on mathematical or statistical techniques, offering a macroscopic perspective. However, these methods often fall short in capturing the nuanced thoughts of individual employees, particularly in large corporations with over 2,000 staff members. This limitation aligns with the challenges in measuring organizational culture identified by Schein (2010) and Smircich (1983).

The integration of Natural Language Processing (NLP) techniques with advanced employee feedback platforms (such as Peakon in our case) presents a promising solution to this limitation. By leveraging NLP models to analyze employee feedback at scale, we can develop a more comprehensive and granular understanding of company culture. This approach builds upon the data-driven NLP paradigm discussed in section 2.2, particularly the advancements in deep learning models (Ruder, 2018).

Applying NLP techniques to organizational culture analysis offers several key benefits. Firstly, it enables the processing of vast amounts of textual data, allowing for a more thorough examination of employee perspectives (Kasneci et al., 2023; Touvron et al., 2023). This depth of insights was previously unattainable with traditional methods. Secondly, deep-learning models like BERT and GPT make it feasible to automate the analysis of feedback from thousands of employees, overcoming traditional time and resource constraints (Radford et al., 2019; Brown et al., 2020). This scalability is crucial for large organizations seeking to understand their culture comprehensively. Lastly, NLP can detect subtle patterns and themes in employee comments that might be missed by conventional methods, providing a more nuanced understanding of the organizational culture (Zhang et al., 2023; Gu et al., 2023)..

This approach has the potential to revolutionize human resource management and enhance the effectiveness of people experience departments. By providing a more accurate and holistic view of organizational culture, NLP-driven analysis can inform targeted interventions and strategic decision-making, addressing the challenges in measuring organizational culture outlined by Kotter (1996) and Schein (2010).

The subsequent sections of this literature review, therefore, will focus on research related to three main areas: the implementation of NLP in Human Resource Management (HRM) systems, specific NLP techniques for analysing employee comments, and case studies supporting the effectiveness of this approach.

2.3.2 NLP Implementation in HRM Systems

The integration of NLP into HRM systems has gained significant attention due to its potential to streamline HR processes and enhance decision-making. The study by Laumer and Morana

(2023) provides a conceptual overview and state-of-the-art discussion on the use of conversational agents and NLP in HRM. They highlight the growing role of conversational agents, such as chatbots and digital assistants, in automating interactions with applicants, employees, and managers. These systems can handle tasks like resume screening, interview scheduling, and onboarding, thereby reducing the administrative burden on HR departments.

Tian et al. (2023) further explore this integration by developing a machine learning-based HR recruitment system using Latent Semantic Analysis (LSA), Bidirectional Encoder Representations from Transformers (BERT), and Support Vector Machines (SVM). Their research demonstrates that combining these NLP techniques can improve the accuracy and efficiency of resume classification, offering a robust alternative to traditional manual screening methods. This aligns with the progression of NLP techniques discussed in section 2.2, particularly the shift towards more sophisticated machine learning and deep learning approaches.

However, a critical limitation noted is the dependency on large and high-quality datasets for training the models. The quality of the output is heavily influenced by the quality of the input data, making it essential to have comprehensive and unbiased datasets, which can be a significant challenge. This echoes the challenges in measuring organizational culture discussed in section 2.1, particularly the dynamic and subjective nature of culture (Kotter, 1996; Smircich, 1983).

2.3.3 Specific NLP Techniques for Analysing Employee Comments

Analysing employee comments and feedback is crucial for understanding organizational culture and employee sentiments. Pandey and Pandey (2017) focus on applying NLP in computerized textual analysis to measure organizational culture, providing empirical evidence supporting the use of NLP in organizational research. By analysing annual letters to shareholders from Fortune 500 companies, they demonstrate how NLP can accurately measure organizational culture dimensions. Their validation tests, including content and external validity checks, confirm the robustness of NLP-enhanced computerized textual analysis in capturing complex organizational constructs. They argue that traditional survey methods are limited by participant biases and recruitment challenges, making NLP a valuable tool for large-scale studies. Their approach involves using multi-word phrase-level analysis to retain the linguistic context and improve the accuracy of textual analysis. However, a limitation noted in their study is the selective nature of the documents analysed, which may not represent the full spectrum of organizational communication.

The use of sentiment analysis is particularly emphasized in the study by Chintalapudi et al. (2021), which applies text mining and sentiment analysis to seafarers' medical documents. This research showcases the effectiveness of lexicon-based and machine learning-based sentiment classification methods in extracting meaningful insights from textual data. Lexicon-based methods are praised for their simplicity and interpretability, while machine learning-based methods, such as Naïve Bayes classifiers, offer higher accuracy and adaptability to different contexts.

These approaches align with the NLP applications discussed in section 2.2, particularly in the areas of sentiment analysis and information extraction. They also address the challenges of measuring organizational culture outlined in section 2.1, by providing more objective and scalable methods for analysing cultural dimensions.

2.3.4 Case Studies and Empirical Evidence

Empirical studies provide strong evidence for the effectiveness of NLP in HRM and related fields. For instance, Apell and Eriksson (2023) investigate the structural and functional dynamics of AI healthcare technology innovations using the Technological Innovation Systems (TIS) framework. Their findings highlight the importance of resource mobilization and communication in improving system performance. The study underscores the potential of AI and NLP technologies in transforming healthcare operations, which can be analogously applied to HRM systems. A limitation noted in their study is the regional focus on West Sweden, which may limit the generalizability of the findings to other regions and contexts.

In another empirical study, Tian et al. (2023) validate their ML-based HR recruitment system using real-world data from Kaggle. Their results show that the system significantly reduces the misclassification rate of resumes compared to traditional methods. This case study demonstrates the practical benefits of integrating NLP and machine learning into HR processes, leading to more efficient and unbiased recruitment.

Chintalapudi et al. (2021) present a compelling case study on the application of sentiment analysis to seafarers' medical documents. Their research shows that sentiment analysis can effectively identify common health issues and correlate them with medical outcomes, providing valuable insights for healthcare management. This study highlights the broader applicability of NLP techniques across different domains, including HRM.

Chapter 3: Methodology

3.1 Data Collection

3.1.1 Data Source:

The dataset for this project consists of several csv files that come from Abcam's engagement questionnaire, which collects employee comments and corresponding questions from Peakon. The questionnaire will be released every month, and while the questions may change, the culture driver will remain the same. As a result, the data will be updated and refreshed each month.

3.1.2 Data Labelling:

The original data frame lacks a labelled cultural dimension for employee comments, which is essential for training a classification model. Therefore, in order to construct and train a customized multi-classifier effectively, it is necessary to manually assign labels to a subset of the data so that the model can learn from and validate.

The original data frame contains the following columns:

Date

Question

Comment

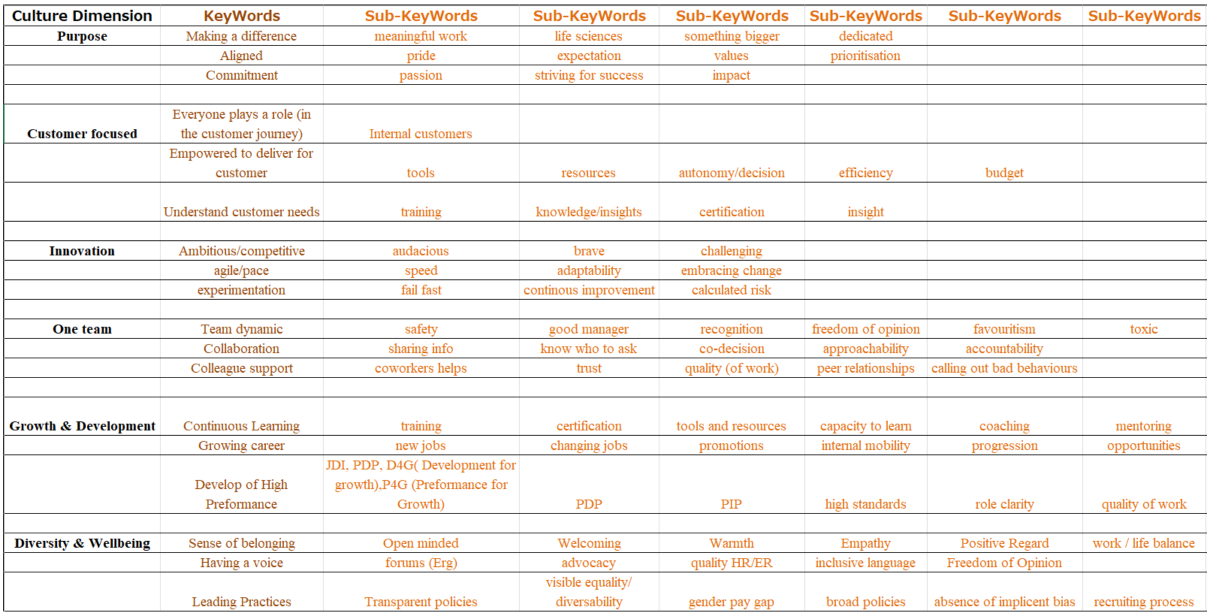
Score

Group

Driver/Value

Given the absence of a pre-existing reference point to rely on for defining Abcam's culture, due to each company having its distinct cultural style (as discussed in 2.1.6), a focus group comprising highly experienced HR professionals and senior managers from Abcam's HRM and people experience team was established to serve the purpose of culture definition and data labelling. This collaborative effort aligns with Pustejovsky and Stubbs' (2012) suggestion in their book *Natural Language Annotation for Machine Learning: A guide to Corpus building for applications*, serving as a guiding principle for future data labelling endeavours, enabling us to effectively capture the unique cultural value specific to Abcam.

Although the HR department possesses a framework and concise descriptions of Abcam's six cultural dimensions (as discussed in 1.3), these words fail to adequately capture the essence of Abcam's culture. The provided descriptions merely offer a broad overview of the six cultural dimensions, which is insufficient for the purpose of manual labelling. Consequently, during initial focus group meetings, participants engaged in discussions and expanded upon the descriptions for each dimension based on Abcam's current cultural context. The focus group incorporated keywords closely associated with each dimension to provide a more comprehensive understanding that can serve as a reference while reviewing comments. The keywords for each dimension and their relationship are shown below (Figure 5 & 6):



Culture Dimension	KeyWords	Sub-KeyWords	Sub-KeyWords	Sub-KeyWords	Sub-KeyWords	Sub-KeyWords	Sub-KeyWords
Purpose	Making a difference	meaningful work	life sciences	something bigger	dedicated		
	Aligned	pride	expectation	values	prioritisation		
	Commitment	passion	striving for success	impact			
Customer focused	Everyone plays a role (in the customer journey)	Internal customers					
	Empowered to deliver for customer	tools	resources	autonomy/decision	efficiency	budget	
	Understand customer needs	training	knowledge/insights	certification	insight		
Innovation	Ambitious/competitive	audacious	brave	challenging			
	agile/pace	speed	adaptability	embracing change			
	experimentation	fail fast	continous improvement	calculated risk			
One team	Team dynamic	safety	good manager	recognition	freedom of opinion	favouritism	toxic
	Collaboration	sharing info	know who to ask	co-decision	approachability	accountability	
	Colleague support	coworkers helps	trust	quality (of work)	peer relationships	calling out bad behaviours	
Growth & Development	Continuous Learning	training	certification	tools and resources	capacity to learn	coaching	mentoring
	Growing career	new jobs	changing jobs	promotions	internal mobility	progression	opportunities
	Develop of High Preformance	JDI, PDP, D4G(Development for growth),P4G (Preformance for Growth)	PDP	PIP	high standards	role clarity	quality of work
Diversity & Wellbeing	Sense of belonging	Open minded	Welcoming	Warmth	Empathy	Positive Regard	work / life balance
	Having a voice	forums (Erg)	advocacy	quality HR/ER	inclusive language	Freedom of Opinion	
	Leading Practices	Transparent policies	visible equality/ diversability	gender pay gap	broad policies	absence of implicent bias	recruiting process

Figure 5 Culture Class Defined by Keywords

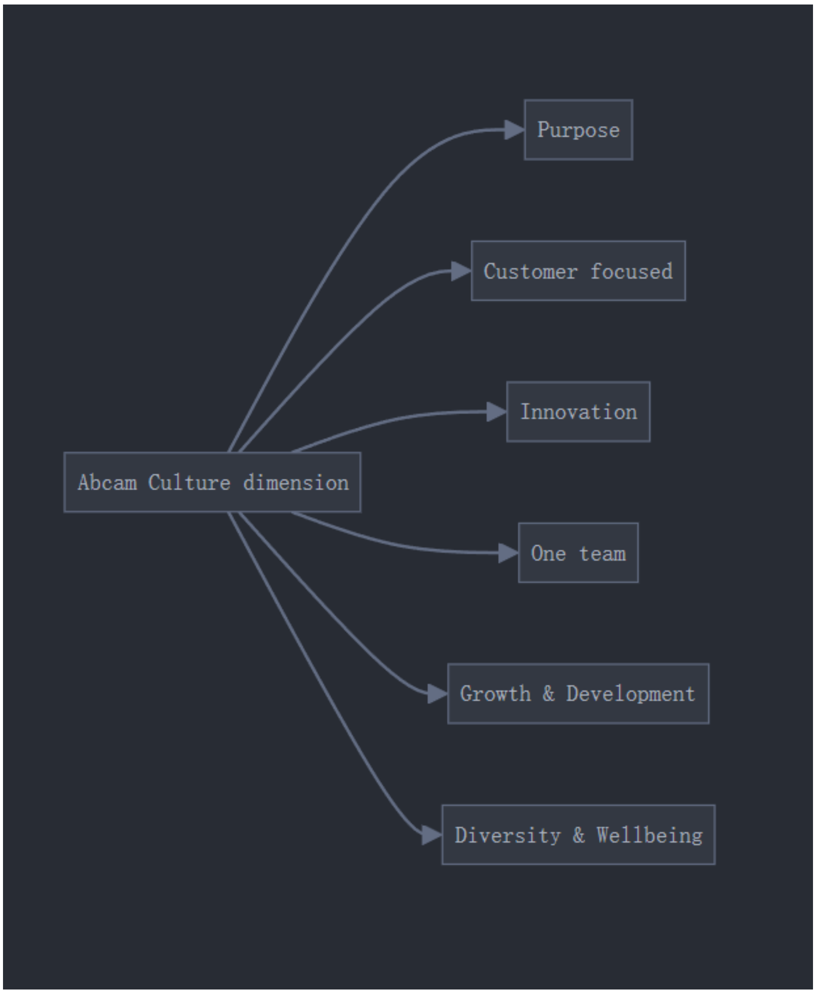
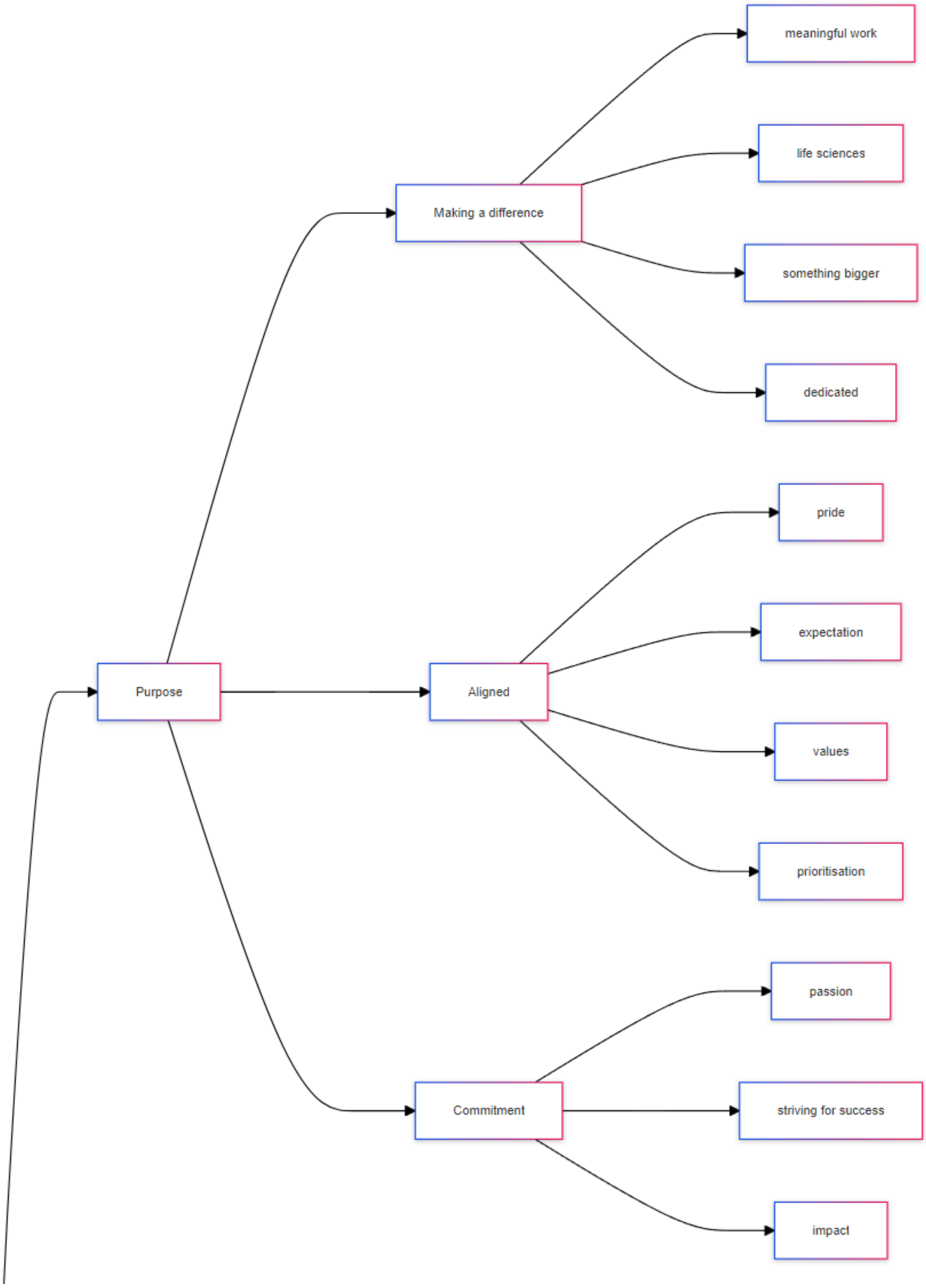


Figure 6 Culture Dimension Branches


meaningful work
life sciences
Making a difference
something bigger
dedicated
pride
expectation
Purpose
Aligned
values
prioritisation
passion
Commitment
striving for success
impact

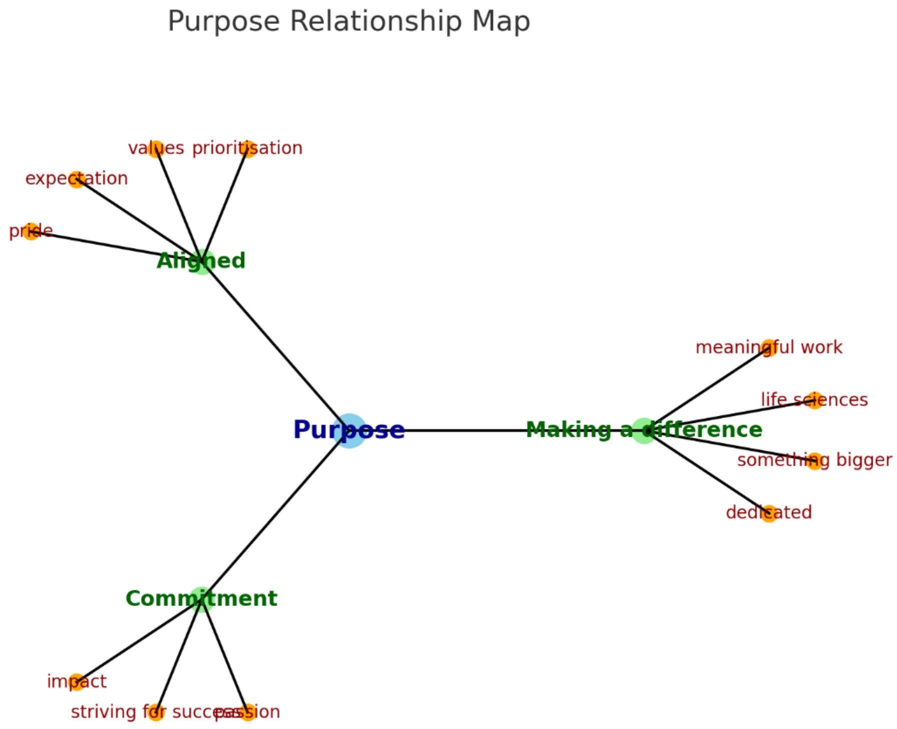


Figure 7 Examples of How Keywords Defines the ‘Purpose’ Culture Dimension

Following the redefinition of Abcam's six cultural dimensions, we have incorporated an additional six columns in the data frame to represent these dimensions alongside their corresponding comments as showed in Figure 7. This approach enhances comprehension and labelling for individuals involved in data annotation. With this approach and the culture reference, focus group members can review the comments and assign a value of '1' if they pertain to one or more culture dimensions; otherwise, they mark it as '0'. For instance, consider the comment: 'Our new project is innovative and encourages excellent teamwork.' In this case, since innovation and teamwork align with two out of the six cultural dimensions, we annotate them as '1', while assigning '0' to the remaining four.

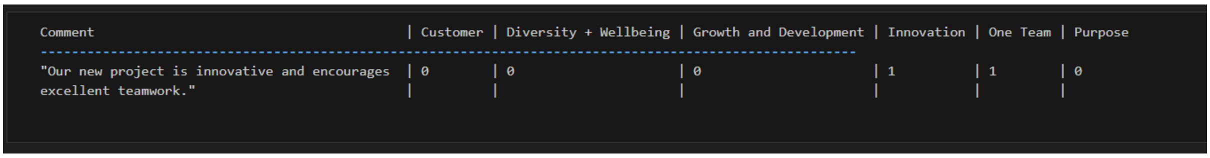
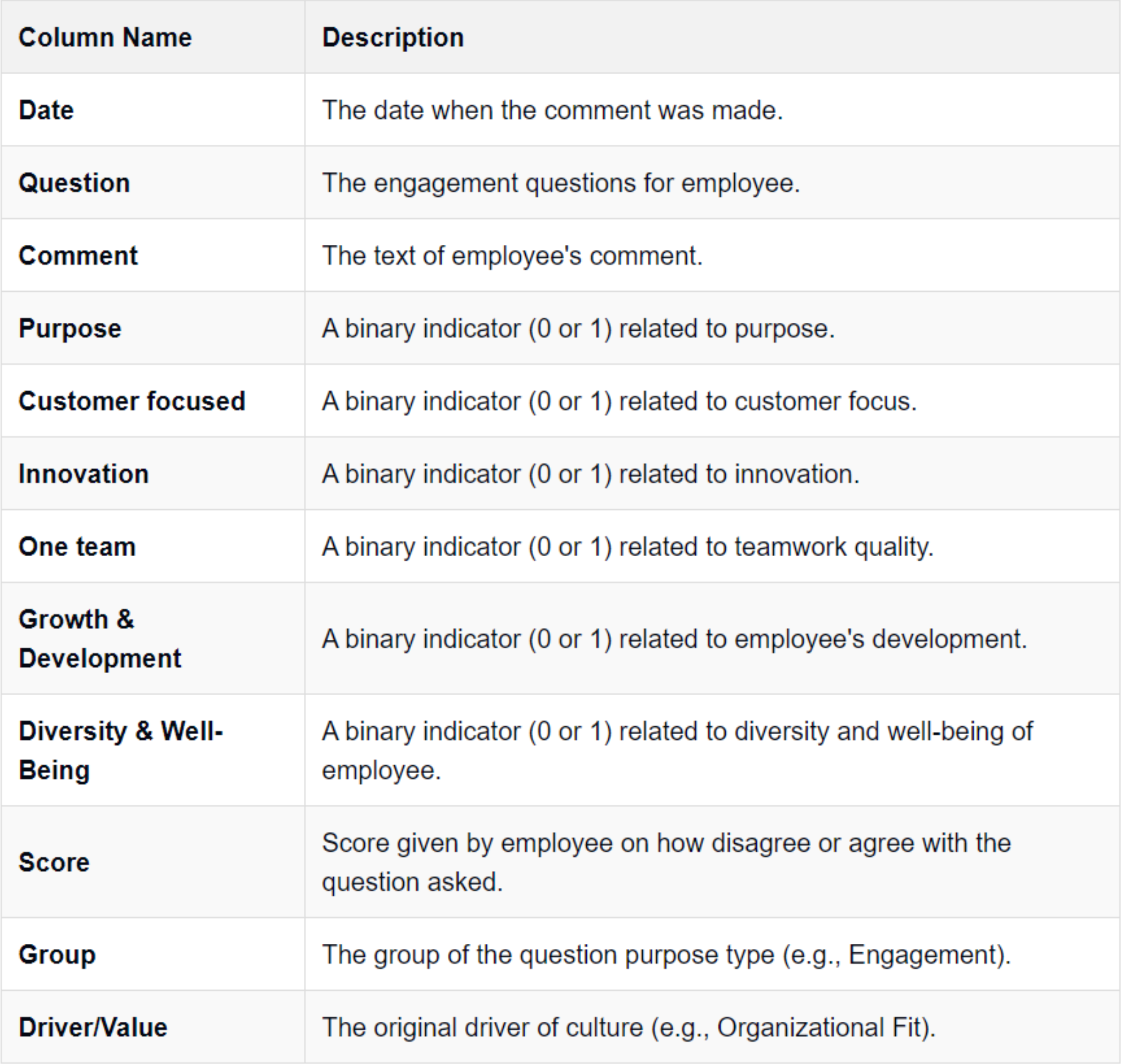

```
Comment                                       | Customer | Diversity + Wellbeing | Growth and Development | Innovation | One Team | Purpose
----------------------------------------------------------------------------------------------------------------
"Our new project is innovative and encourages | 0        | 0                     | 0                      | 1          | 1        | 0
excellent teamwork."                          |          |                       |                        |            |          |
```

Figure 8 Example of Data Labelling

After further discussion and considering the workload and time availability, each member of the focus group was assigned to label 200 rows of comments. The final labelling team consists of 6 members, resulting in a dataset of 1200 rows of comments.

3.1.3 The Structure and Description of The Labelled Dataset:



Column Name	Description
Date	The date when the comment was made.
Question	The engagement questions for employee.
Comment	The text of employee's comment.
Purpose	A binary indicator (0 or 1) related to purpose.
Customer focused	A binary indicator (0 or 1) related to customer focus.
Innovation	A binary indicator (0 or 1) related to innovation.
One team	A binary indicator (0 or 1) related to teamwork quality.
Growth & Development	A binary indicator (0 or 1) related to employee's development.
Diversity & Well-Being	A binary indicator (0 or 1) related to diversity and well-being of employee.
Score	Score given by employee on how disagree or agree with the question asked.
Group	The group of the question purpose type (e.g., Engagement).
Driver/Value	The original driver of culture (e.g., Organizational Fit).

Figure 9 Column Descriptions of Labelled Dataset

The last three data columns won't be used for this project since they are all for the purpose of performing quantitative analysis and do not provide any useful information. However, in the future, they might prove valuable for further research. For instance, the Score column can serve as a supplementary reference for sentiment analysis performance by revealing employees' attitudes towards specific questions (Pang and Lee, 2008) and consequently indicating their sentiments towards those questions as well.

3.2 Data Cleaning

Data cleaning is a crucial step in preparing textual data for analysis, ensuring that the dataset is consistent, accurate, and free from errors. In this project, we first addressed duplicate entries, which can skew results and introduce bias. Using the *pandas* library (McKinney, 2010), we employed the *drop_duplicates()* method to remove any redundant comments. Additionally, handling encoding issues was essential to avoid misinterpretation of characters. This was achieved by converting the text to a consistent encoding format using latin1 and utf-8. The code for this process involved encoding the text with latin1 and then decoding it back to utf-8, ensuring a uniform character set.

Next, we stripped leading and trailing whitespaces from the comments, which can otherwise affect tokenization and word matching. This was accomplished using the *str.strip()* method in pandas. Finally, ensuring consistent date formats is vital for any temporal analysis. By converting the 'Date' column to a datetime format using the *pd.to_datetime()* function, we standardized the date representation across the dataset. These cleaning steps are fundamental as they ensure that the data is in a reliable state for subsequent analysis and modelling.

3.3 Data Preprocessing for Word-Embedding Methods

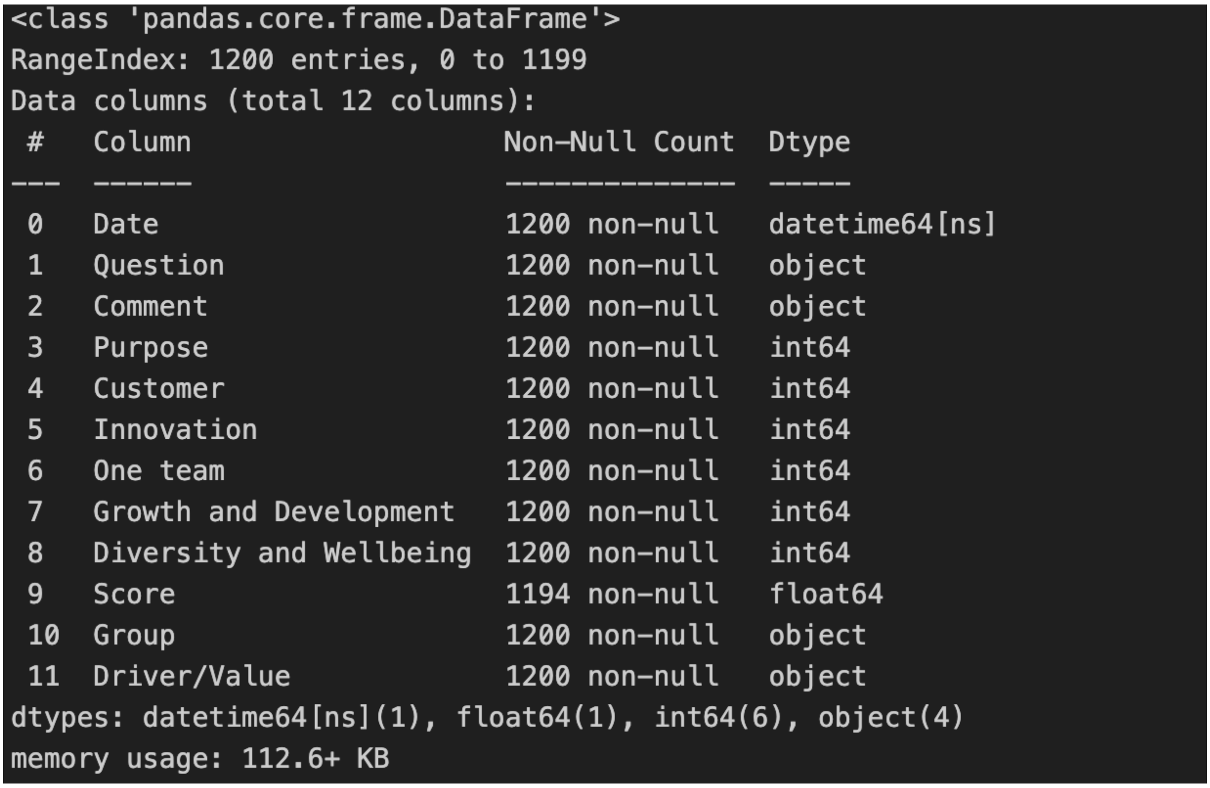
Data preprocessing is essential to transform raw text into a structured format suitable for word embeddings (Manning et al., 2008). This process begins with tokenization, where the text is split into individual words using the *word_tokenize* function from the Natural Language Toolkit (NLTK) and Spacy (Bird et al., 2009; Honnibal and Montani, 2017). Next, all words are converted to lowercase to ensure consistency, implemented with a simple lambda function applied to the tokenized words.

Removing stop words, which are common words like "and" and "the" that do not carry significant meaning, helps to focus on the more meaningful parts of the text (Jurafsky and Martin, 2020). This is done using the stopwords module from NLTK. Finally, lemmatization reduces words to their base or root form, ensuring that different forms of a word (e.g., "running" and "ran") are treated as a single item ("run") (Jurafsky and Martin, 2020). This step uses the WordNetLemmatizer from NLTK.

These preprocessing steps—tokenization, lowercasing, stop word removal, and lemmatization—are crucial for preparing the text data for word embeddings (Alam et al., 2020). By following these steps, we ensure that the text data is consistent, meaningful, and ready for advanced analysis.

3.4 Exploratory Data Analysis

3.4.1 Dataset Overview:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Date                     1200 non-null   datetime64[ns]
 1   Question                 1200 non-null   object
 2   Comment                  1200 non-null   object
 3   Purpose                  1200 non-null   int64
 4   Customer                 1200 non-null   int64
 5   Innovation               1200 non-null   int64
 6   One team                 1200 non-null   int64
 7   Growth and Development   1200 non-null   int64
 8   Diversity and Wellbeing  1200 non-null   int64
 9   Score                    1194 non-null   float64
 10  Group                    1200 non-null   object
 11  Driver/Value             1200 non-null   object
dtypes: datetime64[ns](1), float64(1), int64(6), object(4)
memory usage: 112.6+ KB
```

Figure 10 Overview of Dataset

Based on the provided illustration above, the dataset being analysed is structured as a pandas Data Frame, comprising 1200 entries distributed across 12 columns. The columns encompass diverse data types tailored to their respective content: the 'Date' column employs a datetime64[ns] type for precise temporal tracking, while categorical variables such as 'Purpose', 'Customer', and 'Innovation' utilize 'int64' types. Textual data including 'Question', 'Comment', 'Group', and 'Driver/Value' are stored as object types, enabling flexible string representations.

The dataset is highly complete, with 11 out of 12 columns having non-null values for all 1200 entries. The 'Score' column, represented as float64, has only a few missing values (6 out of 1200), which is insignificant for most analytical purposes and will not affect the overall analysis as it won't be used in this project. This extensive data completeness allows for comprehensive exploratory data analysis without the need for imputation or concern about missing value biases.

3.4.2 Data Exploratory visualizations

The data visualization techniques used for this dataset include distribution plots such as histograms and KDE plots, which help understand the spread, central tendency, and presence of outliers in numerical variables. Bar charts are used to show the frequency distribution of categorical variables, making it easier to identify imbalances between categories. Additionally, comment length analysis and a word-cloud are employed as supplementary methods to explore key characteristics of comments. These visualizations aim to provide a comprehensive understanding of the data, guiding data preprocessing, feature selection, and model building for more informed analysis and decision-making.

3.4.2.1 Comments label distributions plots:

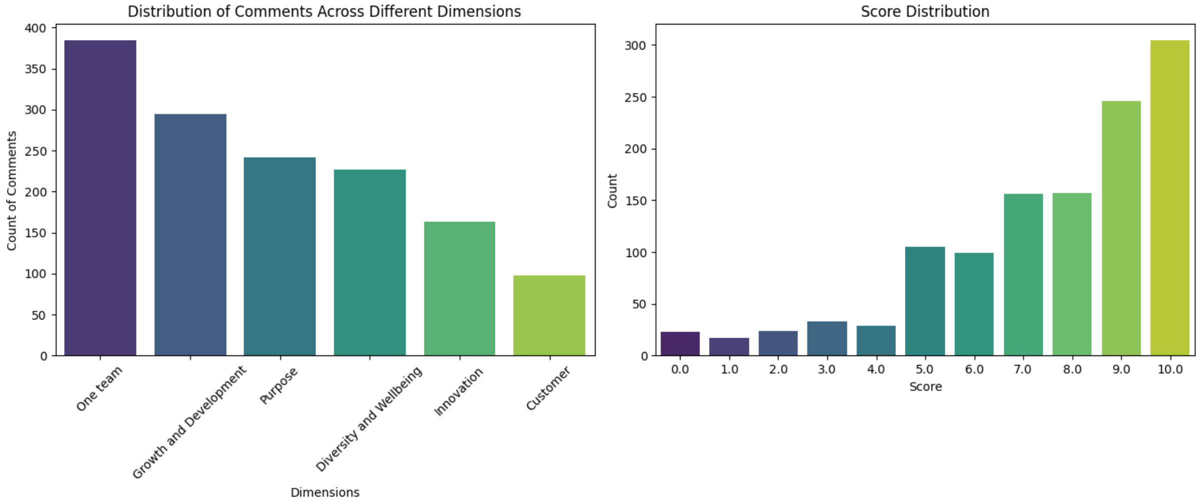


Figure 11 Comments Label Distributions Plots

The distribution of labels across different cultural dimensions in the top left chart indicates that 'One team' has the highest count of comments, suggesting it is a significant area of feedback. Conversely, 'Customer' has the lowest count, indicating less interaction or concern in this area. Other dimension labels such as 'Growth and Development,' 'Purpose,' and 'Diversity and Wellbeing' have moderate levels of comments.

Furthermore, an uneven distribution of comment labels across different dimensions highlights the need to readjust class weights in the loss function for further model training and fine-tuning to give more importance to minority classes during training (He and Garcia, 2009; Buda et al. 2018; Johnson and Khoshgoftaar, 2019).

The bar chart on the right side depicts the distribution of scores across all comments, revealing a pronounced skew towards higher scores, predominantly 10, and fewer comments with lower scores (0-4), indicating limited occurrence of negative feedback. This concentrated prevalence of top scores implies an overall positive sentiment in the comments regarding Abcam's employees.

3.4.2.2 Score Distribution for Each Culture Dimension:

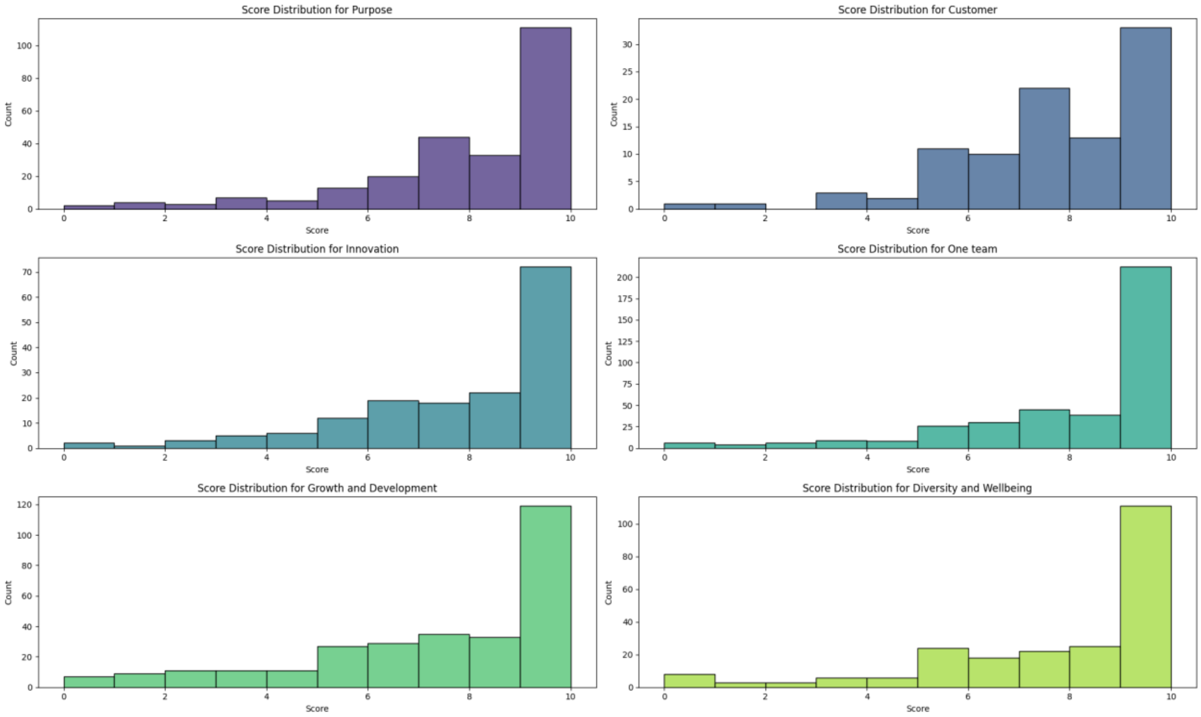


Figure 12 Score Distribution for Each Culture Dimension

From these charts above, the histograms for six culture dimensions show that most scores are skewed towards 10, indicating positive feedback. Purpose and Innovation have predominantly high scores. Customer scores vary but are mostly between 7 and 10. One team shows many scores at 10, indicating highly positive feedback. Growth and Development, and Diversity and Wellbeing, while skewed towards higher scores, display more variability. Overall, the high scores suggest positive feedback across most dimensions, with some variability in Growth and Development and Diversity and Wellbeing.

3.4.2.3 Comment Length Distribution

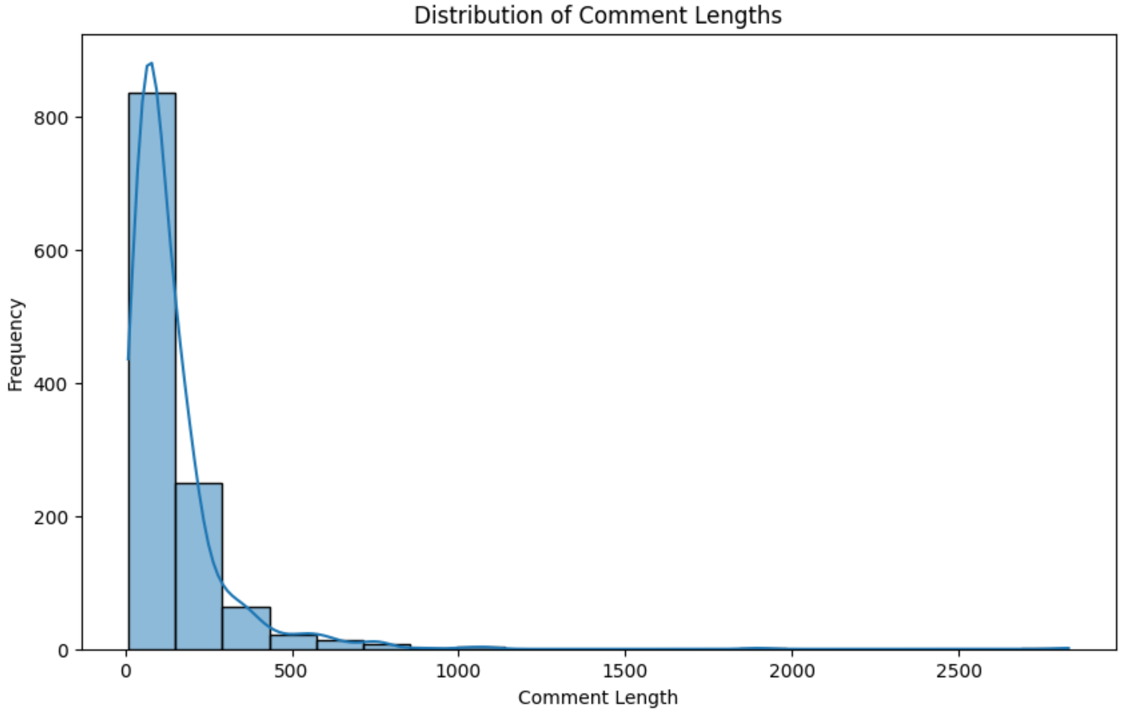


Figure 13 Comment Length Distribution

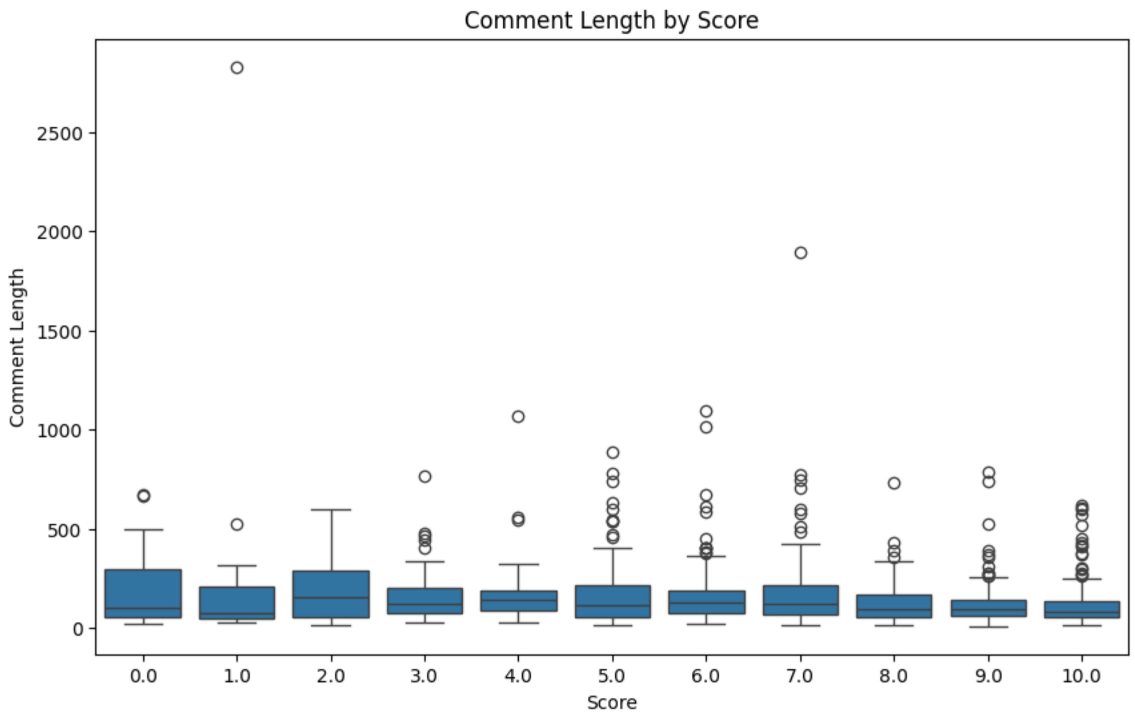


Figure 14 Comment Length Distribution Based on Score

The analysis of comment length distribution reveals that most feedback is concise, with the majority of comments falling within 0-500 characters. This trend towards brevity is consistent across different score categories, with median lengths generally clustered between 0-100 characters. However, there's a notable presence of outliers, particularly for comments associated with middle-range scores (4-7), indicating instances of more detailed feedback. Comments linked to extreme scores (0-1 and 9-10) tend to be shorter, suggesting users provide more direct feedback when highly satisfied or dissatisfied.

These findings have important implications for data processing and analysis. The prevalence of short comments necessitates efficient text processing methods optimized for brief inputs. Simultaneously, the presence of longer outliers, especially in middle-score ranges, highlights the need for techniques that can extract value from more detailed responses. This balanced approach will enable a comprehensive analysis of user feedback, capturing both the dominant trends in concise responses and the nuanced insights found in more elaborate comments.

3.4.2.4 Word Cloud:

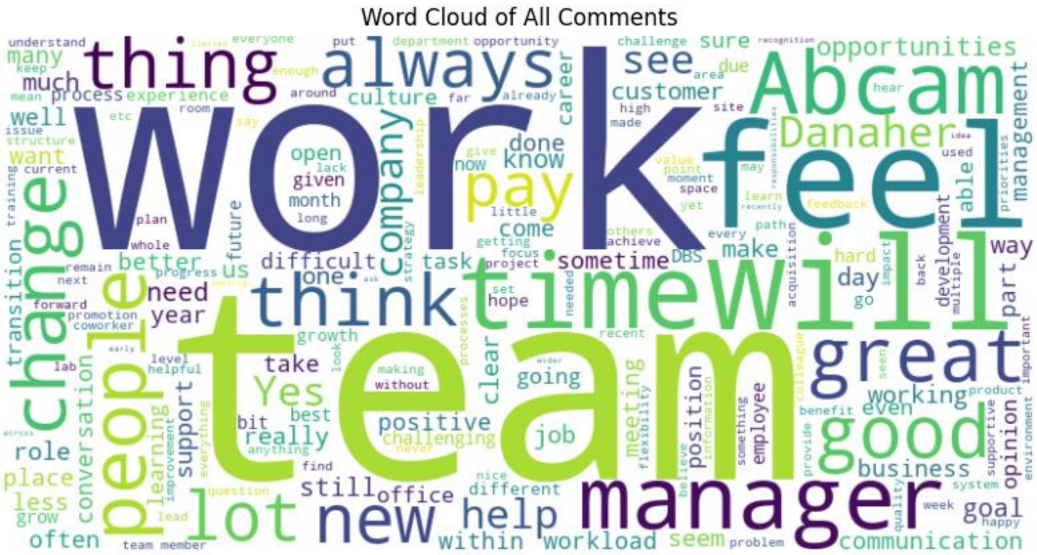

Word Cloud of All Comments

Figure 15 Word Cloud

The word cloud visualizes the frequency of words in the comments, with larger words appearing more often. Prominent words such as "work" and "team" highlight that discussions around work-related topics and teamwork dynamics are predominant in the comments. The frequent mentions of "manager" and "management" indicate that many comments involve discussions about management practices or specific managers. Additionally, the words "feel" and "think" suggest that the comments often contain personal opinions and sentiments.

Other notable words include "great," "good," and "positive," which indicate a significant amount of positive feedback. The words "change" and "new" suggest discussions about organizational changes or new initiatives. Common topics of discussion like "pay," "job," and "help" are likely related to compensation, job roles, and assistance or support within the organization.

3.5 Data Input Strategy:

The available text data for Abcam necessitates the consideration of merging the question column with the comment column, as approximately 5% of comments received during the dataset labelling process consist of brief responses such as 'yes,' 'no,' or very short answers like ‘I think so’ (Figure 16).

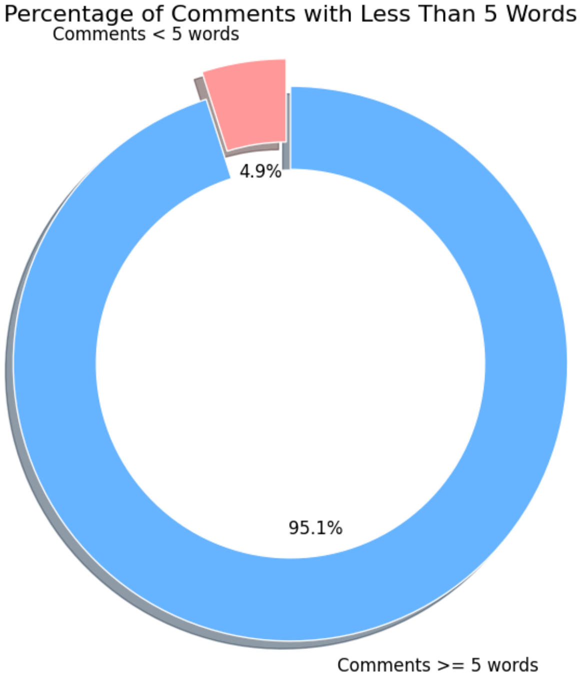


Figure 16 Percentage of Short Comments

The brevity of these comments alone does not provide adequate contextual information for data annotating, necessitating the need to read the corresponding questions. Hence, it is imperative to address this issue in subsequent sections:

Comments Only or Questions + Comments As The Text Input?

After several rounds of discussion, the group decided to adapt a combination of questions and comments as input for model training rather than comments alone. As combining questions with comments provides additional context that significantly enhances the understanding of the comment (Liu, 2020). Questions often set the frame of reference or the topic of discussion, aiding in the accurate interpretation of sentiment and specific nuances within the comments (Poria et al, 2018). Including questions enriches the feature space, capturing more relevant information and improving the model's ability to distinguish between different classes, leading to better performance in classification tasks (Devlin et al., 2018).

Additionally, comments alone as the input may sometimes be ambiguous or lack specificity, whereas the accompanying question can clarify the intent behind the comment, reducing ambiguity and leading to more accurate sentiment analysis and classification (Yang and Cardie, 2014). Empirical evidence from previous analyses and visualizations indicates that incorporating both questions and comments results in better-structured embeddings, which enhance the model's learning capability (Mayur Wankhade, Sekhara and Kulkarni, 2022). Utilizing both questions and comments as input leverages the full context and richness of the data, likely resulting in more robust and accurate models for classification and sentiment analysis (Neeraj Anand Sharma, Ali and Muhammad Ashad Kabir, 2024).

3.6 Questions Need to be Addressed:

Since we have cleaned and processed the data, the next step would be embedding the text data and use it to train classification models. There are two main questions need to be answer on the following sections, which are:

Which word-embedding methods should be selected and why?

Which classification model should be selected and why?

To address these inquiries, we will experiment with various word-embedding techniques and classification models, subsequently comparing their performances to obtain conclusive answers for both questions.

3.7 Word Embedding methods selection

To ensure the utilization of optimal word-embedding methods, this section will explore a range of word embedding techniques, progressing from fundamental to advanced levels. Subsequently, these word embedding outputs will be applied into a simple regression model to evaluate their respective performances. Finally, a comparative analysis will be conducted in order to identify the most suitable embedding method for this project.

The Utilized Word-Embedding Methods and Their Performance Metrics:

BOW (Bag of Words), Harris (1954)

TF-IDF (Term Frequency-Inverse Document Frequency), Sparck Jones (1972)

Word2Vec, Mikolov et al. (2013)

BERT (Bidirectional Encoder Representations from Transformers), Devlin et al. (2018)

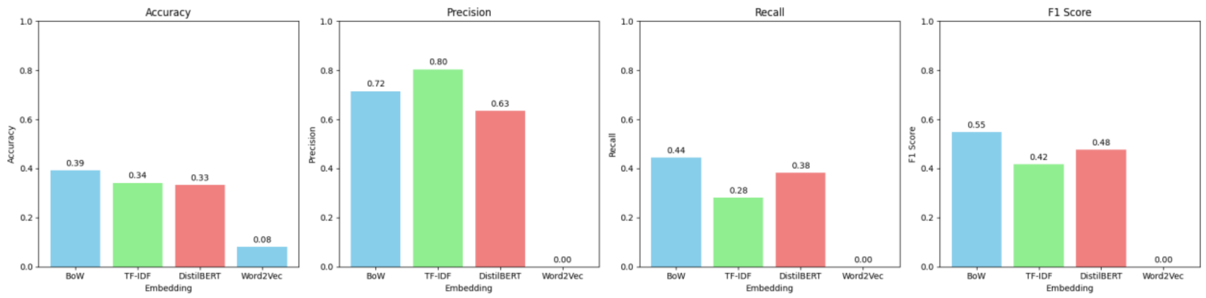


Figure 17 Performance of Different Word-Embedding Methods

Based on Figure 17, the performance of different word embedding techniques (Bag of Words (BoW), TF-IDF, DistilBERT, and Word2Vec) for multi-label classification is evaluated using accuracy, precision, recall, and F1 score. Each method shows varying effectiveness, providing insights into their suitability for sentiment analysis.

From the chart, it is clear to see that BoW demonstrates the best overall performance under this experiment, making it the most effective method for this multi-label classification task. TF-IDF excels in precision but has low recall. DistilBERT, despite being advanced, does not outperform simpler methods significantly. Word2Vec consistently performs poorly, highlighting the importance of selecting suitable embedding techniques based on dataset characteristics and task requirements.

The observed performance differences can be attributed to several factors. The dataset's short and straightforward comments favour simpler models like BoW and TF-IDF, which capture word frequency effectively. In contrast, complex models like Word2Vec and DistilBERT, designed to leverage contextual relationships, underperform due to the limited context and high feature sparsity (Wang et al., 2016). Additionally, the small dataset size exacerbates overfitting in Word2Vec and reduces the effectiveness of DistilBERT's dense embeddings (Kowsari et al., 2019). Consequently, for datasets with short texts and limited samples, simpler models often outperform more advanced techniques.

Considering the performances, we will keep BoW and BERT for word embedding methods, as BoW demonstrated superior performance due to its simplicity and robustness in handling short, straightforward text (Wang et al., 2016). It effectively captured the frequency of terms, which is particularly useful when dealing with smaller datasets where advanced models might overfit or fail to capture meaningful patterns due to insufficient context. The high performance of BoW suggests that for this type of data, where comments are concise, and the vocabulary is relatively limited, simple frequency-based methods are highly effective.

In contrast, DistilBERT, although not surpassing BoW in performance, still demonstrated competitive results. Its utilization of pre-trained language representations enabled a better understanding of contextual information compared to Word2Vec, which struggled with the brevity and limited size of the dataset (Sanh et al., 2019). The performance of DistilBERT suggests that transformer-based models can be advantageous, particularly as the dataset expands and comments become more intricate, providing a richer context for these models to learn from (Devlin et al., 2018. Furthermore, transformer-based models can be fine-tuned, indicating potential for further improvement in this model (Sun et al., 2019). It is worth noting that we incorporated BERT embeddings into a regression model—a technically permissible approach but one that may have an impact on performance (Reimers and Gurevych, 2019; Gao et al., 2019).

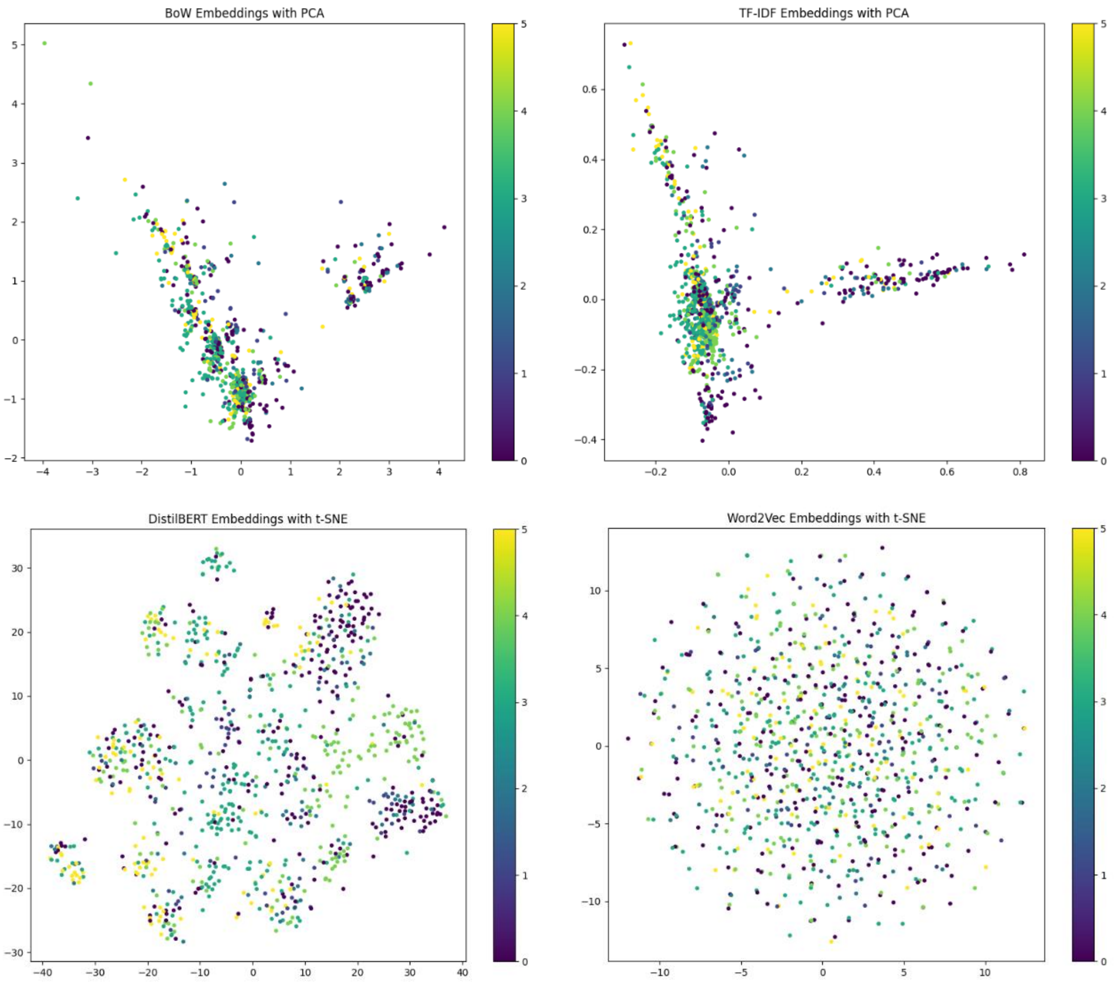


Figure 18 t-SNE & PCA Plots for Different Word-Embedding Methods

Figure 16 reveal the distinct characteristics of text embeddings generated using Word2Vec, Bag of Words (BoW), TF-IDF, and DistilBERT techniques. The t-SNE plot (Maaten and Hinton, 2008) for Word2Vec embeddings shows a broad and diffuse distribution, reflecting its ability to capture a wide range of semantic relationships between words, but lacking clear category distinctions. In contrast, the PCA plots (Wold et al., 1987) for BoW and TF-IDF embeddings display more structured and concentrated clusters, indicating their effectiveness in capturing distinct patterns within the text data.

This structured clustering is likely why BoW, in particular, outperformed other techniques in classification tasks, as it offers clearer distinctions between different categories (Yang and Cardie, 2014). While TF-IDF also captures word importance and aids in classification, its clustering is slightly less distinct than BoW. DistilBERT embeddings, visualized through t-SNE, exhibit more defined clusters compared to Word2Vec but remain less structured than BoW or TF-IDF. This reflects DistilBERT's capacity to capture complex semantic relationships, though it requires more sophisticated models or further tuning to fully leverage its potential in classification tasks. These visualizations elucidate the varying performance of each embedding technique, with BoW and TF-IDF providing more explicit category differentiation, and Word2Vec and DistilBERT capturing deeper semantic nuances (Yang and Cardie, 2014).

In summary, the chosen word-embedding methods for this project are Bag-of-Words (BoW) and BERT. Based on the previous analysis, both BoW and BERT have shown remarkable performances. Furthermore, BoW's simplicity and efficacy in handling short texts establish a strong baseline performance, which aligns well with our dataset characteristics revealed in the Exploratory Data Analysis (3.4.2 EDA) section indicating that most comments have text lengths predominantly distributed within 0-500 characters, with a concentration around 100-300 characters. On the other hand, BERT's contextual understanding offers potential enhancements as the dataset becomes more complex. Additionally, since training the BERT model requires using BERT as the tokenizer, we need to retain the BERT word-embedding for further comparison of model performance.

3.8 Multi-Classifier Models and Cross-Encoder Application:

3.8.1 Adopted Models

The comments in this project will be classified based on cultural dimensions using a multi-label classification methodology. To ensure robust evaluation, the dataset will be partitioned into training, validation, and testing sets. Various models including Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Random Forests (Breiman, 2001), and advanced Neural Networks architectures such as BERT will be explored and tested for their performance. Bag-of-Words (BoW) text embeddings will serve as features (apart from BERT model) while considering cultural dimensions as target labels during the training of the selected model.

The BERT model exclusively accepts the BERT tokenizer as the word embedding method, rendering BoW inapplicable to the BERT model. Instead, we independently trained the BERT model end-to-end and employed a cross-encoder architecture (Humeau et al., 2020) to concatenate questions with employee comments in conjunction with the BERT model.

3.8.2 Reasons to Apply Cross-Encoder Architecture

Firstly, cross-encoders excel in understanding context by capturing intricate relationships between questions and corresponding comments (Humeau et al., 2020). Unlike bi-encoder architectures that process inputs independently, cross-encoders analyse question-comment pairs holistically. This allows the model to grasp the semantic interplay between a question's context and the nuances of an employee's response, capturing implicit references and contextual clues that might otherwise be lost. This is crucial in HRM, where subtle linguistic cues significantly affect feedback interpretation and classification into culture dimensions (Otter et al., 2021; Cheng et al., 2021).

Secondly, cross-encoder architecture enhances feature interaction by facilitating deep interaction between features from both the question and the comment at multiple levels within the transformer layers (Vaswani et al., 2017). This allows for modelling complex, non-linear relationships between the question's intent and the comment's content. Attention mechanisms can focus on relevant parts of both inputs simultaneously, uncovering latent patterns indicative of specific culture dimensions (Tay et al., 2020). This is especially beneficial in HRM applications, where relationships between questions about company policy and employee responses may not be immediately apparent but are crucial for accurate classification (Noorbehbahani and Kargar, 2019; Cheng et al., 2021).

Additionally, cross-encoders handle ambiguity in employee responses effectively, which often contain multi-faceted opinions. By considering the full context provided by both the question and the comment, cross-encoders weigh different aspects of the response against the specific framing of the question (Liu, 2020). This capability is invaluable in culture dimension classification, where comments may touch upon multiple dimensions, requiring nuanced interpretation (Hofstede et al., 2010; Otter et al., 2021).

In scenarios where a limited set of recurring questions are employed, cross-encoders offer distinct advantages by facilitating the acquisition of question-specific patterns and their correlation with different cultural dimensions through repeated exposure (Cer et al., 2018). This renders the utilization of cross-encoders highly suitable for our dataset, as we have observed variations in employee comments while maintaining recurrent questioning (Bao et al., 2014).

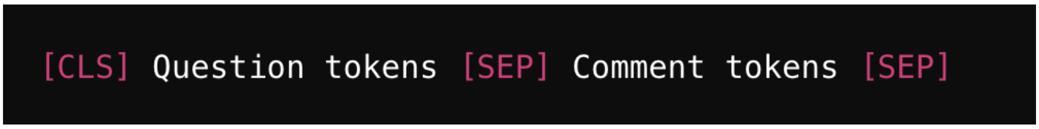
3.8.3 Process of Applying Cross-Encoder

To merge the Questions and Comments together with Cross-Encoder, first, both the question and comment are tokenized using the RoBERTa tokenizer. After tokenization, the two sequences (question and comment) are combined into a single sequence. The special tokens [CLS] and [SEP] are utilized:

[CLS]: This token is added at the beginning of the sequence and is used by the model to aggregate information from the entire sequence. It is typically used for classification tasks.

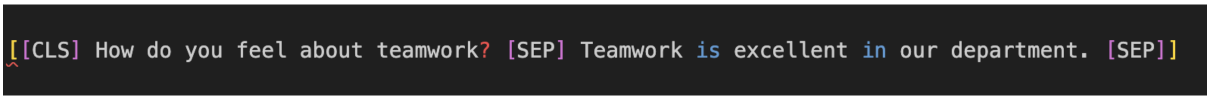
[SEP]: This token is used to separate the two segments (question and comment) within the sequence.

The concatenated sequence takes this form:



```
[CLS] Question tokens [SEP] Comment tokens [SEP]
```

Example of Cross-Encoder applied for this project:



```
[[CLS] How do you feel about teamwork? [SEP] Teamwork is excellent in our department. [SEP]]
```

This format allows the model to process the question and comment together, considering the context provided by the question when interpreting the comment. After the formatting, the entire sequence is then passed through the cross-encoder (in this case, RoBERTa), which generates a contextualized representation of the combined input.

Chapter 4: Result Analysis

4.1 Model Performance Comparison

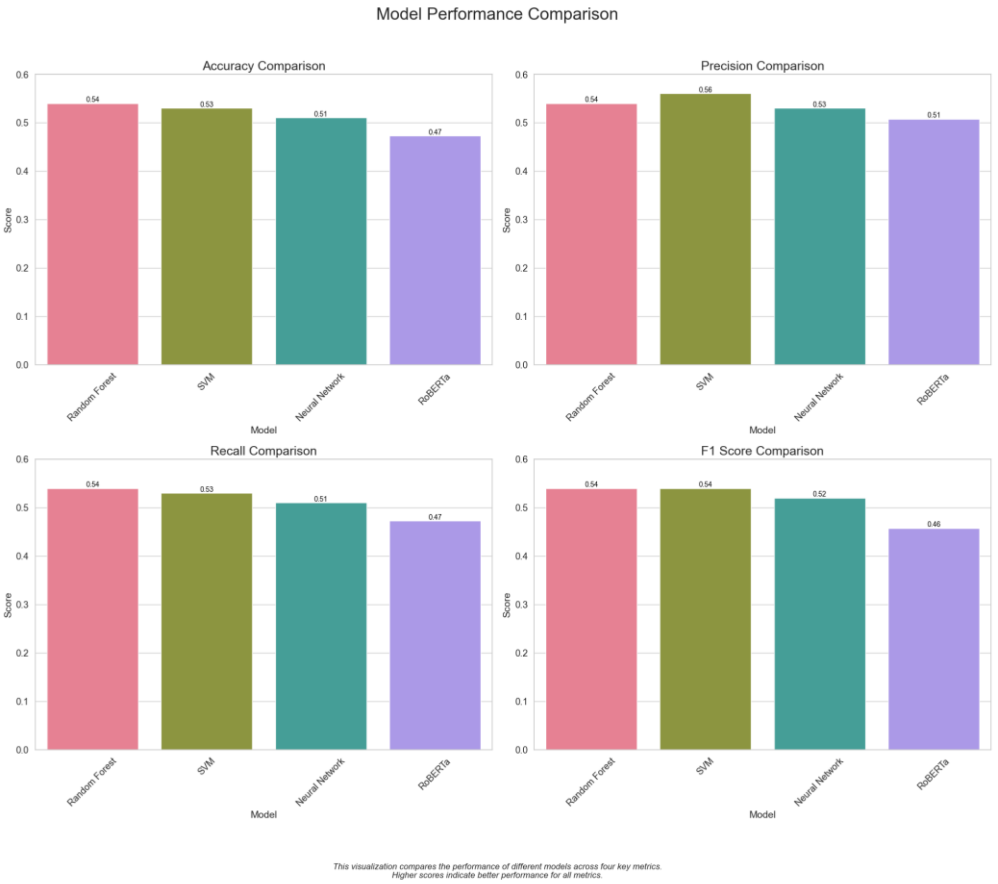


Figure 19 Model Performance Comparison

Upon thorough examination of the model performance comparison visualization, we can discern distinct patterns of efficacy across four machine learning models: Random Forest, Support Vector Machine (SVM), Neural Network, and RoBERTa, evaluated against four critical performance metrics.

4.1.1 Accuracy Comparison:

The Random Forest algorithm demonstrates superior performance in terms of accuracy, achieving a score of 0.54. This is followed closely by SVM (0.53) and Neural Network (0.51), while the RoBERTa model lags behind with an accuracy of 0.47. This suggests that the ensemble-based Random Forest method exhibits a marginally enhanced capability in correct classification across all instances.

4.1.2 Precision Comparison:

In the domain of precision, which evaluates the model's ability to avoid false positives, the SVM model excels with a score of 0.56 (Sokolova and Lapalme, 2009). This is followed by Random Forest (0.54), Neural Network (0.53), and RoBERTa (0.51). The SVM's superior precision indicates its propensity for making more reliable positive predictions, which could be particularly valuable in scenarios where false positives carry significant consequences.

4.1.3 Recall Comparison:

The recall metric, assessing the model's proficiency in identifying all relevant instances, reveals Random Forest (0.54) outperforming its counterparts. SVM follows closely (0.53), with Neural Network (0.51) and RoBERTa (0.47) showing diminished performance. This suggests that the Random Forest model is more adept at capturing a higher proportion of true positive cases, which is crucial in applications where missing positive instances is particularly problematic.

4.1.4 F1 Score Comparison:

The F1 score, representing the harmonic mean of precision and recall, shows Random Forest and SVM models achieving parity (both at 0.54). The Neural Network model follows (0.52), while RoBERTa trails noticeably (0.46). This metric provides a balanced assessment of the models' overall performance, considering both false positives and false negatives.

In conclusion, the traditional machine learning approaches, particularly Random Forest and SVM, demonstrate consistent superiority across all metrics in this specific classification task. The Neural Network model, while competitive, generally underperforms compared to these methods. Notably, the RoBERTa model, despite its sophisticated architecture, exhibits the least favourable performance across all metrics.

These findings suggest that for this type of multi-classification problem, the complexity of transformer-based models like RoBERTa may not offer advantages over simpler, more traditional machine learning approaches. This underscores the importance of comprehensive model evaluation and the potential superiority of less complex models in certain scenarios, aligning with the No Free Lunch Theorem for model selection (Wolpert and Macready, 1997). Further investigation into feature engineering, hyperparameter optimization, and dataset characteristics would be prudent to fully understand the performance discrepancies and potentially enhance the efficacy of the more sophisticated models.

4.2 Model Fine-Tuning

In this section, a robust strategy is employed to fine-tune a cross-encoder model using RoBERTa-base for accomplishing a multi-label classification task. The approach incorporates weight adjustment techniques to address class imbalance and utilizes hyperparameter tuning with Optuna (Akiba et al., 2019) for optimizing the performance of the model.

4.2.1 Weight Adjustment for Class Imbalance

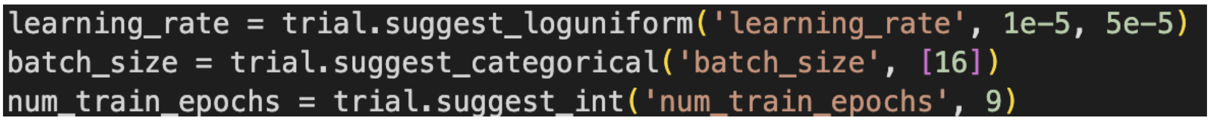
The importance of weight adjustment is emphasized by the distribution of comment labels across different dimensions in the dataset. As discussed in the exploratory data analysis (3.4.2) section, the label distribution reveals a significantly higher count for the "One team" label compared to the "Customer" label. Consequently, during subsequent model training, there is a risk that the model may excessively focus on learning from instances with the "One team" class while neglecting instances related to Customer and other class dimensions (Cui et al., 2019; Johnson and Khoshgoftaar, 2019). This biased learning process could potentially impact model training.

To address this issue, we compute class weights using the *compute_class_weight* function from *sklearn*. This function calculates weights inversely proportional to class frequencies, ensuring that underrepresented classes receive higher weights. These weights are then integrated into the model's loss function, penalizing misclassifications of minority classes more heavily and guiding the model to focus on these classes.

The use of class weights mitigates the impact of class imbalance, ensuring that the model does not favour majority classes at the expense of minority ones (Krawczyk, 2016). This leads to a more balanced performance across all classes.

4.2.2 Test Run to Decide The Initial Hyperparameter Input

Prior to commencing fine-tuning, a preliminary exploration was conducted using a limited set of random parameters to establish the initial starting point for the subsequent rigorous hyperparameter optimization process (Bergstra and Bengio, 2012; Feurer and Hutter, 2019). This approach aligns with established practices in hyperparameter tuning, where initial exploration can inform more targeted optimization strategies (Li et al., 2017). The blind testing encompassed ten trials, and the ensuing training results are presented in the following charts



```
learning_rate = trial.suggest_loguniform('learning_rate', 1e-5, 5e-5)
batch_size = trial.suggest_categorical('batch_size', [16])
num_train_epochs = trial.suggest_int('num_train_epochs', 9)
```

Figure 20 Hyperparameters Before Optuna Fine-Tuning

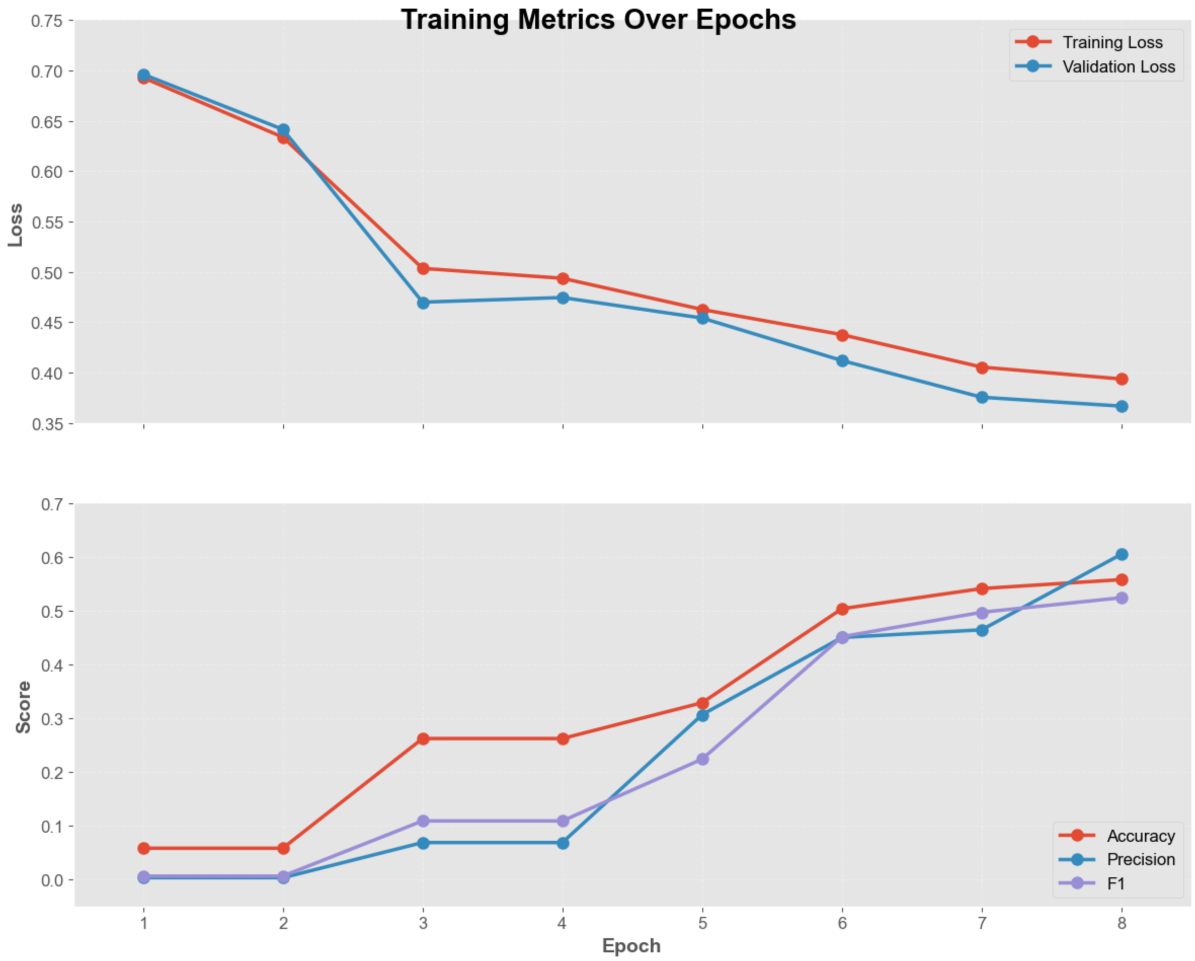


Figure 21 Training Process Metrics Over Epochs

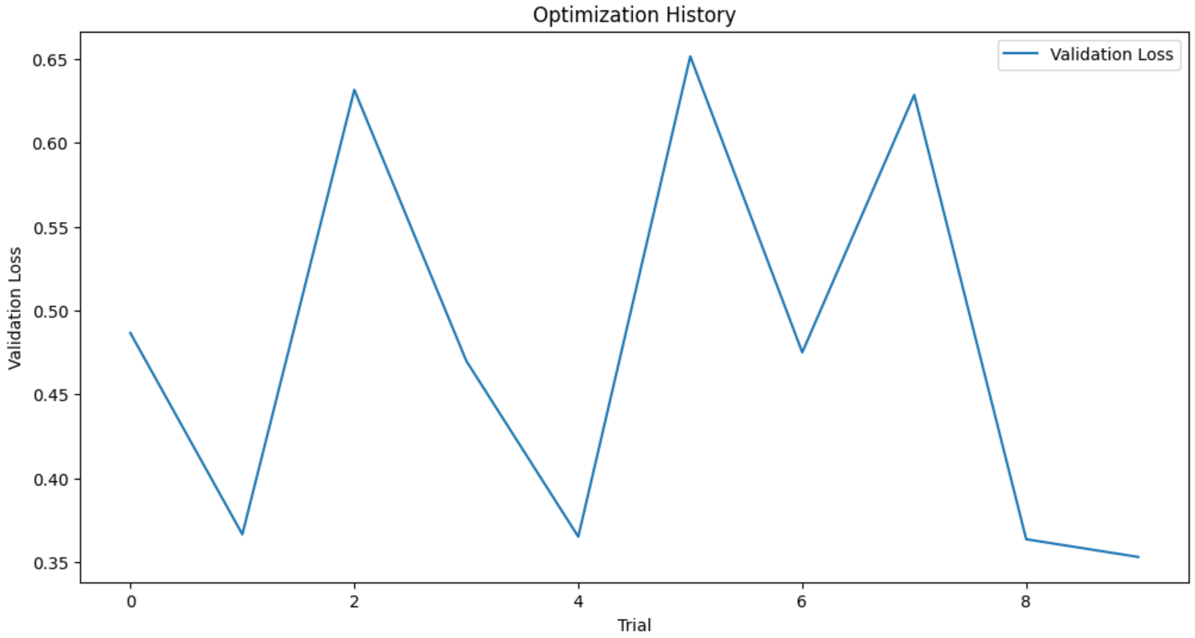
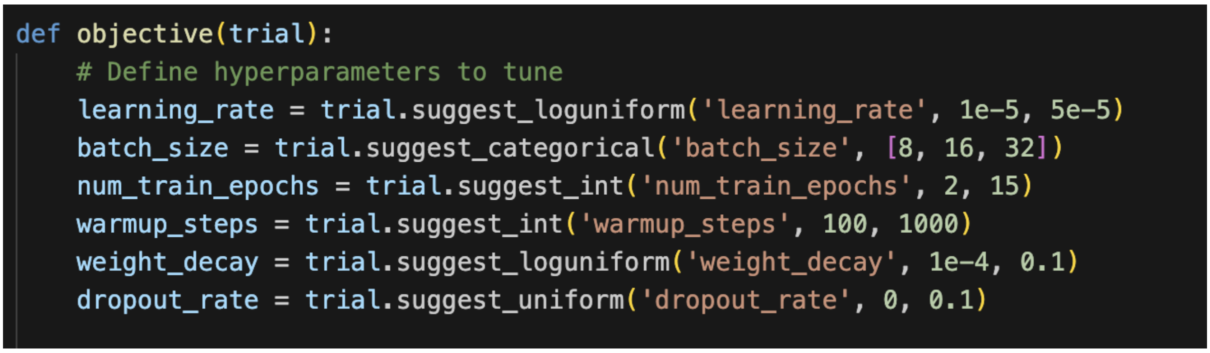


Figure 22 Validation Loss Through Training Trials

The loss curves for both the training and validation sets consistently decline (Figure 21), converging noticeably after the third epoch, indicating effective learning without significant overfitting. Notably, the performance metrics (accuracy, precision, and F1 score) show substantial improvement between the fifth and sixth epochs, suggesting a potential breakthrough in the model's learning process. Moreover, the oscillating pattern observed in the optimization history chart (Figure 22) of validation loss indicates exploration of diverse hyperparameter configurations during search while consistently decreasing. Consequently, it is advisable to consider larger values for both epoch number (>9) and trial number (>10) during hyperparameter tuning.

4.2.3 Hyperparameter tuning using Optuna:

The search space is defined for key hyperparameters, including learning rate, batch size, number of training epochs, warmup steps, weight decay, and dropout rate. Optuna's '*study. Optimize'* function is utilized to conduct a series of trials where each trial represents a different set of hyperparameters. Each trial is evaluated based on the validation loss, and the search strategy is iteratively updated to converge on the optimal set of hyperparameters. Below is the hyperparameter defined for tuning:



```python
def objective(trial):
    # Define hyperparameters to tune
    learning_rate = trial.suggest_loguniform('learning_rate', 1e-5, 5e-5)
    batch_size = trial.suggest_categorical('batch_size', [8, 16, 32])
    num_train_epochs = trial.suggest_int('num_train_epochs', 2, 15)
    warmup_steps = trial.suggest_int('warmup_steps', 100, 1000)
    weight_decay = trial.suggest_loguniform('weight_decay', 1e-4, 0.1)
    dropout_rate = trial.suggest_uniform('dropout_rate', 0, 0.1)
```

Figure 23 Hyperparameter Defined For Optuna

4.2.4 Training History Analysis:

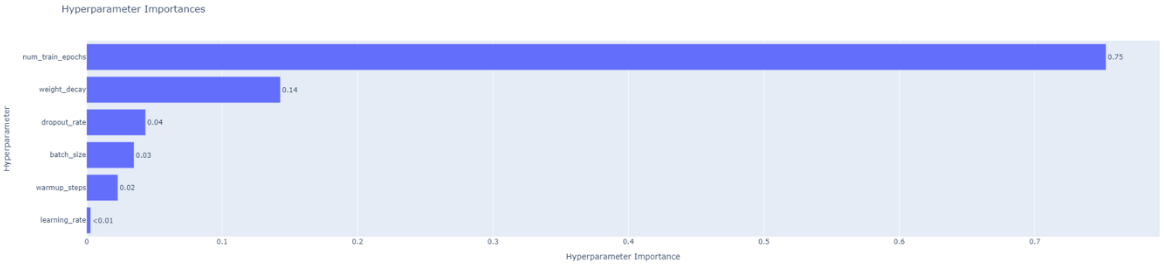


Figure 24 Hyperparameter Importances Plot

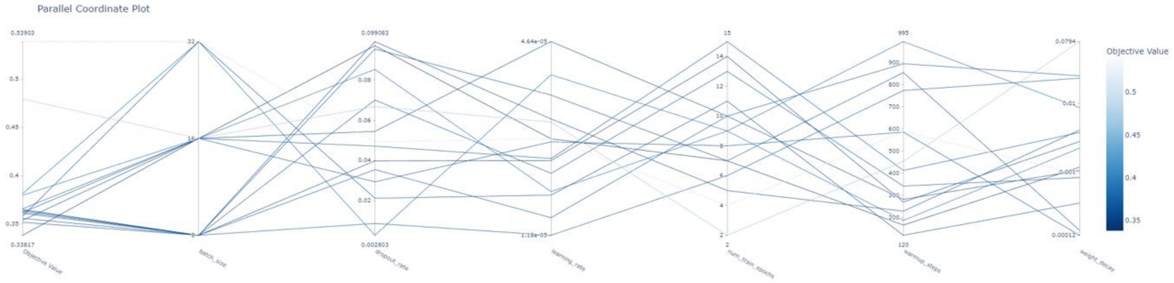


Figure 25 Parallel Coordinate Plot

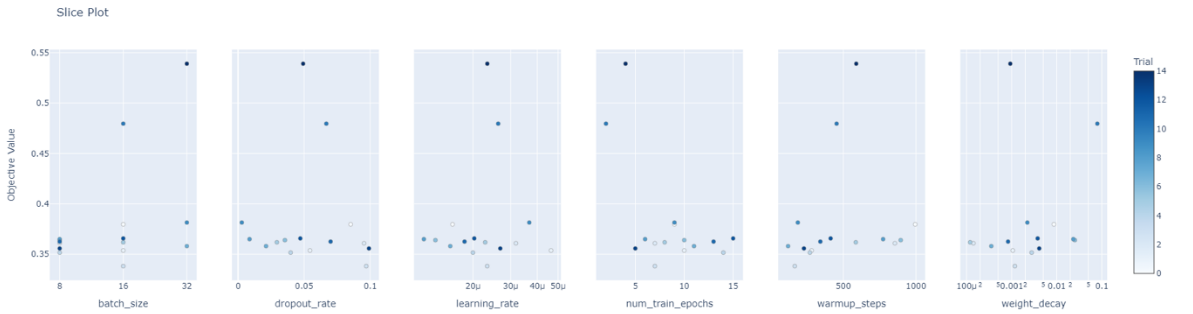


Figure 26 The Slice Plot

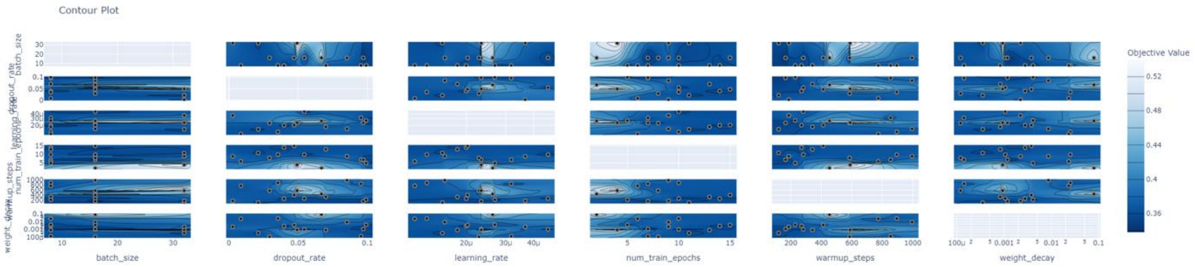


Figure 27 The Contour Plot

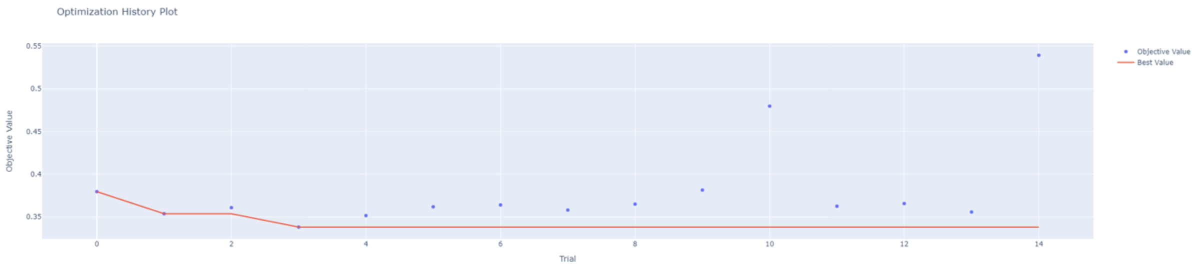


Figure 28 Optimization History Plot

The Hyperparameter Importances plot (Figure 24) clearly demonstrates that the number of training epochs is the most crucial factor, with an importance score of 0.75 (Bischl et al., 2017). This is followed by weight decay, with a score of 0.14. The remaining parameters - dropout rate, batch size, warmup steps, and learning rate - have considerably lower importance scores, all below 0.05.

The Parallel Coordinate Plot (Figure 25) provides a holistic view of how different hyperparameter combinations affect the objective value (Akiba et al., 2019). While it's challenging to discern clear patterns due to the plot's complexity, it appears that higher numbers of training epochs and lower learning rates tend to yield better results (lower objective values).

The Slice Plot (Figure 26) offers individual perspectives on each hyperparameter's impact on the objective value (Feurer and Hutter, 2019). For batch size, 16 or 32 seem to perform better than 8. Lower dropout rates generally show improved performance. Learning rates around 2e-5 to 3e-5 appear optimal. Higher numbers of training epochs correlate with better performance. Warmup steps show no clear trend, but values around 500-600 seem favourable. Weight decay values between 0.01 and 0.1 appear to work well.

The Contour Plot (Figure 27) illustrates the relationships between pairs of hyperparameters and their combined effect on the objective value (Li et al., 2017). The darker blue areas, indicating better performance, provide insights into optimal combinations of hyperparameters. However, the complexity of this plot makes it challenging to draw definitive conclusions without further analysis.

The Optimization History Plot (Figure 28) shows the progression of the best objective value over the course of the trials (Snoek et al., 2012). There is a rapid improvement in the first few trials, followed by a more gradual optimization. The best value appears to stabilize around 0.34 after approximately 4-5 trials.

4.2.5 Extract Best Hyperparameter and Train The Model

Based on the hyperparameter tuning process using Optuna, we got the best trial (Trial 22) achieved a validation loss of 0.3479 with the following hyperparameters:

Learning rate: 3.438751388233202e-05

Batch size: 16

Number of training epochs: 9

Warmup steps: 339

Weight decay: 0.00025884870760933614

Dropout rate: 0.04924173930490047

After applying these best hyperparameters into our model, we got:

Final Model - Accuracy: 0.5476190476190477

Final Model - Precision: 0.5529696772649662

Final Model - Recall: 0.5476190476190477

Final Model - F1 Score: 0.537791559299353

4.2.5.1 Compared to the Roberta model without fine-tuning:

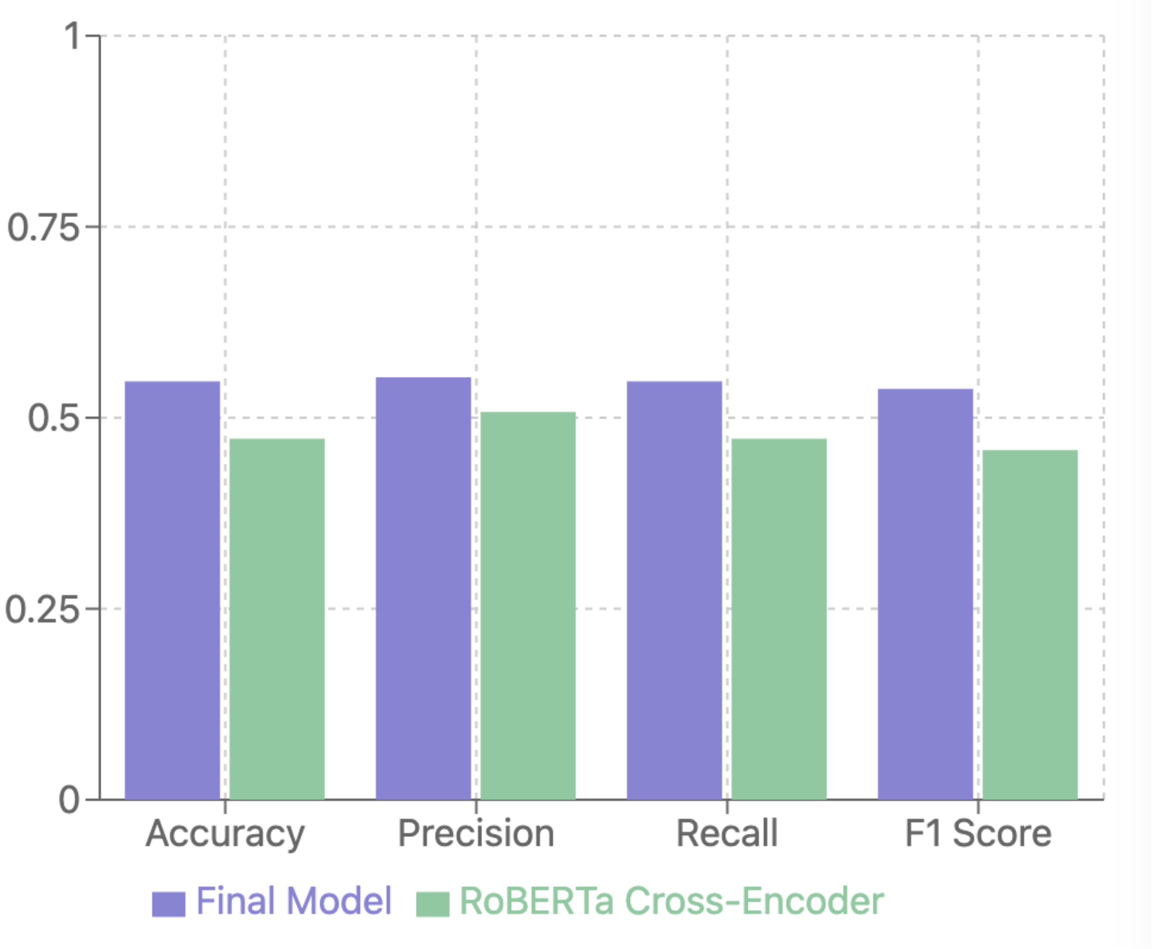


Figure 29 Final RoBERTa Performance Compared with RoBERTa Without Fine-tuned

The performance metrics of the Final Model (Figure 29) are compared with those of RoBERTa Cross-Encoder without fine-tuning in this bar chart. It is evident from the chart that the Final Model consistently outperforms raw RoBERTa across all metrics by approximately 10%, showcasing a well-balanced performance with scores around 55%. In contrast, RoBERTa exhibits more variability, reaching its peak precision at 51%.

Overall, both models exhibit moderate performance overall, with the Final Model's superiority most pronounced in accuracy, recall, and F1 score. Precision emerges as the strongest metric for both, suggesting a shared tendency to minimize false positives. The performance gap is smallest in precision, indicating potential areas for targeted improvement in RoBERTa. Despite the Final Model's clear advantage, both models' scores, ranging from 46% to 55%, indicate significant room for enhancement.

4.2.5.2 Decision Making for Model Selection

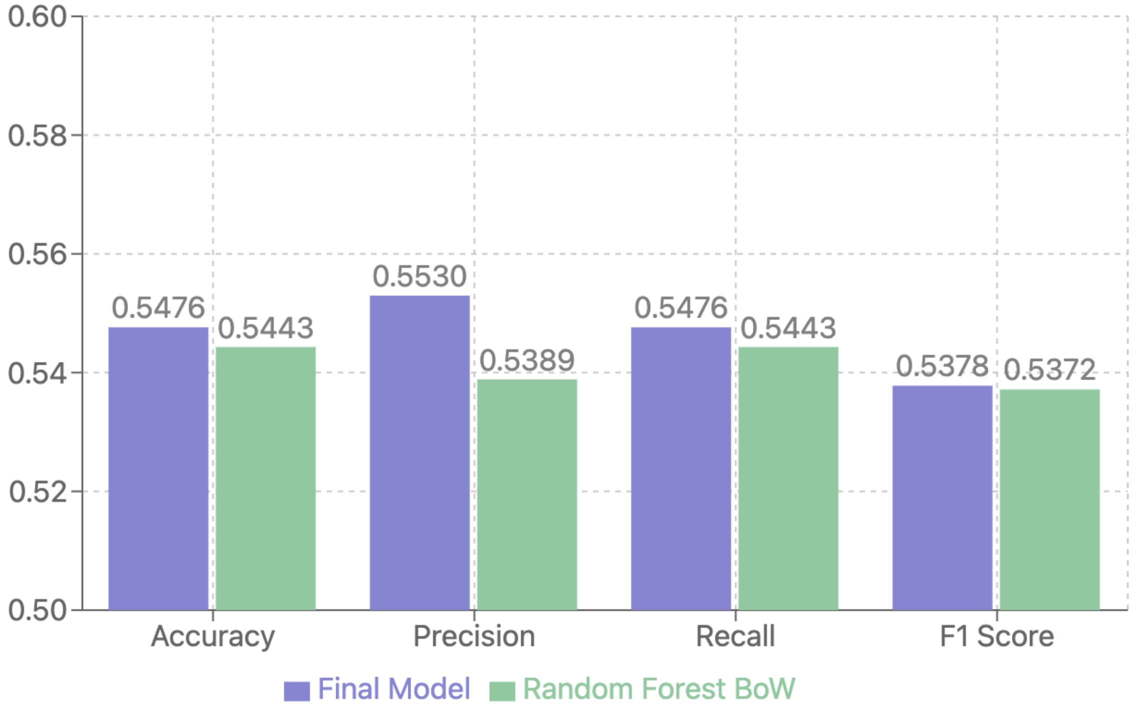


Figure 30 Final RoBERTa Performance Matrix Compared to Random Forest

After a series test for various word-embedding and models, the model adoption strategy for Abcam is to focus on improving the final model, with Random Forest + BoW as a fallback.

The selection of the final model is based not only on its current superior performance metrics, but also on the demonstrated potential for improvement through fine-tuning and architectural modifications in BERT-based models. Furthermore, an analysis of the training history suggests that there is more room for enhancement in the BERT model compared to the Random Forest approach. Additionally, BERT models have exhibited remarkable advancements with larger datasets and sophisticated training techniques, indicating a higher long-term performance ceiling. Moreover, if further improvements in the final model reach a plateau, the Random Forest model presents a competitive alternative with potentially lower computational requirements.

In conclusion, this strategy allows Abcam for pushing the boundaries of model performance with BERT while maintaining a solid, competitive alternative in Random Forest. It balances innovation with pragmatism, ensuring the project has both a high-potential path forward and a reliable fallback option.

Chapter 5: Discussion

5.1 Methodological Review and Considerations

The methodology employed in this study for developing an NLP solution to analyse Abcam's employee feedback has yielded significant insights, yet several aspects need critical examination and potential refinement.

5.1.1 Evaluation Metrics for Word Embedding Methods and Models

The current study primarily utilized accuracy, precision, recall, and F1 score as performance metrics for evaluating word embedding methods and classification models. While these metrics provide valuable insights, they may not fully encapsulate the complexities inherent in multi-label classification tasks. Schütze et al. (2008) argue that a more comprehensive evaluation framework is crucial for nuanced understanding of model performance. To address this limitation, future iterations of this research could benefit from incorporating additional evaluation metrics. Hamming Loss, as described by Tsoumakas and Katakis (2007), quantifies the fraction of misclassified labels, offering insights into the overall error rate across all labels. Given the potential overlap in cultural dimensions, the Jaccard Similarity Score, as discussed by Sorower (2010), could provide a more nuanced measure of prediction accuracy.

Macro and Micro-averaged Metrics, as proposed by Yang (1999), could offer a more balanced view of performance across all classes, especially given the class imbalance in our dataset. Additionally, Chicco and Jurman (2020) argue that Matthew's Correlation Coefficient (MCC) provides a more informative and truthful score in evaluating binary classifications, which could be extended to our multi-label scenario. The integration of these metrics could facilitate a more robust evaluation framework, potentially guiding more informed decisions in model selection and optimization.

5.1.2 Data Split Strategy Review

The current study employed an 80-20 train-test split, a common practice in machine learning. However, this approach may present limitations, particularly given the relatively small size of our labelled dataset (n=1,200). Kohavi (1995) suggests that for smaller datasets, this splitting strategy may lead to high variance in performance estimates.

To address these concerns, it is worth considering adjusting the train-test ratio and evaluating model performance. Reducing the size of the training set while increasing the test set size to 30% or 40% could provide a more robust estimation of real-world performance. However, as cautioned by Guyon (1997), careful evaluation of the trade-offs between training set size and estimate reliability is necessary, ensuring that the test set does not exceed a certain threshold that may negatively impact model training.

5.2 Model Performance Analysis and Future Directions

While the fine-tuned RoBERTa-base model demonstrated improvement over the baseline, its performance did not meet initial expectations. This section critically examines potential reasons for this suboptimal performance and proposes avenues for future improvement.

5.2.1 Factors Influencing Model Performance

Several factors may have influenced the model's performance. Firstly, the limited dataset, consisting of only 1,200 labelled examples, might have restricted the model's ability to learn complex patterns required for high-performance multi-label classification. Sun et al. (2017) emphasizes the crucial role of dataset size in deep learning performance. Despite attempts to address class imbalance through weight adjustment, the significant imbalance in cultural dimension labels could still impact the model's performance, as highlighted by Buda et al. (2018). The classification problem being multi-label in nature and involving nuanced and potentially overlapping cultural dimensions presents a challenging task. Zhang and Zhou (2014) note that multi-label classification inherently involves higher complexity compared to traditional single-label classification. Additionally, our exploratory data analysis reveals a prevalence of short comments which may limit the contextual information available to the model—a challenge discussed by Shen et al. (2021) in short text classification.

5.2.2 Proposed Enhancements for Future Research

The current findings of this project suggest several potential improvements for future research. Experimenting with larger pre-trained models, such as RoBERTa-large or GPT-3, has the potential to capture more intricate patterns in the data, as demonstrated by Brown et al. (2020) who showcased the superior performance of larger language models across various NLP tasks. In addition, Fine-tuning RoBERTa on a substantial corpus of HR-related or company-specific text could enhance the model's understanding of domain-specific language and context, as illustrated by Gururangan et al. (2020). Moreover, augmenting our training data by increasing its size and diversity through techniques like back-translation or synthetic example generation could also be beneficial, given the effectiveness demonstrated by Wei and Zou (2019). Combining different models using ensemble methodologies, such as our top-performing Random Forest and fine-tuned RoBERTa, may lead to improved overall performance based on Sagi and Rokach's (2018) review. Lastly, implementing an active learning approach where the model identifies informative examples for human labelling could facilitate more efficient expansion of our labelled dataset according to Settles' (2009) survey.

In conclusion, while the current model shows promise, there remains significant scope for enhancement. Future research should prioritize expanding the dataset, experimenting with more advanced modelling techniques, and refining the evaluation methodology to better capture the nuances of this complex classification task.

Chapter 6: Conclusion

This project set out to develop an advanced Natural Language Processing (NLP) solution for analysing employee feedback at Abcam, a prominent life sciences company. The primary objective was to automatically classify in employee comments from engagement surveys, aligning with Abcam's six cultural dimensions: Purpose, Customer Focus, Innovation, One Team, Growth and Development, and Diversity & Well-being.

The methodology commenced with rigorous data collection and preprocessing. Employee comments and corresponding questions from Peakon, Abcam's engagement platform, were aggregated. A focus group comprising experienced HR professionals and senior managers participated in defining and elaborating upon Abcam's cultural dimensions, establishing a framework for the manual labelling of 1,200 comments. This labelled dataset served as the foundation for model training and evaluation.

Exploratory Data Analysis (EDA) yielded several significant insights. The distribution of labels across cultural dimensions revealed "One Team" as the most frequently discussed topic, while "Customer" had the lowest representation. Generally, the comments exhibited a positive sentiment bias, with high scores predominating across all dimensions. Analysis of comment length indicated that most feedback was concise, with the majority falling within the 0–500-character range.

In addressing the challenge of processing textual data, various word embedding techniques were explored, including Bag of Words (BoW), TF-IDF, Word2Vec, and BERT. Performance evaluation utilizing a simple regression model demonstrated that BoW outperformed other methods, likely due to its efficacy in handling short, straightforward text. BERT, while not surpassing BoW, exhibited competitive results and potential for improvement with fine-tuning.

For the classification task, multiple models were tested, including Support Vector Machines (SVM), Random Forests, Neural Networks, and RoBERTa. Interestingly, traditional machine learning approaches, particularly Random Forest and SVM, consistently outperformed more complex models like Neural Networks and RoBERTa across all metrics (accuracy, precision, recall, and F1 score). The dataset mostly consists of short texts that do not require the model to capture a strong semantic relationship, which is likely the reason for this phenomenon.

To enhance the performance of the deep learning approach, a cross-encoder architecture using RoBERTa was implemented, allowing for improved contextual understanding by analysing question-comment pairs holistically. Class imbalance was addressed through weight adjustment, and Optuna was employed for hyperparameter tuning. This fine-tuned model demonstrated a significant improvement of approximately 10% across all metrics compared to the base RoBERTa model.

The final fine-tuned model achieved an accuracy of 0.5476, precision of 0.5530, recall of 0.5476, and an F1 score of 0.5378. While these results show clear improvement over the baseline, they indicate that there is still considerable room for enhancement in the model's performance.

Key findings from this project include the effectiveness of simpler models (BoW, Random Forest, SVM) in handling short, straightforward text data typical of employee feedback. The potential of advanced models like BERT and RoBERTa was also highlighted, especially when fine-tuned and adapted to the specific context of organizational culture analysis. The importance of addressing class imbalance and careful hyperparameter tuning in improving model performance was underscored. Furthermore, the value of combining questions with comments to provide additional context for more accurate classification was demonstrated.

This research exemplifies the feasibility of utilizing NLP techniques to analyse organizational culture through employee feedback. It provides Abcam with a foundation for automating the analysis of large volumes of employee comments, potentially leading to more data-driven decision-making in human resource management.

As we discussed at Chapter 5, future research directions could focus on expanding the labelled dataset, exploring more advanced NLP techniques or larger BERT models, and integrating the model into Abcam's existing HR systems. Additionally, investigating methods to improve performance on underrepresented cultural dimensions and incorporating longitudinal analysis could provide deeper insights into Abcam's evolving organizational culture

In addition, this project's approach was significantly informed by a comprehensive literature review that spanned three interconnected domains: organizational culture, Natural Language Processing (NLP), and their intersection in HR analytics. The review traced the evolution of organizational culture concepts, from early studies on group norms to complex frameworks like the Competing Values Framework, providing a solid theoretical foundation for our cultural dimension classification. In the NLP domain, our exploration of its historical development, from rule-based systems to modern deep learning approaches, guided our choice of advanced techniques like BERT and RoBERTa. The review of NLP applications in organizational culture analysis revealed a gap in leveraging these advanced models for multi-label classification of employee feedback, which our project directly addressed. This interdisciplinary literature review not only contextualized our research within broader academic discourse but also informed our methodological choices, such as the implementation of cross-encoder architecture and the focus on addressing class imbalance. By bridging theoretical frameworks of organizational culture with cutting-edge NLP techniques, our project contributes to the growing field of AI-driven organizational analysis, offering both academic insights and practical applications for HR management in the life sciences industry.

In conclusion, while the current model's performance leaves room for improvement, this project represents a significant advancement in leveraging artificial intelligence and deep learning model to understand and shape organizational culture. It opens up new possibilities for data-driven HR practices at Abcam and potentially in the broader life sciences industry. The findings not only contribute to the expanding body of literature on the application of Natural Language Processing (NLP) in organizational studies but also present a well-developed transformer-based multi-classifier, offering a practical framework for companies aiming to leverage the potential of Artificial Intelligence (AI) in organisational culture management.

Bibliography

Abcam (2022). Antibodies Proteins Kits and Reagents for Life Science | Abcam. [online] Abcam.com. Available at: https://www.abcam.com/engb?gclsrc=aw.ds&gad_source=1&gclid=CjwKCAjw4f6zBhBVEiwATEHFVnE_mv8o7SbvW6pIMjpPKv3pH2VjEEibxWCmpNbm_VDPJVkquziG-hoC0DAQAvD_BwE&gclsrc=aw.ds [Accessed 30 Jun. 2024].

Agarwal A., Xie B., Vovsha I., Rambow O., & Passonneau R. (2011). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011), 30--38.

Akiba T., Sano S., Yanase T., Ohta T., & Koyama M. (2019). Optuna: A Next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Alemzadeh H., & Devarakonda M. (2017). An NLP-based cognitive system for disease status identification in electronic health records. In 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI).

Angrave D., Charlwood A., Kirkpatrick I., Lawrence M., & Stuart M. (2016). HR and analytics: why HR is set to fail the big data challenge. Human Resource Management Journal, 26(1), pp.1-11.

Apell D., & Eriksson Y. (2023). AI Healthcare Technology Innovation Systems: A Case Study of West Sweden. Sustainability, 15(5), p.4491.

Bao Y., Fang H., & Zhang J. (2014). TopicMF: Simultaneously exploiting ratings and reviews for recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 28, No. 1).

Baye Z., & Yusuf K. (2023). Top management support and business model impact on biotechnology venture performance: highlighting mediation of company culture and moderation of innovation. Journal of Commercial Biotechnology, 28(2).

Bergstra J., & Bengio Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Bird S., Klein E., & Loper E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media Inc.

Bischl B., Richter J., Bossek J., Horn D., Thomas J., & Lang M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. arXiv preprint arXiv:1703.03373.

Bollen J., Mao H., & Zeng X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

Breiman L. (2001). Random forests. Machine learning, 45(1), pp.5-32.

Brockopp D.Y. (1983). What is NLP? The American Journal of Nursing, 83(7), pp.1012--1014.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., & Agarwal S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Buda M., Maki A., & Mazurowski M.A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, pp.249-259.

Cameron K.S., & Quinn R.E. (2011). Diagnosing and Changing Organizational Culture: Based on the Competing Values Framework. 3rd ed. San Francisco: Jossey-Bass.

Cer D., Yang Y., Kong S.Y., Hua N., Limtiaco N., St. John R., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., & Sung Y.H. (2018). Universal sentence encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 169-174).

Chatman J., & Choi A. (2019). Measuring Organizational Culture: Converging on Definitions and Approaches to Advance the Paradigm. [online] Available at: https://faculty.haas.berkeley.edu/chatman/papers/Chatman_Choi_Measuring%20Organizational%20Culture_2019.pdf.

Cheng M., Hackett R.D., & Wang D. (2021). A multi-level framework for understanding the role of artificial intelligence in human resource management practices: an employee perspective. The International Journal of Human Resource Management, 32(19), pp.4111-4142.

Chicco D., & Jurman G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), pp.1-13.

Chintalapudi N., Battineni G., Di Canio M., Sagaro G.G., & Amenta F. (2021). Text mining with sentiment analysis on seafarers' medical documents. International journal of information management data insights, 1(1), p.100005.

Cooke R.A., & Rousseau D.M. (1988). Behavioral norms and expectations: A quantitative approach to the assessment of organizational culture. Group & Organization Studies, 13(3), 245-273.

Cortes C., & Vapnik V. (1995). Support-vector networks. Machine learning, 20(3), pp.273-297.

Creswell J.W., & Plano Clark V.L. (2017). Designing and Conducting Mixed Methods Research. 3rd ed. Thousand Oaks: SAGE Publications.

Cui Y., Jia M., Lin T.Y., Song Y., & Belongie S. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9268-9277).

Cyr J. (2015). The Pitfalls and Promise of Focus Groups as a Data Collection Method. Sociological Methods & Research [online] 45(2), pp.231--259. doi:https://doi.org/10.1177/0049124115570065.

Das B.L., & Tripathy S. (2022). Factors affecting employee engagement in IT industry: A systematic literature review. International Journal of Human Capital and Information Technology Professionals, 13(1), pp.1-23.

Demner-Fushman D., Chapman W.W., & McDonald C.J. (2009). What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 42(5), 760-772.

Denison D. (1990). Corporate culture and organizational. New York: Wiley. Dike P. (2013). The impact of workplace diversity on organizations. Dobbin F. & Jung J. (2010). Corporate board gender diversity and stock performance: The competence gap or institutional investor bias. NCL Rev* 89 p.809.

Denison D., Nieminen L., & Kotrba L. (2012). Diagnosing organizational cultures: A conceptual and empirical review of culture effectiveness surveys. European Journal of Work and Organizational Psychology, 23(1), pp.145--161. doi:https://doi.org/10.1080/1359432x.2012.713173.

Devlin J., Chang M.W., Lee K., & Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dorr B.J. (1993). Machine translation: A view from the lexicon. MIT Press.

Efron B., & Tibshirani R.J. (1994). An introduction to the bootstrap. CRC press.

Feurer M., & Hutter F. (2019). Hyperparameter optimization. In Automated Machine Learning (pp. 3-33). Springer Cham.

Finch D., Choi Y., & Kirsch P. (2019). AI chatbots in healthcare. HIMSS.

Ghahramani Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452-459.

Gonzalez J., Hakkani-Tur D., & Mehryar M. (2008). Call routing based on automatic identification of topics in telephonic conversations. IEEE Transactions on Audio Speech and Language Processing.

Green B.F., Wolf A.K., Chomsky C., & Laughery K. (1961). BASEBALL: An automatic question-answerer. Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference, 219-224.

Gu Y., Han Z., Zhou Y., Xie K., Zhang T., & Lv Z. (2023). Intelligent analysis of online comments based on deep learning and natural language processing. Expert Systems with Applications, 219, p.119451.

Gururangan S., Marasović A., Swayamdipta S., Lo K., Beltagy I., Downey D., & Smith N.A. (2020). Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Guyon I. (1997). A scaling law for the validation-set training-set size ratio. AT&T Bell Laboratories, pp.1-11.

Harris Z.S. (1954). Distributional structure. Word, 10(2-3), pp.146-162.

Harris Z.S. (1984). Distributional structure. In The Philosophy of Linguistics (pp. 26-47). Oxford University Press.

Hays D.G. (1967). Introduction to computational linguistics. Macdonald.

He H., & Garcia E.A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), pp.1263-1284.

Hellriegel D., & Slocum J.W. (2011). Organizational behavior. 13th ed. Mason, OH: South-Western Cengage Learning.

Heskett J. (2011). The Culture Cycle: How to Shape the Unseen Force that Transforms Performance. 1st edition ed. [online] PH Professional Business. Available at: https://ucl.alma.exlibrisgroup.com/view/action/uresolver.do?operation=resolveService&package_service_id=6998814360004761&amp;institutionId=4761&amp;customerId=4760.

Hirschberg J., & Manning C.D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Hofstede G. (1991). Cultures and Organizations: Software of the Mind. London: McGraw-Hill.

Hofstede G., Hofstede G.J., & Minkov M. (2010). Cultures and organizations: Software of the mind. Revised and expanded. McGraw-Hill.

Honnibal M., & Montani I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks, and incremental parsing. To appear, 7(1), pp.411-420.

Hopkins A. (2006). Studying organisational cultures and their effects on safety. Safety Science [online] 44(10), pp.875--889. doi:https://doi.org/10.1016/j.ssci.2006.05.005.

Humeau S., Shuster K., Lachaux M.A., & Weston J. (2020). Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969.

Hutchins W.J. (1986). Machine translation: Past, present, future. Ellis Horwood.

Hutchins W.J., & Somers H.L. (1992). An introduction to machine translation. Academic Press.

Isensee C., Teuteberg F., Griese K.M., & Topi C. (2020). The relationship between organizational culture, sustainability, and digitalization in SMEs: A systematic review. Journal of Cleaner Production, 275, p.122944.

Jaakson K., Jørgensen F., Tamm D., & Hämmal G. (2011). Investigating cultural influences on innovation: A comparison of Estonian and Danish biotechnology organizations. In: Innovation Systems in Small Catching-Up Economies: New Perspectives on Practice and Policy. Springer, pp.197--213.

Jain R., Jayakumar M., Christy V., Singh G., & Inamdar A.M. (2023). The effect of organizational culture on employee engagement and job satisfaction: A HR perspective. Journal of Survey in Fisheries Sciences, 10(1S), pp.6212--6225.

James C., & Collins C. (1994). Built to last: successful habits of visionary companies.

Johnson J.M., & Khoshgoftaar T.M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), pp.1-54.

Jones K.S. (1994). Natural language processing: a historical review. In Current issues in computational linguistics: in honour of Don Walker, pp.3--16.

Jurafsky D., & Martin J.H. (2018). Speech and language processing (2nd ed.). Prentice-Hall.

Katz D., & Kahn R. (2015). The social psychology of organizations. In: Organizational behavior 2. Routledge, pp.152–168.

Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., & Brown D. (2019). Text classification algorithms: A survey. Information, 10(4), p.150.

Krawczyk B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), pp.221-232.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Laumer S., & Maier C. (2023). Conversational Agents and Natural Language Processing in Information Systems and Human Resource Management: A Literature Review and Research Agenda. Communications of the Association for Information Systems, 52(1), p.9.

Laumer S., & Morana S. (2022). HR natural language processing: conceptual overview and state of the art on conversational agents in human resources management. In Handbook of research on artificial intelligence in human resource management, pp.226--242.

Lee M.Y. (2021). The relationship between organizational culture and job satisfaction: A multilevel analysis. Journal of Management & Organization, 27(1), pp.145-166.

Lewin K., Lippitt R., & White R.K. (1939). Patterns of aggressive behavior in experimentally created 'social climates'. The Journal of social psychology, 10(2), pp.269--299.

Li L., Jamieson K., DeSalvo G., Rostamizadeh A., & Talwalkar A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1), pp.6765-6816.

Lin, H.E. and McDonough, F. (2011). Investigating the Role of Leadership and Organizational Culture in Fostering Innovation Ambidexterity. *IEEE Transactions on Engineering Management*, 58(3), pp.497--509.doi:https://doi.org/10.1109/TEM.2010.2092781.

Lincoln J.R., Olson J., & Hanada M. (1978). Cultural effects on organizational structure: The case of Japanese firms in the United States. American Sociological Review, pp.829--847.

Liu B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

Manetje O., & Martins N. (2009). The relationship between organisational culture and organisational commitment. Southern African Business Review, 13(1), pp.87--111.

Martin J. (1980). Stories and scripts in organizational settings. Graduate School of Business, Stanford University.

Martin J. (2002). Organizational Culture: Mapping the Terrain. Thousand Oaks: SAGE Publications.

Mayur Wankhade, Sekhara C., & Kulkarni C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review [online] 55(7), pp.5731--5780. doi:https://doi.org/10.1007/s10462-022-10144-1.

McKinney W. (2010). Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 51-56.

Mihalcea R., Liu H., & Lieberman H. (2006). NLP (natural language processing) for NLP (natural language programming). In: Computational Linguistics and Intelligent Text Processing: 7th International Conference CICLing 2006, Mexico City, Mexico, February 19-25, 2006. Proceedings 7. Springer, pp.319--330.

Mikolov T., Chen K., Corrado G., & Dean J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mittra J. (2007). Life science innovation and the restructuring of the pharmaceutical industry: Merger, acquisition, and strategic alliance behaviour of large firms. Technology Analysis & Strategic Management, 19(3), pp.279--301.

Mroczkowski T. (2011). The new players in life science innovation: best practices in R&D from around the world. Pearson Education.

Nadkarni P.M., Ohno-Machado L., & Chapman W.W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), pp.544--551.

Neeraj A., Sharma Ali, & Muhammad A. Kabir (2024). A review of sentiment analysis: tasks, applications, and deep learning techniques. International Journal of Data Science and Analytics. [online] doi:https://doi.org/10.1007/s41060-024-00594-x.

Niosi J., & McKelvey M. (2018). Relating business model innovations and innovation cascades: the case of biotechnology. Journal of Evolutionary Economics, 28(5), pp.1081-1109.

Noorbehbahani F., & Kargar A. (2019). A novel approach for feature selection based on the deep learning in text classification. International Journal of Machine Learning and Cybernetics, 10(9), pp.2367-2383.

O'Connor J., & McDermott I. (2013). Principles of NLP: What it is, how it works. Singing Dragon.

Otter D.W., Medina J.R., & Kalita J.K. (2021). A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 32(2), pp.604-624.

Ouchi W. (1981). Theory Z: How American business can meet the Japanese challenge. Business Horizons, 24(6), pp.82–83.

Pandey S., & Pandey S.K. (2019). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. Organizational Research Methods, 22(3), pp.765--797.

Pang B., & Lee L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1--2), pp.1--135.

Pascale R.T., & Athos A.G. (1981). The art of Japanese management. Business Horizons, 24(6), pp.83-85.

Poria S., Majumder N., Hazarika D., Cambria E., Gelbukh A., & Hussain A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. IEEE Intelligent Systems, 33(6), pp.17-25.

Pustejovsky J., & Stubbs A. (2012). Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media Inc.".

Quiñonero-Candela J., Sugiyama M., Schwaighofer A., & Lawrence N.D. (2022). Dataset shift in machine learning. MIT Press.

Radford A., Wu J., Child R., Luan D., Amodei D., & Sutskever I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9.

Radu C. (2023). Fostering a positive workplace culture: Impacts on performance and agility. In: Human Resource Management-An Update. IntechOpen.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., & Liu P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, pp.1-67.

Reshamwala A., Mishra D., & Pawar P. (2013). Review on natural language processing. IRACST Engineering Science and Technology: An International Journal (ESTIJ), 3(1), pp.113--116.

Resnik P., & Lin J. (2010). Evaluation of NLP systems. In The handbook of computational linguistics and natural language processing, pp.271--295.

Ruder S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Ruder S. (2018). Neural transfer learning for natural language processing. Ph.D. dissertation, National University of Ireland, Galway.

Sagi O., & Rokach L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1249.

Sanh V., Debut L., Chaumond J., & Wolf T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schein E.H. (1990). Organizational culture. American Psychological Association.

Schein E.H. (2010). Organizational Culture and Leadership. [online] Available at: https://ia800809.us.archive.org/14/items/EdgarHScheinOrganizationalCultureAndLeadership/Edgar_H_Schein_Organizational_culture_and_leadership.pdf.

Schütze H., Manning C.D., & Raghavan P. (2008). Introduction to information retrieval (Vol. 39). Cambridge University Press.

Sechidis K., Tsoumakas G., & Vlahavas I. (2011). On the stratification of multi-label data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 145-158). Springer, Berlin, Heidelberg.

Serpa S. (2016). An overview of the concept of organisational culture. International business management, 10(1), pp.51--61.

Settles B. (2009). Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.

Shen T., Ott M., Ainslie J., & Ravi S. (2021). Convolutional Neural Networks for Short Text Understanding. arXiv preprint arXiv:2102.09570.

Smircich L. (1983). Concepts of culture and organizational analysis. Administrative Science Quarterly, 28(3), 339-358.

Snoek J., Larochelle H., & Adams R.P. (2012). Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25.

Sokolova M., & Lapalme G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), pp.427-437.

Sorower M.S. (2010). A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis, 18, pp.1-25.

Sparck Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), pp.11-21.

Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), 36(2), pp.111-133.

Sun C., Qiu X., Xu Y., & Huang X. (2019). How to fine-tune bert for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194-206). Springer, Cham.

Sun C., Shrivastava A., Singh S., & Gupta A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 843-852).

Tait J. (2007). Systemic interactions in life science innovation. Technology Analysis & Strategic Management, 19(3), pp.257--277.

Tay Y., Dehghani M., Bahri D., & Metzler D. (2020). Efficient transformers: A survey. arXiv preprint arXiv:2009.06732.

Tian X., Tan X., Xue J., Qi L., & Jiang C. (2023). Understanding Human Resource Management: Designing a Machine Learning-Based HR Recruitment System Using LSA, BERT, and SVM. Computers, Materials & Continua, 74(3), pp.6389-6405.

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., & Rodriguez A. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Tsoumakas G., & Katakis I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3(3), pp.1-13.

Tumasjan A., Sprenger T.O., Sandner P.G., & Welpe I.M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Van der Maaten L., & Hinton G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), pp.2579-2605.

Van Maanen J., & Barley S.R. (1984). Occupational communities: Culture and control in organizations. Research in organizational behavior, 6, pp.287-365.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., & Polosukhin I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Verbeke W., Volgering M., & Hessels M. (1998). Exploring the Conceptual Expansion within the Field of Organizational Behaviour: Organizational Climate and Organizational Culture. Journal of Management Studies [online] 35(3), pp.303--329. doi:https://doi.org/10.1111/1467-6486.00095.

Wang B., Wang A., Chen F., Wang Y., & Kuo C.C.J. (2019). Evaluating word embedding models: methods and experimental results. APSIPA Transactions on Signal and Information Processing [online] 8, p.e19. doi:https://doi.org/10.1017/ATSIP.2019.12.

Wang Y., Huang M., Zhao L., & Zhu X. (2016). Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing, pp.606-615.

Wei J.W., & Zou K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

Wilkins A.L., & Ouchi W.G. (1983). Efficient cultures: Exploring the relationship between culture and organizational performance. Administrative science quarterly, pp.468--481.

Wilson T., Wiebe J., & Hoffmann P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp.347--354.

Wold S., Esbensen K., & Geladi P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), pp.37-52.

Wolpert D.H., & Macready W.G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), pp.67-82.

Workday (2023). Workday Peakon Employee Voice | Workday. [online] Workday.com. Available at: https://www.workday.com/en-gb/products/employee-voice/overview.html [Accessed 30 Jun. 2024].

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., ... & Dean J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Yang B., & Cardie C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.325--335.

Yang P., Sun X., Li W., Ma S., Wu W., & Wang H. (2018). SGM: sequence generation model for multi-label classification. arXiv preprint arXiv:1806.04822.

Yang Y. (1999). An evaluation of statistical approaches to text categorization. Information retrieval, 1(1), pp.69-90.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R.R., & Le Q.V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

Yin R.K. (2018). Case Study Research and Applications: Design and Methods. 6th ed. Thousand Oaks: SAGE Publications.

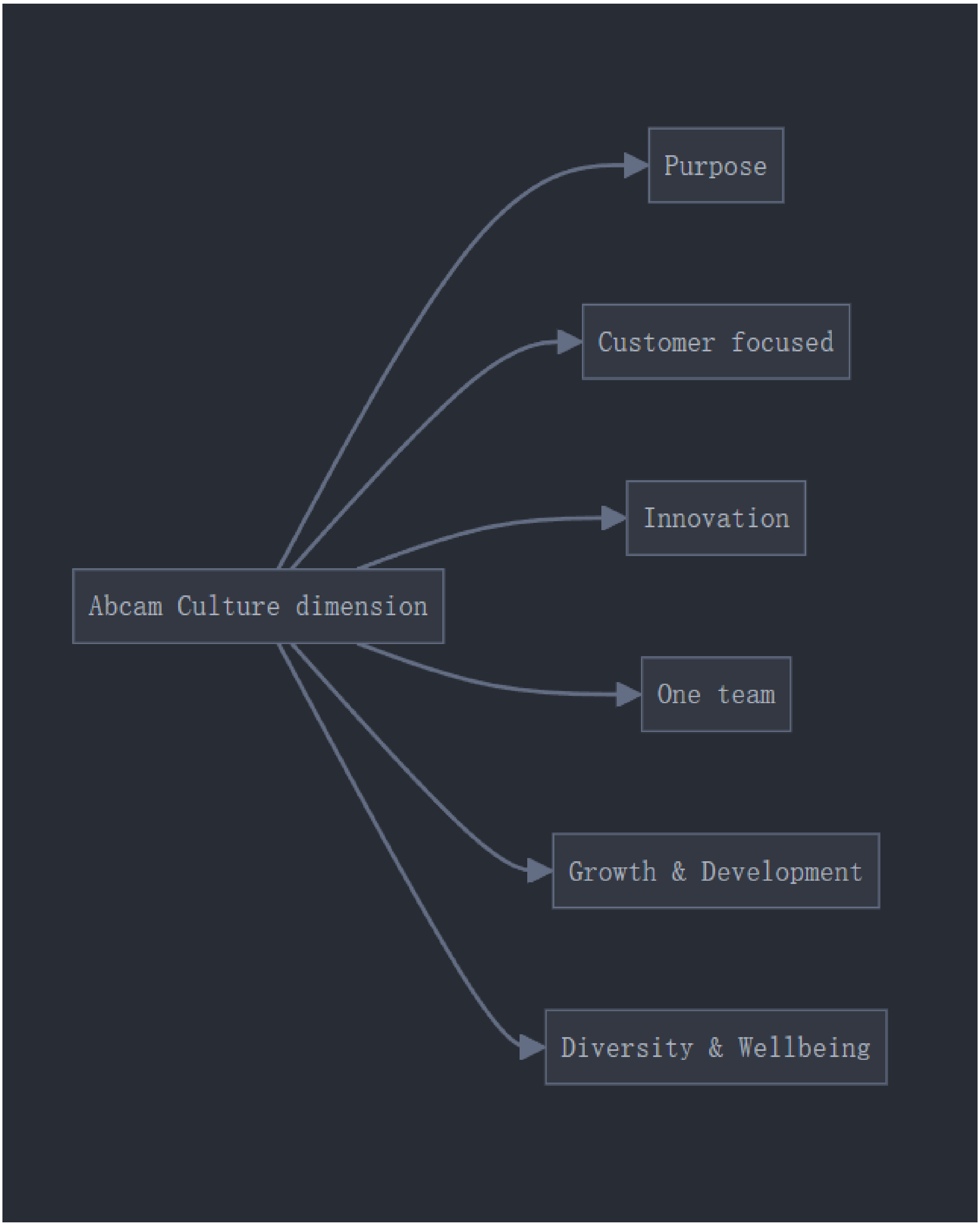
Zhang L., Wu Z., & Zhao Y. (2023). A Review of Machine Learning and Deep Learning Techniques for Sentiment Analysis. Applied Sciences, 13(7), p.4605.

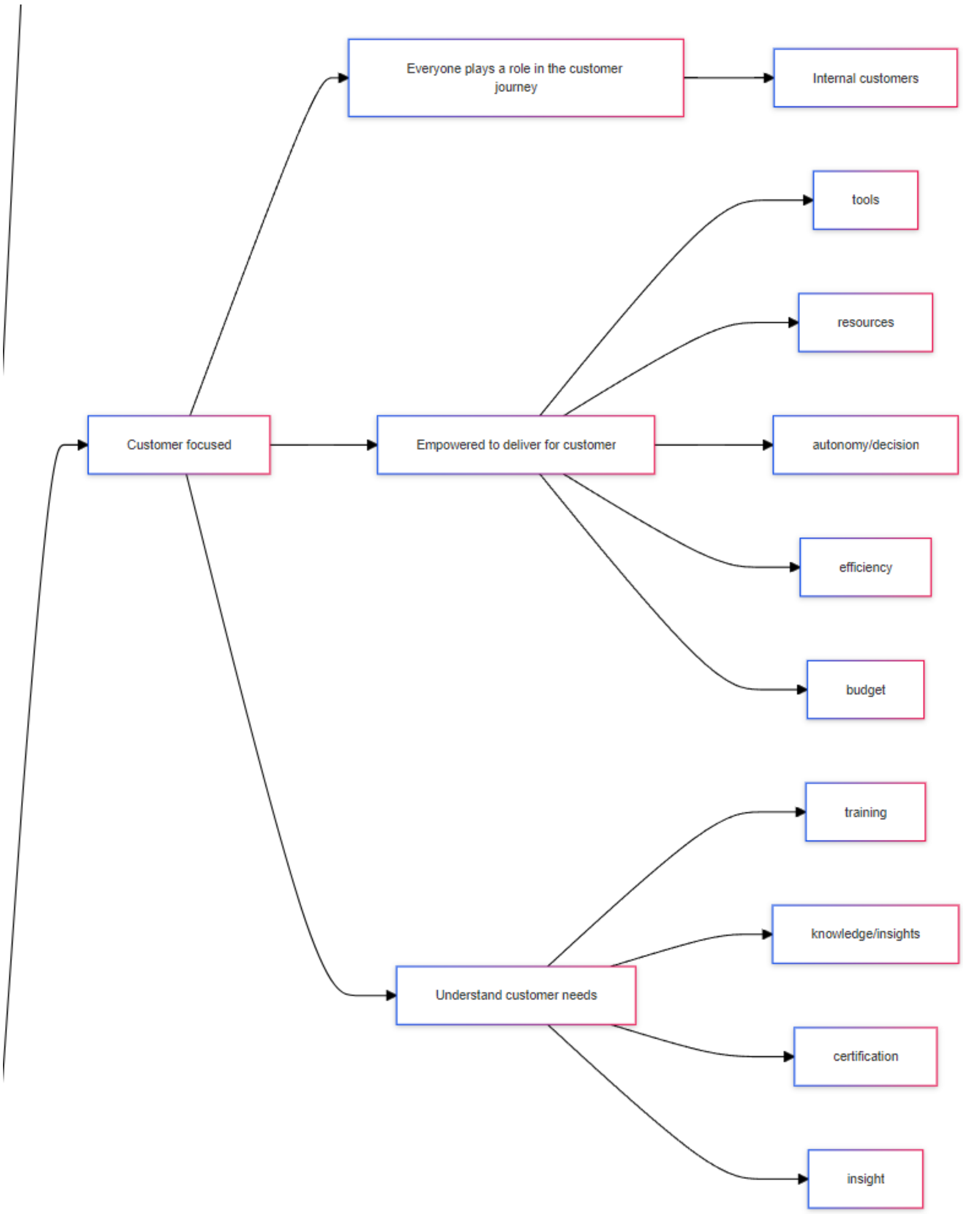
Zhang M.L., & Zhou Z.H. (2014). A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8), pp.1819-1837.

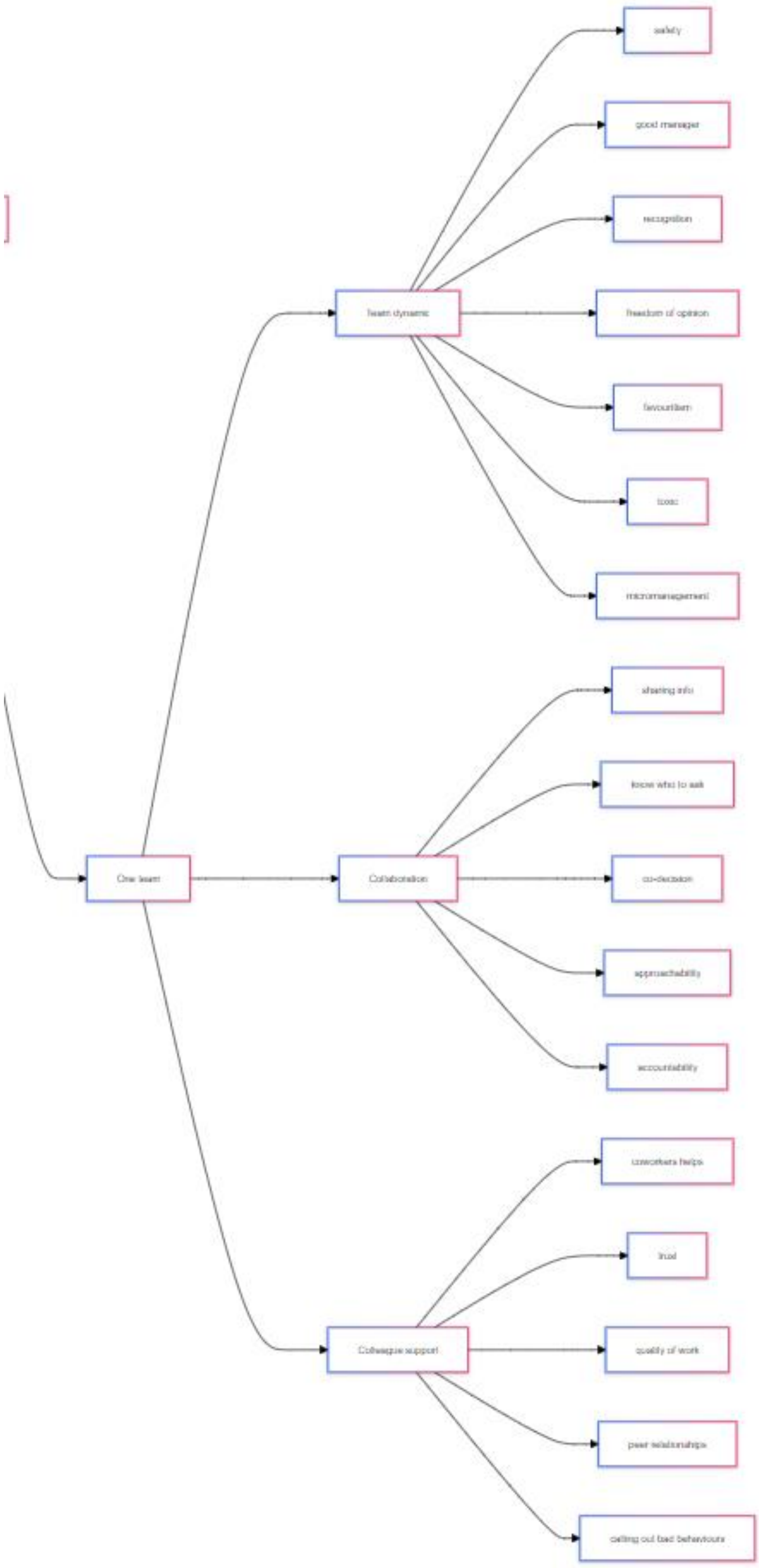
Zhang S., Gu Y., Usuyama N., Woldeselassie T., Yu Y., Liang P., & Wu T. (2023). Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. arXiv preprint arXiv:2307.06439.

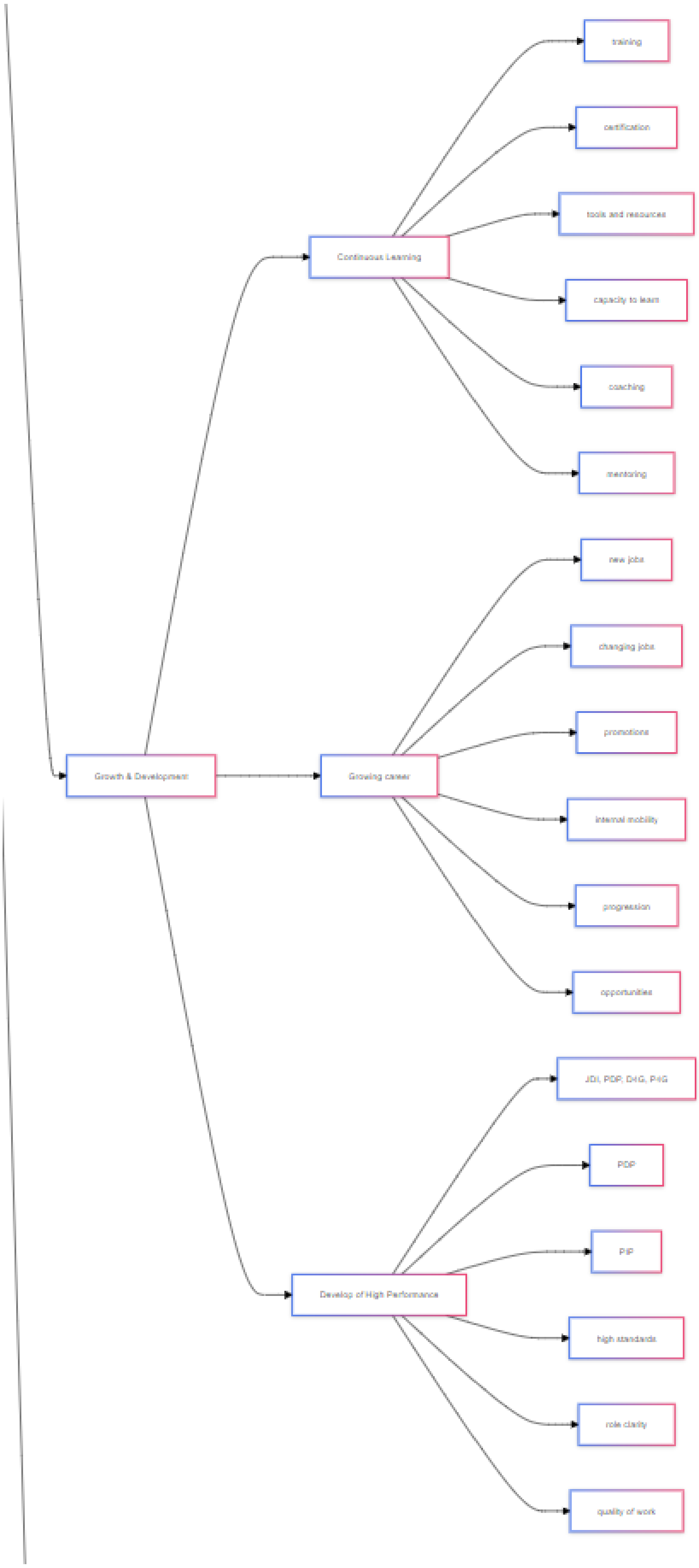
Appendix

Relationship Diagrams Between Abcam Culture Dimensions and Keywords

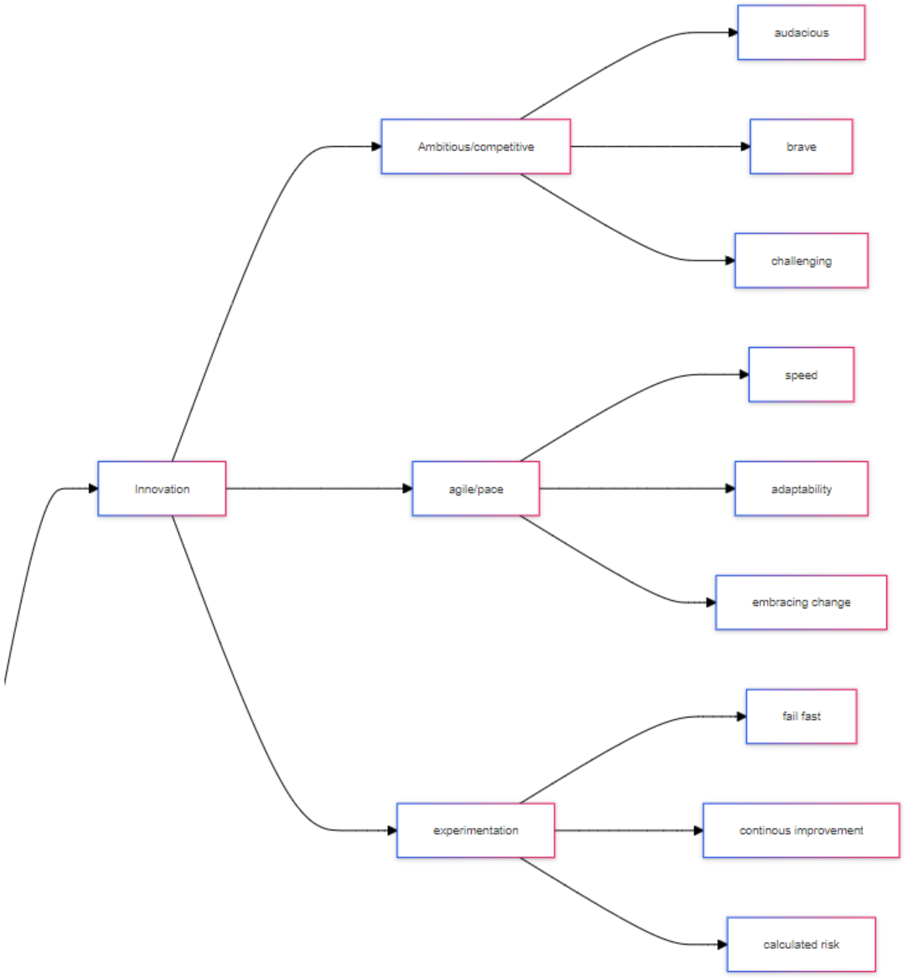


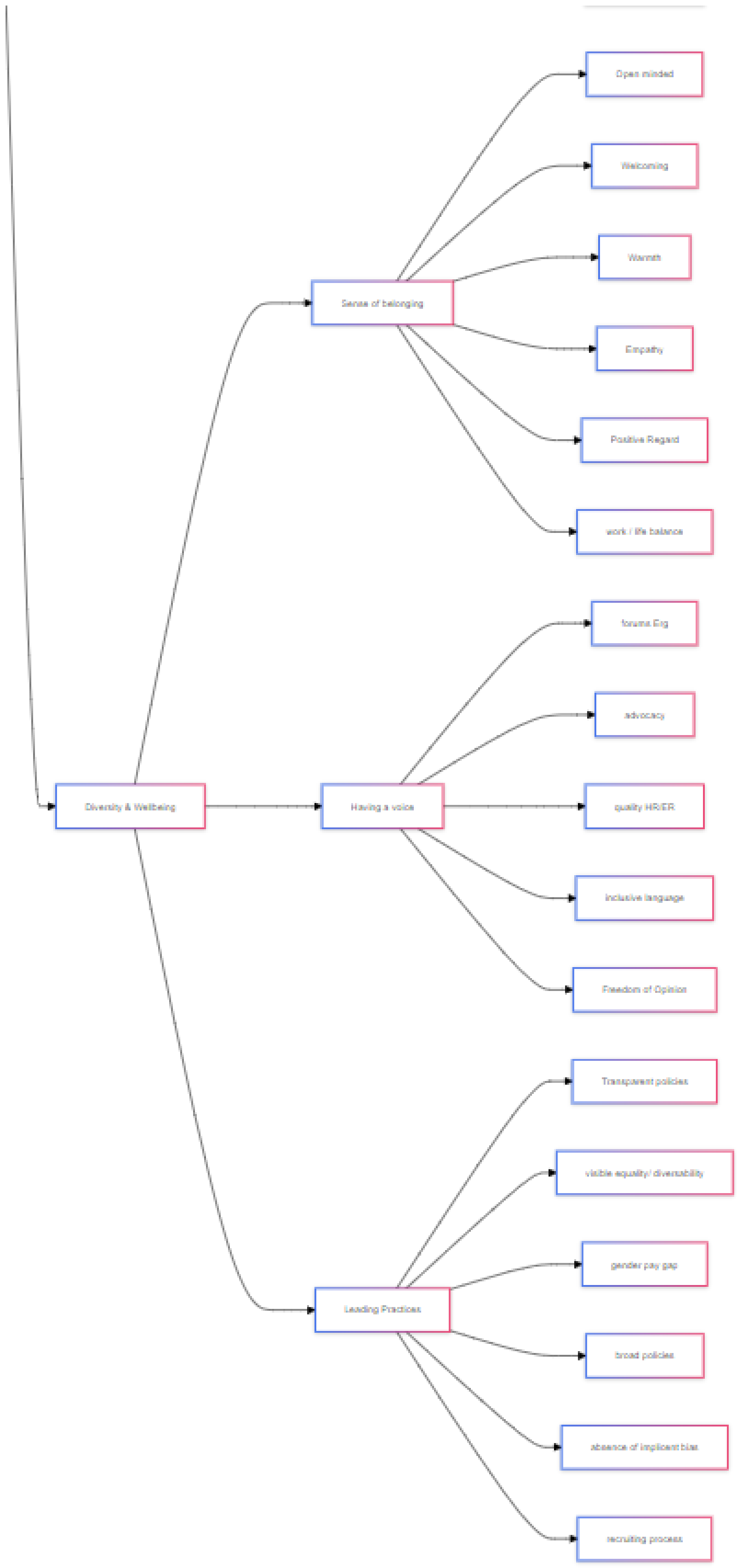

Everyone plays a role in the customer journey
Internal customers
tools
resources
Customer focused
Empowered to deliver for customer
autonomy/decision
efficiency
budget
training
knowledge/insights
Understand customer needs
certification
insight


safety
good manager
recognition
team dynamic
freedom of opinion
favouritism
toxic
micromanagement
sharing info
know who to ask
One team
Collaboration
co-decision
approachability
accountability
coworkers helps
trust
Colleague support
quality of work
peer relationships
calling out bad behaviours



training
certification
tools and resources
Continuous Learning
capacity to learn
coaching
mentoring
new jobs
changing jobs
promotions
Growth & Development
Growing career
internal mobility
progression
opportunities
JDI, PDP, D4G, P4G
PDP
PIP
Develop of High Performance
high standards
role clarity
quality of work


Innovation
Ambitious/competitive
audacious
brave
challenging
agile/pace
speed
adaptability
embracing change
experimentation
fail fast
continous improvement
calculated risk


Open minded
Welcoming
Warmth
Sense of belonging
Empathy
Positive Regard
work / life balance
forums Erg
advocacy
Diversity & Wellbeing
Having a voice
quality HR/ER
inclusive language
Freedom of Opinion
Transparent policies
visible equality/ diversability
gender pay gap
Leading Practices
broad policies
absence of implicent bias
recruiting process