Master Computer Science

Discovering Organizational Culture from Text:
An NLP Pipeline for Cultural Analysis

Name:	Ivar Martens
Student ID:	s2990806
Date:	22/04/2026
Specialisation:	Artificial Intelligence
1st supervisor:	Flor Miriam Plaza del Arco
2nd supervisor:	Christoph Johann Stettina

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Organizational culture is recognized as a key determinant of employee performance, wellbeing, and long-term organizational performance. Despite its importance, culture remains difficult to measure accurately in practice. Traditional employee surveys often reduce complex experiences to numerical scores, while qualitative interview methods provide richer insight but are labour-intensive, costly, and difficult to scale. This thesis investigates how Natural Language Processing (NLP) can be used to quantify organizational culture from employee interview data and translate these findings into actionable organizational insights.

To address this challenge, an end-to-end NLP pipeline was developed using interview transcripts from three organizations, comprising 1,383 employee responses. The pipeline combines three complementary analytical components: First, supervised multi-label text classification maps responses onto the six dimensions of the Organizational Culture Survey (OCS). Second, topic modelling identifies emergent and organization-specific themes beyond predefined frameworks. Third, sentiment analysis estimates the evaluative tone of employee statements, enabling prioritization of positive and negative cultural patterns.

Across the three NLP components, results show a clear trade-off between predictive performance, interpretability, and computational efficiency. For classification, fine-tuned transformer models, particularly BERTje with threshold calibration, achieved the most reliable results, demonstrating that validated culture dimensions can be operationalized accurately at scale. For topic modelling, GPT-based methods produced the most meaningful and business-relevant themes, outperforming traditional approaches in capturing nuanced organizational narratives despite higher computational cost and less coherent topic structures. For sentiment analysis, GPT also achieved the strongest performance, especially in detecting subtle ordinal differences in employee feedback, indicating that zero-shot LLM sentiment analysis is already a viable option when labelled domain data are unavailable.

Finally, the outputs of all components were integrated into a prototype dashboard that visualizes cultural dimensions, recurring themes, sentiment patterns, and temporal developments. Overall, this thesis demonstrates that NLP can provide a scalable, theory-grounded, and practically relevant approach to organizational culture diagnostics. The findings further show that model choice should depend on task requirements: fine-tuned transformers remain strongest for supervised classification, while LLMs offer clear advantages for nuanced interpretation and low-resource tasks.

Contents

Chapter 1 Introduction

"Culture is to an organization what personality is to an individual" (Van Der Post et al., 1997). This organizational 'personality' is not merely descriptive but economically consequential. Organisational culture (OC) shapes how employees interpret situations, make decisions, and coordinate their actions, thereby exerting a measurable influence on job performance. Empirical evidence shows that stronger, well-aligned organizational cultures are associated with higher employee performance. In turn, improved performance contributes to greater job satisfaction and overall life satisfaction within the organization (Aggarwal, 2024). Beyond its direct effects on individuals, OC constitutes a strategic asset at the firm level. Because culture is inherently socially complex and causally ambiguous, it resists imitation by competitors. As a result, the advantages it generates are not only impactful but also durable, positioning culture as a critical source of sustained competitive advantage and long-term superior performance (Yilmaz and Ergun, 2008).

Despite its importance, gaining a clear understanding of an organization's true culture remains a critical yet challenging task for management. Employees' perceptions are often difficult to capture due to dependencies inherent in the employer–employee relationship. Often organizations try to overcome these barriers through satisfaction surveys. However, traditional satisfaction surveys typically reduce complex experiences to numerical ratings on standardized scales (e.g., 1–5). Although these instruments are easy to analyse and can provide broad indicators, they rarely uncover participants' genuine opinions (Hansen and Świderska, 2024). As a result, they often fail to reveal the underlying cultural dynamics within an organization. External, independent facilitators, such as De Hofnar[1], seek to bridge this gap by engaging employees in open dialogue about their lived experiences and reflecting these perspectives back to management in an anonymous and unbiased manner. Such interviews and discussions generate rich qualitative data that can surface unspoken cultural dynamics otherwise hidden in survey metrics. However, as Evans (2014) note, conventional qualitative methods are "labor-intensive and do not scale well beyond small datasets". And reviews of large qualitative projects warn of intensified burdens of data management, data overload, time-intensive coding and analysis, and team coordination as scale increases (Hunt et al., 2011).

The reliance on manual interpretation of rich textual data creates several barriers to effective elicitation of OC. First, consistency is difficult to maintain across projects, as human interpretation is inherently subjective and the amount of information a person can reliably process is limited. An issue that becomes particularly overwhelming when analysing data from more than 100 interviews in a single project. Second, scalability is constrained: as the number of interviews and clients increases, analysis becomes increasingly time-consuming and costly. Third, the current approach does not readily translate into a software-supported product, which restricts De Hofnar's capacity to innovate within the field of human capital management. Together, these challenges hinder the development of efficient, future-proof cultural diagnostics capable of providing organizations with timely and actionable insights into the state of their human capital.

[1] https://www.de-hofnar.nl/

An automated pipeline for interpreting qualitative data offers a promising methodological alternative. Natural Language Processing (NLP), a subfield of artificial intelligence concerned with enabling machines to understand, interpret, and generate human language, provides computational methods capable of extracting structure from large volumes of unstructured text. As stated before, while skilled analysts are capable of identifying themes, sentiment, contradictions, and other cultural signals, doing so manually is highly time-consuming and becomes increasingly impractical as datasets grow. NLP techniques address this bottleneck by transforming rich interview transcripts into quantifiable indicators that can be processed consistently and efficiently. This enables organisations to analyse qualitative data at a scale and speed that manual approaches cannot feasibly match, while still preserving the depth of insight that open-ended interviews provide. Applied to OC research, NLP therefore lays the foundation for a scalable, reproducible, and data-driven assessment framework that supports De Hofnar's ambition to deliver timely and actionable cultural insights.

The goal of this thesis is to investigate how OC can be quantified using NLP techniques applied to employee text data. To this end, an automated pipeline is developed and evaluated as a means to explore how qualitative cultural signals can be translated into actionable and interpretable insights. That is uncovering what shared meanings, values, and practices exist, and to what extent they represent cultural strengths or sources of friction. The research proceeds in three main steps. First, we identify key dimensions of OC through a literature review. Second, we research, apply and evaluate NLP techniques to evaluate the companies culture. In this step we compare traditional language models with Large Language models (LLMs). Finally, we aggregate the results into a single Cultural Dashboard. The dashboard is designed to be both explainable and trackable over time, thereby providing organizations with actionable insights into cultural strengths and potential friction points.

Research Question

> *How can NLP be used to quantify organizational culture using text data and translate it into actionable insights?*

Sub-questions

1. What cultural model is best suited to measure OC?
2. Which NLP methods are most effective for detecting signals of culture in employee interviews?
3. How do traditional language models and LLMs compare in terms of performance for OC analysis?
4. How can model outputs be aggregated into interpretable and actionable organizational insights?

Chapter 2 Related Work

2.1 Organizational Culture

In this section, we outline the research on OC. We discuss the definitions, how it is measured and explore different cultural models.

2.1.1 Definition

OC is a concept with multiple, often overlapping definitions in the literature. A widely cited and influential perspective is provided by Schein, who defines OC as:

> "*Organizational culture is the pattern of basic assumptions that a given group has invented, discovered, or developed in learning to cope with its problems of external adaptation and internal integration, and that has worked well enough to be considered valid, and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems.*" (Schein, 1984)

Building on this foundation, many scholars describe OC as the set of shared characteristics of individuals within an organization (Tadesse Bogale and Debela, 2024). Van Der Post et al. (1997) highlight this collective dimension by portraying culture as the "personality" of the organization: a hidden but unifying system of shared meanings, values, and practices that provides direction and identity. Similarly, Akhavan et al. (2014) emphasize that OC significantly influences interpersonal interactions, behaviours, and communication in day-to-day work. This work is also confirmed by a review from Bellot (2011), who describes the academic consensus as 1) OC exists, 2) cultures are inherently fuzzy, 3) culture is socially constructed, and 4) each organization's culture is relatively unique. Together, these perspectives underscore that OC is both deeply embedded and socially constructed, shaping how members interpret their environment and coordinate collective action.

2.1.2 Cultural Models

Most relevant to this thesis is the OC literature on models culture. A number of models have been proposed to capture the dimensions of OC, with no consensus on what the best dimensions are. In this section, we detail the most cited and influential models, and discuss, wherever possible, how the models are constructed/validated.

CVF The Competing Values Framework (CVF), developed by Quinn and Rohrbaugh (1981), explains OC through three core dimensions that contrast (1) an internal with an external orientation, (2) stability and control with flexibility and change, and (3) means-oriented with ends-oriented criteria. The framework was derived through a multistage procedure in which Quinn and Rohrbaugh reduced thirty effectiveness criteria to sixteen with expert input and compared them pairwise for conceptual similarity. They then applied the INDSCAL multidimensional scaling algorithm, which revealed the three underlying value dimensions. The solution showed strong expert agreement, with an overall correlation of $r \approx 0.72$ and individual correlations between $r \approx 0.64$ and $r \approx 0.87$, supporting the validity of the CVF structure.

Excellent Companies Peters and Waterman (1982) popularised the idea that corporate culture can be a source of competitive advantage. In their work they identified eight "excellence attributes" through a qualitative study of 43 high-performing U.S. companies. These attributes include (1) a bias for action, (2) customer closeness, (3) autonomy and entrepreneurship, (4) productivity through people, (5) hands-on and value-driven leadership, (6)focus on core business, (7) simple structure with lean staffing, and (8) simultaneous loose–tight properties. Peters and Waterman constructed this framework using a grounded, comparative case approach: they selected firms through expert nominations and performance screening, conducted extensive interviews and archival analysis, and derived the eight attributes through cross-case pattern recognition rather than statistical modelling. Their work became highly influential in management practice, although its academic validity has been widely debated.

OCS The Organizational Culture Survey (OCS), developed by Glaser et al. (1987), is an early quantitative instrument designed to translate the qualitative notion of OC into empirically verifiable dimensions. They identified six core factors: (1) teamwork and conflict, (2) climate and morale, (3) information flow, (4) involvement, (5) supervision, and (6) meetings. The model was constructed by first administering a 62-item version covering five categories to employees in a wood-products company ($N = 267$), then refining it to 31 items and re-testing it in a manufacturing firm ($N = 138$) and a government agency ($N = 195$). A principal components analysis on the latter sample produced the six factors, and validation analyses demonstrated strong internal consistency ($\alpha = 0.72$–0.88), high test–retest reliability ($p < .001$), and significant satisfaction/dissatisfaction ratios for most dimensions, confirming the OCS as a reliable and valid measure across organizational contexts.

OCI The Organizational Culture Inventory (OCI), developed by Cooke and Lafferty (1983), extends the Life Styles Inventory (LSI) by reframing its 12 individual thinking and behavioural styles as shared organizational norms, resulting in 12 cultural dimensions: (1) Humanistic-Encouraging, (2) Affiliative, (3) Approval, (4) Conventional, (5) Dependent, (6) Avoidance, (7) Oppositional, (8) Power, (9) Competitive, (10) Perfectionistic, (11) Achievement, and (12) Self-Actualizing. These dimensions form a circumplex defined by people- versus task-focus and satisfaction versus security needs, yielding three cultural types: Constructive, Passive/Defensive, and Aggressive/Defensive. Validation by Cooke and Szumal (1993) using approximately $N \approx 4{,}890$ respondents showed strong reliability ($\alpha = .65$–$.95$; $\eta^2 = .12$–$.40$), stable test–retest correlations over 21–24 months, and a three-factor structure explaining 65–73% of variance, with criterion analyses linking Constructive cultures to satisfaction and innovation and Defensive cultures to stress and turnover.

Hofstede Hofstede et al. (1990) identified six dichotomous dimensions of OC through a three-stage study of 20 units across 10 Danish and Dutch organizations: (1) process- vs. results-orientation, (2) employee- vs. job-orientation, (3) parochial vs. professional, (4) open vs. closed system, (5) loose vs. tight control, and (6) normative vs. pragmatic. The model was developed using 180 exploratory interviews, a survey of N=1,295 employees, and an ecological principal component analysis of practice items that explained 73% of the variance (rising to 86% with three core items per dimension). Validation through additional unit-level interviews showed meaningful correlations with organizational characteristics, such as specialization, formalization, and hierarchy, confirming the robustness of the six-dimension structure.

OCP O'Reilly et al. (1991) developed the Organizational Culture Profile (OCP) as a quantitative measure of person–organization fit based on seven value dimensions: (1) innovation, (2) stability, (3) respect for people, (4) outcome orientation, (5) attention to detail, (6) team orientation, and (7) aggressiveness. The framework was built by refining an initial pool of 110 value statements to 54 items, using Q-sort procedures to create organizational and individual value profiles, and applying factor analyses across five samples (N=1,395) to derive the final dimensions. Reliability was high ($\alpha \approx$ .84-.90), test–retest stability remained strong over twelve months ($r \approx .73$), and predictive validity was confirmed through correlations with commitment, job satisfaction, intent to leave, and two-year tenure outcomes.

Denison Denison and Mishra (1995) developed a performance-linked model of OC based on five qualitative case studies, identifying four core dimensions: (1) involvement, (2) consistency, (3) adaptability, and (4) mission. This framework was later formalized into the Denison Organizational Culture Survey (DOCS) and validated across large-scale quantitative studies, with confirmatory factor analyses supporting its hierarchical structure (CFI = .98; RMSEA = .04) and showing strong reliability across twelve indexes ($\alpha = .70$–$.85$) and high interrater agreement ($r_{wg} = .85$–$.89$) (Denison et al., 2014). In a sample of 160 organizations ($N = 35{,}474$), the model demonstrated excellent reliability (ICC(2) = .93–.96) and predictive validity, with culture–performance correlations ranging from $r = .44$ to $r = .68$, where involvement and adaptability predicted growth and consistency and mission predicted profitability.

Van der Post Van Der Post et al. (1997) synthesized prior culture research into a coherent measurement framework by reviewing 114 reported cultural dimensions and grouping them through expert card-sorting into 15 constructs: (1) conflict resolution, (2) culture management, (3) customer orientation, (4) disposition toward change, (5) employee participation, (6) goal clarity, (7) human resource orientation, (8) identification with the organization, (9) locus of authority, (10) management style, (11) organizational focus, (12) organizational integration, (13) performance orientation, (14) reward orientation, and (15) task structure. A 169-item questionnaire was developed and refined through pilot testing and item–total correlation analysis, resulting in a 97-item instrument with strong reliabilities ($\alpha = 0.788$–0.932). Construct validity was confirmed through factor analysis using principal factor extraction with varimax rotation, which produced 15 factors with eigenvalues greater than 1.0 and loadings between 0.39 and 0.84, aligning closely with the theoretical structure.

Culture 500 The Culture 500, developed by MIT Sloan Management Review and Glassdoor, is a large-scale benchmarking project that measures OC using employee-generated text data rather than survey-based constructs (Sull et al., 2019). It identifies nine cultural values, the "Big 9": (1) agility, (2) collaboration, (3) customer orientation, (4) diversity, (5) execution, (6) innovation, (7) integrity, (8) performance, and (9) respect. The framework was built using more than 1.2 million Glassdoor reviews from over 500 major U.S. firms, combining human annotation and NLP to extract more than 90 culture-related topics that were aggregated into the Big 9 values. For each company, researchers computed the frequency and sentiment of value-related mentions. Although the approach does not involve traditional psychometric validation, it has high face validity, aligns strongly with firms' stated values, and shows meaningful associations with employee ratings, reputation, and financial performance, supporting its credibility as a data-driven measure of OC.

2.1.3 How Culture Is Measured

Like there is no consensus of the definition or conceptualisation of OC, there is also no consensus in the literature on how to measure culture or aspects thereof. There are three approaches to exploring OC (Jung et al., 2009). The approaches can broadly be divided into dimensional, emergent, and typological methods. Dimensional approaches assess the degree to which predefined cultural traits are present within an organization. In contrast, emergent approaches take a bottom-up perspective, allowing cultural dimensions to arise from participants' own descriptions of their organization before being clustered and analyzed. Lastly, typological approaches go a step further by categorizing organizations into overarching cultural types or archetypes, such as the Clan, Adhocracy, Market, or Hierarchy cultures, an idea built from the Competing Values Framework (Quinn and Rohrbaugh, 1981).

Druckman et al. (1997) describe 3 methods for data gathering. The first is a holistic approach, where the researcher becomes involved in the organization of interest, such as the interview-based approach used by De Hofnar. Secondly the metaphorical or language approaches, in which the researcher uses language patterns in documents, reports, stories, and conversations in order to uncover cultural patterns. Finally there are quantitative approaches, in which the investigator uses questionnaires or interviews to measure particular dimensions of cultures, with numerical scales or codes. Jung et al. (2009) identified 3 common methods for quantitative approaches. The most common techniques include Likert scales, Q-methodology, and ipsative measures. Likert scales present respondents with predefined statements and ask them to indicate their level of agreement on a graded scale, offering a straightforward way to quantify attitudes. Q-methodology instead requires participants to sort a set of value statements along a continuum from least to most characteristic, providing a more nuanced measure of subjective viewpoints. Ipsative measures, often used in the Competing Values Framework, ask participants to allocate a fixed number of points among several statements, forcing trade-offs that reveal dominant cultural preferences.

The question remains as to which method is most appropriate for measuring OC. Jung et al. (2009). argue that the choice of method should be guided by the specific aims of the study. Druckman et al. (1997) add further considerations to this debate. The first concerns the dimensional approach: a common critique is that survey instruments predetermine cultural dimensions rather than capturing those that exist a priori within the organization. As a result, questionnaires may miss or distort the culture's actual structure. One way to mitigate this issue is to construct survey items based on ethnographic observation, ensuring that they reflect dimensions grounded in qualitative insight. The second consideration draws on the proverb "fish discover water last," emphasizing that individuals are often unaware of their own culture until it is explicitly described to them. Consequently, some approaches ask respondents to indicate the extent to which written scenarios or descriptions reflect their organization's culture to mitigate this problem.

2.1.4 Synthesis of Culture Research

In this section, we explain the reasoning behind our chosen cultural model. We also highlight additional considerations from the literature that guides our research design and the development of our analysis pipeline.

Table 1: Summary of Cultural Models

Model	Dimensions	Qualitative Grounding	Exploratory Factor Analysis	Reliability	Predictive Validity	External Validity
CVF	3 Scales	✓	✓	✓	✓	✓
Excellent Companies	8 Bins	✓	-	-	-	-
OCS	6 Bins	✓	✓	✓	✓	✓
OCI	12 Bins	✓	✓	✓	✓	✓
Hofstede	6 Scales	✓	✓	✓	✓	✓
OCP	7 Bins	✓	✓	✓	✓	✓
Denison	4 Bins	✓	✓	✓	✓	✓
Van der Post	15 Bins	✓	✓	✓	✓	✓
Culture 500	9 Values	✓	-	-	✓	✓

Model Choice Our NLP pipeline aims to automatically analyse OC. To ensure conceptual validity and interpretability, this analysis must be grounded in an established and empirically supported culture model. As shown in Table 1, most widely used models, with the exception of Culture 500 and the Excellent Companies framework, demonstrate robust validation across multiple criteria. These include qualitative grounding through literature review or case studies, exploratory factor analysis to verify dimensional structure, reliability testing, and both criterion and external validity assessments. Because several, well validated models were available, we consulted an expert from De Hofnar, who conducted all interviews, to identify the most suitable model for this dataset and application.

We selected the OCS by Glaser et al. (1987) for three primary reasons. First, the OCS offers concrete and intuitively interpretable dimensions. Its six categories are conceptually clear and can be easily attributed to concrete behaviours and practices within the organization, which supports both manual labelling and downstream interpretation. In contrast, models such as the Organizational Culture Inventory (OCI) (Cooke and Lafferty, 1983), with dimensions like "humanistic-encouraging" or "oppositional", rely on more abstract constructs that are harder to consistently annotate and less straightforward to communicate in an applied setting.

Second, the OCS strikes a desirable balance in model scope. With six dimensions, it provides sufficient breadth to capture specific aspects of culture without becoming overly complex. Models such as the Competing Values Framework (CVF) (Quinn and Rohrbaugh, 1981) or Denison and Mishra's framework (Denison and Mishra, 1995), with only 4 dimensions, were considered less complete for the purposes of De Hofnar's work while larger models, such as the fifteen-dimensional

instrument of Van Der Post et al. (1997), risk including dimensions that may not surface in every interview corpus. A concise model increases the likelihood that all dimensions are adequately represented across companies.

Finally, the OCS consists of categorical bins rather than bipolar continua. This structural property aligns well with the requirements of the pipeline: individual text segments can be assigned to one or more discrete categories without needing to position them along latent scales such as those used by Hofstede et al. (1990) or the CVF (for example, internal versus external focus, or tight versus loose control). A categorical structure simplifies annotation, model training, and interpretation of the output dashboard.

Other Considerations In this research, OC is conceptualized in line with the definition proposed by Van Der Post et al. (1997), namely as the identification of shared meanings, values, and practices within the workplace. Although the OCS model has been extensively validated, we also acknowledge the concerns raised by Druckman et al. (1997), who argue that a fixed set of dimensions may be too restrictive to fully capture the cultural reality of a specific organization. To address this limitation, the dimensional approach is complemented with topic modeling, as discussed in Section 2.2, which allows for the identification of company-specific cultural elements. To further support interpretability and explainability, representative example quotes are also presented.

2.2 Natural Language Processing

In this section, we describe several use cases for NLP, both in general domains and in the context of OC. We also outline the main NLP techniques used in this thesis, explain what they do, which models are available, and summarize relevant best practices.

2.2.1 General Use Cases

In this section, we explore how NLP is applied across different sectors to understand what techniques are available, how they are commonly used, and which approaches may be relevant or adaptable to our use case. By examining applications in domains such as healthcare, psychology, education, and customer service we aim to identify methodological patterns and extract practical inspiration for designing an NLP-driven pipeline for OC analysis.

Healthcare The first approach is the study by Cammel et al. (2020) show how NLP can transform open-ended patient survey responses into actionable insights. Using Non-Negative Matrix Factorization (NMF) topic modelling on data from two hospitals, the authors identified themes and refined them with n-grams to add contextual specificity. Each response received a sentiment score, and topics were prioritised through an impact metric that combined sentiment and frequency. The model transferred successfully to a second hospital, demonstrating cross-institution applicability. Strengths include reduced human bias, clear prioritization through the impact score, and demonstrated transferability. Limitations involve sentiment bias from survey phrasing, single-topic assignment per response, and challenges in capturing nuanced findings.

Psychology Guo et al. (2024) use Doc2Vec (Mikolov et al., 2013) embeddings to predict Big Five personality traits from open-ended interview responses. Dense text vectors were fed into ridge regression models, achieving an average correlation of $r \approx 0.28$, comparable to or better than lexical and n-gram baselines. Strengths include modelling nuanced language in small datasets and demonstrating how text can complement traditional assessments. Limitations relate to moderate predictive accuracy, reliance on self-reported validation data, and restricted generalizability due to controlled prompts.

Education Parker et al. (2025) demonstrate how LLMs such as GPT-3.5 and GPT-4 can analyse open-ended educational survey comments using a zero-shot, multi-step workflow. Across 2,500 student responses, LLMs performed classification, extraction, thematic analysis, and sentiment scoring with near-human accuracy while drastically reducing analysis time. Chain-of-thought reasoning improved transparency and error inspection. Strengths include accessibility, flexibility, and interpretability without domain-specific training data. Limitations involve strong prompt dependence and uncertain generalizability beyond English educational contexts.

Customer satisfaction Aldunate et al. (2022) propose a deep learning framework for identifying customer-satisfaction drivers in 25,943 survey responses from 39 service companies. They reformulate the task as multi-label classification using eleven predefined drivers and fine-tune Bidirectional Encoder Representations from Transformers (BERT)(Devlin et al., 2018) within a CRISP-DM–based workflow. A hybrid BERT + term frequency-inverse document frequency (TF-IDF) model achieved a weighted F1-score of 0.72, outperforming SVMs and shallow neural networks. Strengths include transferability, strong performance across thirteen sectors, and actionable managerial insights. Limitations concern reliance on labelled data, uneven driver representation, and reduced performance in smaller subsets.

Across the literature, including the four studies discussed above, most text-based analytical approaches can be grouped into four dominant methodological families: embedding, topic modelling, text classification, and sentiment analysis. These constitute the core of contemporary text-mining research, providing complementary ways to uncover, categorize, and evaluate meaning in large collections of unstructured data. As Hassani et al. (2020) note, these four methods underpin the majority of applications across domains ranging from political discourse and marketing analytics to social media and academic review mining. Other techniques, such as argument and knowledge extraction, keyword identification, dimensionality reduction, and network-based modelling, play a more specialized but supporting role, often enhancing the interpretability or structure of the main analytical outputs rather than replacing them altogether.

2.2.2 Organizational Culture Use Cases

In this section, we focus on NLP approaches that directly address the assessment of OC. Unlike the previous section, which explored NLP applications across diverse domains, the studies reviewed here examine how culture can be inferred, classified, or quantified from natural language data such as employee reviews, organizational documents, and interview transcripts.

Schachner et al. (Schachner et al., 2024) introduce the Dictionary of Organizational Culture and Practices (DOCP), a theory-driven tool for measuring OC directly from text. Based on Hofstede et al. (1990) six cultural dimensions, the dictionary maps over 2,800 validated n-grams to dimension poles using a corpus of 26 million words and expert-based refinement. Applied through tools such as LIWC or Quanteda, the DOCP computes dimension-specific word frequencies to generate culture profiles. Validation showed strong criterion validity with Li et al. (2018) and high congruence with survey-based Hofstede scores ($> .9$), demonstrating that theoretically grounded culture dimensions can be reliably extracted from natural language at scale.

Ginting et al. (2023) develop a multi-label text classification approach to measure corporate culture as an alternative to self-report instruments. Using 14,042 expert-verified sentences from organizational documents, interviews, and testimonials, the authors operationalize 21 culture sub-dimensions across five categories. Twelve models were evaluated, with the combination of Label Powerset–Linear SVC model performing the best, with precision 0.89, recall 0.81, and F1 0.85. A pilot test confirmed organizational applicability, although broader validation is recommended by the authors. The study provides a systematic procedure for quantifying culture via machine-based text analysis.

Koch and Pasch (2023) present CultureBERT, a transformer-based approach to measuring culture in employee reviews using fine-tuned LLMs. A dataset of 2,000 manually coded reviews was annotated according to the CVF (Quinn and Rohrbaugh, 1981). Fine-tuned BERT(Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models were compared with dictionary-based and TF-IDF classifiers. RoBERTa-large achieved the highest accuracy, outperforming dictionary methods by 17–30 percentage points and TF-IDF models by 3–15 points. The results highlight transformers' contextual strengths, particularly in handling polysemy, negation, and syntax, and demonstrate their superior validity for automated culture measurement.

Gupta et al. (2022) use Embedded Topic Modeling (ETM) to extract cultural elements from 24,965 Glassdoor reviews in the operations management domain. Compared with Latent Dirichlet Allocation (LDA) (Blei et al., 2001), ETM achieved better fit through lower perplexity and higher topic coherence. The authors tested bag-of-words, n-gram, and recurrent neural network inputs to assess robustness. Nineteen cultural elements emerged, including leadership, communication, compensation, and work environment, with both shared and organization-specific themes. Mapping these themes onto established frameworks showed that text-based methods surface cultural aspects often missed by traditional surveys, positioning employee-generated text as a scalable resource for studying culture dynamically.

Together, these studies show that culture is typically inferred either by classifying text into predefined cultural dimensions or by using topic modeling to uncover prevalent themes. This aligns with the patterns identified in Section 2.2.1. The Culture 500 project, described in Section 2.1.2, offers a further example of this approach.

2.2.3 Text Classification

Text classification is a core NLP task concerned with assigning predefined labels to textual input. It underpins a wide range of applications, including sentiment analysis, topic labelling, intent detection, and document categorisation. In recent years, the task has been dominated by pretrained transformer-based language model families, which consistently outperform traditional machine learning approaches.

Survey work by Gasparetto et al. (2022) demonstrates that encoder-based transformer model families achieve state-of-the-art performance across diverse text classification benchmarks. These models benefit from contextualized token representations and typically require minimal preprocessing; in particular, stopword removal is unnecessary and often undesirable, as transformer architectures are trained to exploit full linguistic structure. In contrast, smaller or non-contextual model families may still benefit from more aggressive preprocessing.

Chae and Davidson (2026) compare ten language models across zero-shot prompting, few-shot prompting, fine-tuning, and instruction-tuning for text classification. Using stance detection as a case study, they find that the largest LLMs often achieve the highest performance with minimal examples, while fine-tuned smaller traditional language models remain competitive due to their favourable accuracy-cost trade-off. They further conclude that instruction-tuning expands the applicability of LLMs to more complex classification tasks, highlighting the importance of matching model strategy to task requirements.

Beyond fine-tuning, large language model families enable text classification through prompt-based inference. Sun et al. (2023) propose Clue-and-Reasoning Prompting (CARP), in which prompts are structured into three components: a task description, a small set of annotated demonstrations, and the input instance to be classified. Their results show that, with careful prompt design, LLMs can match or slightly outperform fine-tuned encoder-based transformer families, particularly in low-data scenarios.

From a system design perspective, Minaee et al. (2022) provide a practical framework for model selection in text classification. They recommend starting from a pre-trained language model family, optionally continuing pre-training on domain-specific corpora, adapting the architecture to the task, fine-tuning on labelled data, and finally compressing the model to meet latency constraints. Across their experiments, encoder-based transformer families emerge as consistently strong choices, reinforcing their status as reliable baselines for classification tasks.

2.2.4 Topic Modelling

Topic modelling aims to uncover latent semantic themes in a corpus without relying on labelled data. A topic model represents each topic as a structured semantic entity, such as a distribution over words or a region in embedding space, and associates each document[2] with one or more topics. Classical formulations treat documents as mixtures of topics, while more recent approaches cluster

[2]In the context of text mining, a document denotes a single unit of analysis, which may vary in length from a short sentence or message to an extended multi-page text.

documents in a semantic embedding space. A comprehensive overview by Abdelrazek et al. (2022) categorizes topic modeling approaches into algebraic, fuzzy, Bayesian probabilistic, and neural methods.

Neural and embedding-based approaches have become increasingly prominent. Wang et al. (2023) compare prompt-based topic modelling using large language model families with embedding-based clustering frameworks. While LLM-based methods perform competitively, embedding-based clustering approaches achieve slightly better overall performance. To enable comparison, they use class-based TF–IDF (c-TF-IDF) to extract representative topic terms, as prompt-based topic models do not inherently provide explicit word distributions.

More recently, Pham et al. (2024) introduce a generative large language model–based topic generation framework. Their results show substantial improvements over both classical probabilistic topic models and embedding-based clustering approaches, particularly in topic coherence and interpretability. However, systematic comparisons across multiple topic modelling paradigms remain scarce in the literature. LDA (Blei et al., 2001) and BERTopic (Grootendorst, 2022) continue to serve as widely adopted models in empirical research.

Despite methodological advances, evaluating topic models remains challenging. Wu et al. (2024) show that commonly used metrics such as topic coherence and topic diversity often correlate poorly with human judgment. They argue for semantic-aware diversity, emphasizing that word meaning is context-dependent; for example, the interpretation of a term such as "apple" shifts depending on surrounding concepts. This observation is particularly relevant in domains where abstract or polysemous concepts recur, such as OC or workplace dynamics.

2.2.5 Sentiment Analysis

Sentiment analysis focuses on identifying the sentiment expressed in text, typically modelling sentiment as positive, negative, or neutral, or along a more fine-grained intensity scale. More advanced formulations, such as aspect-based sentiment analysis, further decompose sentiment by target or topic, allowing multiple attitudes to be expressed within a single document.

Bashiri and Naderi (2024) conduct a broad comparison of sentiment models and show that encoder–decoder transformer families offer strong and stable performance across diverse sentiment tasks. Certain autoregressive transformer variants excel in specialized settings such as irony detection and product-focused sentiment, whereas distilled or compressed encoder-based transformer families trade accuracy for computational efficiency. Discriminative pretraining variants stand out for their ability to reduce false positives and accurately identify negative sentiment, making them particularly suitable for high-precision applications.

Recent work increasingly contrasts encoder–decoder transformer families and large language model families with traditional encoder-based transformer classifiers. Zhang et al. (2024) compare encoder–decoder transformers and LLMs across zero-shot and few-shot sentiment analysis tasks. Their findings indicate that LLMs perform well in zero-shot settings for simple sentiment distinctions

but struggle with more complex or nuanced sentiment tasks. In few-shot settings, however, LLMs consistently outperform smaller pretrained model families, highlighting their adaptability when limited labeled data is available.

A large-scale empirical study by Shafikuzzaman et al. (2024) evaluates sentiment analysis in software engineering text using three modeling paradigms: traditionally fine-tuned pretrained transformer families, few-shot sentence-embedding–based models, and zero-shot large language model families. Across six datasets, they show that performance is strongly influenced by dataset size and class imbalance. Encoder-based transformer families achieve the highest macro F1-scores on large, well-balanced datasets, while lightweight embedding-based models perform best on smaller datasets with limited labeled data. LLMs are the strongest zero-shot models but generally underperform fine-tuned and few-shot approaches when supervised data is available.

Complementing these findings, Bashiri and Naderi (2024) similarly report that encoder–decoder transformer families offer stable cross-task performance, while other transformer variants trade off efficiency and precision depending on the application context.

2.3 Summary of Related Work

In summary, prior research demonstrates that OC can be operationalised through theoretically grounded dimensions and extracted from text using modern NLP techniques. Building on this foundation, our pipeline integrates three complementary cultural signals. First, we perform multi-label classification within the OCS (Glaser et al., 1987) framework, directly comparing fine-tuned BERT-based (Devlin et al., 2018) models with instruction-tuned LLMs to assess their ability to detect predefined cultural dimensions. Second, we apply topic modelling to uncover emergent, organization-specific themes, contrasting LDA (Blei et al., 2001) and BERTopic (Grootendorst, 2022) with LLM-based topic generation. Third, we incorporate sentiment analysis as a contextual layer, comparing a pre-trained BERT (Devlin et al., 2018) sentiment model with an LLM to evaluate affective tone. Together, these three components combine theory-driven categorisation, data-driven theme discovery, and evaluative sentiment signals, forming a structured and empirically grounded basis for quantifying OC in an automated pipeline.

Chapter 3 Data

This section outlines the data used throughout the experiments. We begin with the primary dataset, consisting of employee interview transcripts, and describe its key characteristics across companies, including structure, length, and word variability, which directly inform subsequent modelling and preprocessing decisions. We then introduce the annotated datasets used for the classification and sentiment analysis tasks, detailing the annotation procedures, labelling schemes, and quality considerations. Finally, we examine the resulting label distributions to provide insight into class imbalance and potential sources of bias in the evaluation.

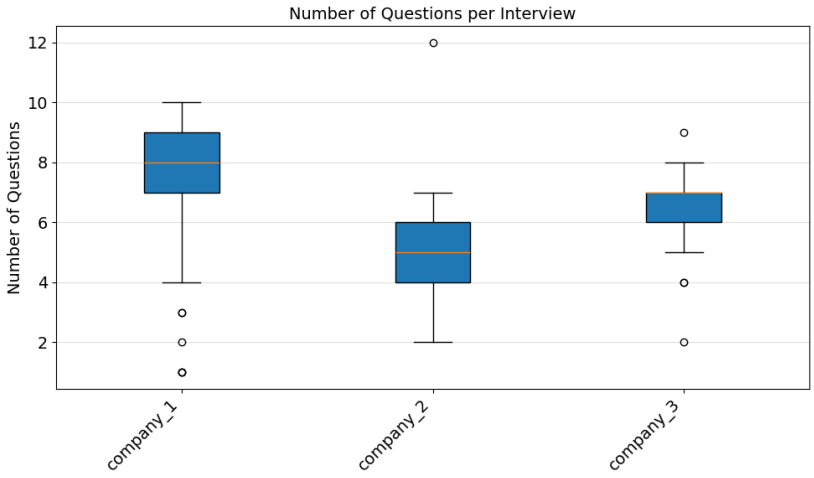


Figure 1: Box plot of the number of questions per interview for each company

3.1 Interview Data

De Hofnar[3] is a startup that aims to surface unspoken organizational dynamics and present them to management in a constructive and unbiased manner. To this end, De Hofnar conducts in-depth interviews with employees within organizations. These interviews are transcribed manually by the interviewer in real time, implying that the resulting transcripts are not fully verbatim. As a consequence, the data may contain omissions, paraphrasing, and other forms of interpretation reflecting the interviewer's judgment regarding what to record and how to formulate it. All interviews are conducted by the same interviewer and are held in Dutch, ensuring consistency in interview style and transcription practices across interviewees and organizations. However, the questions asked differ based on the specific needs and challenges of each organisation.

Company 1 In this organization[4], De Hofnar conducted a total of 127 interviews, which yielded a total of 1,000 question-answer pairs. All interviews were conducted in the context of a recent organizational transition. Compared to the other organizations in the dataset, interviews in Company 1 contained the highest number of questions, with an average of 7.3 main questions per interview, as shown in Figure 1 and Table 2. The corresponding answers have an median length of 55 words (see Table 4). To summarize, the dataset comprises of 127 interviews, with 1,000 question-answer pairs, and 72,983 words (5,103 unique words) for this company.

Table 2: Distributional statistics for number of questions per interview

Company	n	Min	Q1	Median	Q3	Max	Whisker$_L$	Whisker$_U$	Outl. (%)
Company 1	127	1.00	7.00	8.00	9.00	10.00	4.00	10.00	4.72
Company 2	15	2.00	4.00	5.00	6.00	12.00	2.00	7.00	6.67
Company 3	42	2.00	6.00	7.00	7.00	9.00	5.00	8.00	9.52

[3] https://www.de-hofnar.nl/

[4] For reasons of confidentiality, participating companies are not mentioned by name.

Company 2 In Company 2, a total of 15 interviews were conducted with employees who had recently joined the organization, which yielded a total of 80 question-answer pairs. These interviews focused on the company's ongoing efforts to improve its onboarding procedures. On average, fewer questions were asked per interview than in Company 1, with a mean of 5.3 main questions as can be seen in Table 2. However, answers in Company 2 are substantially longer. As shown in Figure 3 and Table 4, answers contain a mean of 136 words. To summarise, the dataset comprises of 15 interviews, with 80 question-answer pairs, and 13,272 words (2,314 unique words) for this company.

Company 3 In Company 3, a total of 42 interviews were conducted, yielding 275 question–answer pairs. On average, each interview consisted of 6.55 main questions. Compared to Company 1 and 2, responses in Company 3 are longer on average, containing approximately 202.55 words per response. However, the mean answer length is lower than Company 2, as can be seen in Figure 3 and Table 4. To summarize, the dataset comprises of 42 interviews, with 275 question-answer pairs, and 55,189 words (4,483 unique words) for this company.

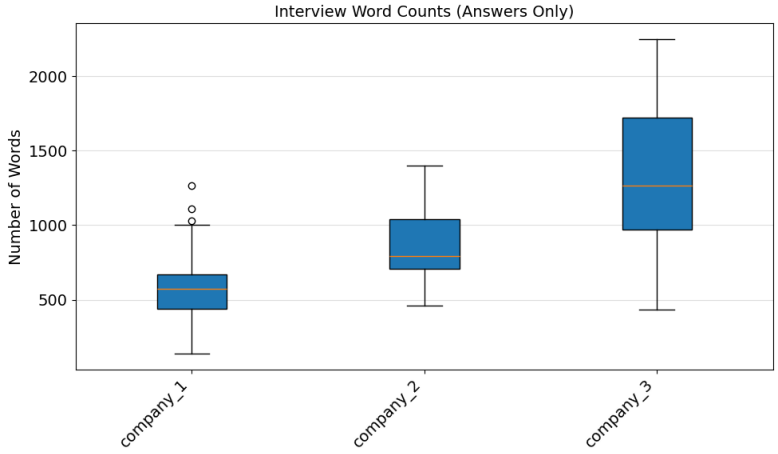


Figure 2: Number of words per interview per company. Questions not included.

Table 3: Distributional statistics for number of words per interview per company (questions excluded).

Company	n	Min	Q1	Median	Q3	Max	Whisker$_L$	Whisker$_U$	Outl. (%)
Company 1	127	138.00	440.50	572.00	671.50	1264.00	138.00	1004.00	2.36
Company 2	15	461.00	709.00	795.00	1039.50	1398.00	461.00	1398.00	0.00
Company 3	42	432.00	968.50	1267.50	1721.25	2248.00	432.00	2248.00	0.00

In total, this results in a dataset of 1,383 documents. The dataset is unevenly distributed across organizations, with Company 1 contributing the majority of interviews. Across all documents, the corpus comprises 143,181 words and a vocabulary of 8,086 unique words. Document lengths vary substantially, with a mean of 103.5 words per document and a median of 65 words, reflecting a right-skewed distribution of textual length. No empty documents are present in the dataset.

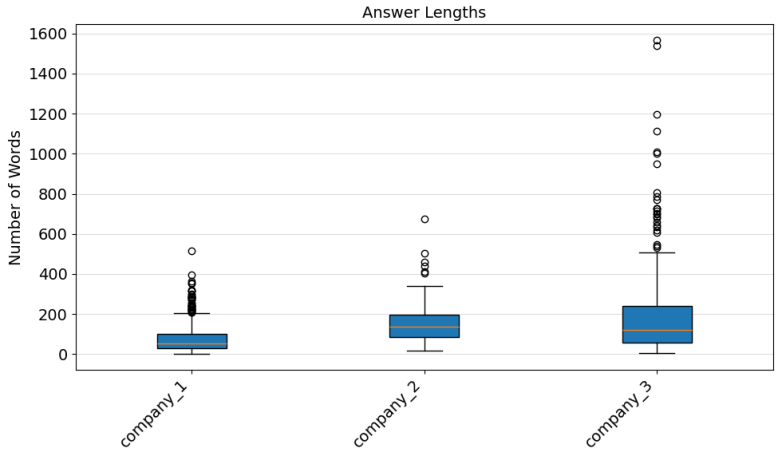


Figure 3: Number of words per answer per company.

Table 4: Distributional statistics for number of words per answer per company.

Company	n	Min	Q1	Median	Q3	Max	Whisker$_L$	Whisker$_U$	Outl. (%)
Company 1	1000	1.00	28.00	55.00	99.00	516.00	1.00	205.00	4.00
Company 2	80	18.00	84.25	136.00	196.25	674.00	18.00	341.00	7.50
Company 3	275	5.00	58.00	121.00	239.50	1567.00	5.00	506.00	9.45

From a modelling perspective, an important consideration concerns the length of the textual data in relation to model constraints. As shown in Figure 2, complete interviews are substantially longer than the maximum context window supported by BERT-based models. Interview transcripts have mean lengths ranging from 569 words in Company 1 to 1326 words in Company 3, with considerable dispersion as reflected by interquartile ranges between 231 and 753 words. When converted to tokens, these interview lengths typically correspond to approximately 760–1,770 tokens[5] , exceeding the 512-token limit of BERT (Devlin et al., 2018) for the majority of interviews. Consequently, modelling full interviews would require truncation, risking systematic information loss, or the application of sliding-window approaches, which introduce additional computational complexity and complicate interpretation. This constraint is largely mitigated when analysing individual answers instead of complete interviews. As illustrated in Figure 3 and Table 4, answer-level responses are substantially shorter across all companies, with mean lengths of 72, 161, and 203 words for Companies 1, 2, and 3 respectively. Although Company 3 exhibits greater variability, with an interquartile range of 182 words and a higher proportion of long responses, the prevalence of extreme outliers remains limited: fewer than 10% of answers exceed the standard boxplot outlier threshold in any company. As a result, the vast majority of answers fall comfortably within the 512-token limit, making answer-level analysis feasible without truncation for most observations. These characteristics support the methodological decision to operate at the answer level, balancing contextual completeness with compatibility with transformer-based language models.

[5]Estimated using https://www.gptcostcalculator.com/open-ai-token-calculator.

3.2 Classification Annotation Data

In order to train, and evaluate classification models we use manual annotation of the dataset according to the OCS framework. All annotations are performed by a single annotator (the author). This was a deliberate design decision for two reasons. First, organizational constraints limited the availability of additional trained annotators. Second, the interview data contain highly sensitive information, making outsourcing annotation infeasible. The detailed annotation guidelines are provided in Appendix A. These guidelines were derived from the original OCS working definitions and associated questionnaire items, displayed in Appendix B.

The full dataset, 1383 documents, is annotated. The annotation process is designed as a multi-label classification task, allowing each document to be assigned to one or more OC categories. The majority of documents (79.7%) are associated with a single label, while a substantial minority (19.2%) contain two labels, and a small fraction (1.1%) exhibit three labels. This distribution indicates that, although most employee statements focus on a single cultural dimension, a non-negligible portion reflects overlapping themes, justifying the multi-label formulation. The class distribution in Table 5 further highlights imbalance across categories, with *Teamwork and Conflict* (43.0%) and *Climate and Morale* (28.3%) dominating the dataset, while categories such as *Meetings* (3.5%) and *Involvement* (5.6%) are underrepresented. This imbalance has direct implications for model training and evaluation, necessitating techniques that account for skewed label frequencies. Finally, the co-occurrence patterns in Figure 4 reveal systematic relationships between labels, particularly between *Teamwork and Conflict* and *Climate and Morale*, as well as *Information Flow*. These dependencies suggest that certain cultural dimensions are not independent but instead tend to manifest jointly within employee narratives. Consequently, the annotation schema captures both the primary classification structure and the latent interdependencies between categories, reinforcing the need for models capable of handling correlated multi-label outputs.

Table 5: Class distribution indicating the percentage of documents that contain each label.

Label	**Documents (%)**
Teamwork and Conflict	43.0
Climate and Morale	28.3
Information Flow	20.7
Involvement	5.6
Supervision	6.8
Meetings	3.5
Other	13.5

Note: Percentages do not sum to 100% due to the multi-label nature of the task.

3.3 Sentiment Annotation Data

To evaluate sentiment model performance, a manually annotated subset comprising 10%, 138 documents, of the dataset is constructed as ground truth. This subset consists of randomly selected

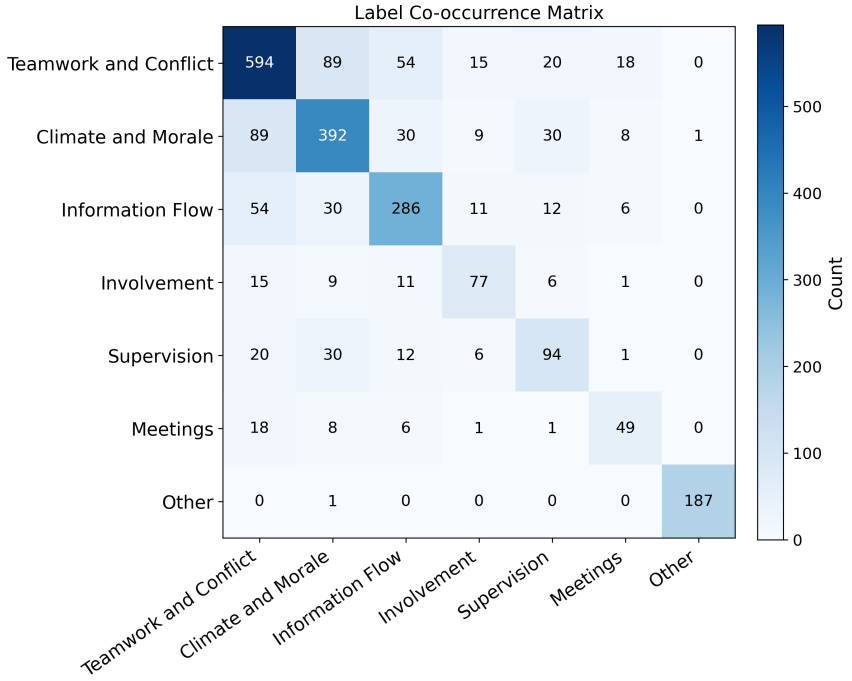


Figure 4: Label co-occurrence matrix showing how often categories appear together.

documents. Each document is assigned a sentiment label on an ordinal scale ranging from -1.0 (very negative) to $+1.0$ (very positive), with intermediate values of -0.5 (somewhat negative), 0.0 (neutral or mixed), and 0.5 (somewhat positive). This scale is chosen for its interpretability and its alignment with the directional nature of sentiment, where negative values indicate unfavourable sentiment, positive values indicate favourable sentiment, and the magnitude reflects intensity. The ordinal formulation enables the use of both categorical metrics and distance-based measures such as Mean Absolute Error (MAE), which capture the severity of prediction deviations. The selected subset size reflects a trade-off between annotation effort and statistical reliability, while preserving sufficient variability in sentiment expressions for robust evaluation.

Figure 5 shows that the annotated labels exhibit a clear central tendency, with the majority of samples concentrated around the neutral/mixed category ($n = 59$). The distribution is moderately symmetric, with a gradual decrease toward the extremes, although slightly skewed toward the positive side (34 somewhat positive and 14 very positive versus 30 somewhat negative and only 1 very negative). This imbalance indicates that strongly negative sentiment is under-represented, which may limit the models' ability to robustly be evaluated on extreme negative cases. At the same time, the dominance of neutral and mildly valenced samples reflects the nature of organizational communication, where sentiment is often implicit, nuanced, and less polarized.

Chapter 4 Methodology

In this section, we describe our methodological design choices. Our pipeline consists of the three most dominant approaches in the literature discussed in Sections 2.2.1, 2.2.2: classification, topic

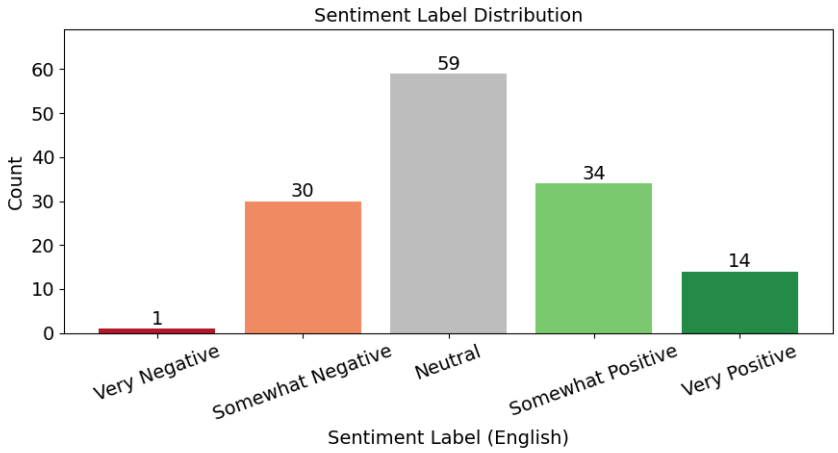


Figure 5: Sentiment label distribution showing how often each label appears.

modelling, and sentiment analysis. Each component of the pipeline is discussed separately.

A central principle guiding our experiment design is the balance between simplicity and performance. From a business perspective, locally deployed and computationally efficient models offer greater independence and cost predictability. Avoiding external APIs reduces exposure to vendor lock-in, pricing volatility, and contractual changes affecting access or data handling. It also enables more reliable budgeting by eliminating usage-based inference fees. In addition, in-house deployment strengthens control over data governance, compliance, and model update cycles, which is particularly important when processing sensitive organizational data. Thus, in this research we compare performance of traditional language models with LLMs for our three culture signals: classification in the OCS, topic modelling and sentiment classification.

4.1 Pre-Processing

The interview transcripts were converted from PDF to structured text and subsequently segmented into predefined components, including general notes, interview sections, question–answer pairs, and additional reflections. Special sections were automatically detected using rule-based patterns. This structured segmentation enables flexible use of the data within the analysis pipeline, allowing both full-document processing and fine-grained, component-level analysis for specific NLP tasks.

Within the pipeline, each individual answer is treated as a separate document. This design choice is driven by both technical and methodological considerations. As discussed in Section 3.1, BERT-based models impose a maximum input length of 512 tokens, which is frequently exceeded when processing entire interviews. More importantly, full interviews often contain multiple themes and conversational transitions, resulting in heterogeneous topic mixtures that hinder precise topic modeling and classification. In contrast, individual answers tend to be more topically focused and semantically coherent, providing a more granular and analytically useful representation of organizational discourse. Consequently, answer-level segmentation ensures compatibility with model constraints while improving the quality of downstream analyses.

To further protect privacy and confidentiality, all documents are anonymized through systematic placeholder substitution. Company names are replaced with *company_x*, project names with *project_x*, and personal names with *person_x*. Importantly, these placeholders remain unique identifiers within the dataset, allowing models to preserve distinctions between entities without exposing sensitive information. This approach balances privacy preservation with analytical utility.

Finally, for the topic modeling experiment using LDA, a cleaned version of each document is constructed to enhance topic quality. Each segment is first lowercased to ensure lexical consistency. A regular expression is then applied to remove non-alphanumeric characters while preserving accented Dutch characters. The text is tokenized on whitespace, after which Dutch stopwords are removed using the predefined list from `NLTK`, eliminating high-frequency function words with limited semantic value. The remaining tokens are subsequently recombined into a single string, resulting in a compact, noise-reduced representation that improves the ability of LDA to capture the latent thematic structure of the corpus.

4.2 Classification

To introduce construct validity into the NLP pipeline, we incorporate a supervised classification task grounded in an established theoretical framework. Specifically, each document is mapped onto one or more dimensions of the OCS framework (Glaser et al., 1987). Framing the task as multi-label classification reflects the theoretical assumption that cultural dimensions are not mutually exclusive and may co-occur within a single interview answer. By operationalizing cultural constructs through an externally validated model, this step anchors the computational analysis in established organizational theory.

4.2.1 Experimental Setup

The automated classification task is formulated as a multi-label text classification problem. While the original OCS framework comprises six predefined categories, the annotation process revealed a non-trivial subset of documents that could not be meaningfully assigned to any of these dimensions. Forcing such documents into existing categories would risk introducing noise and undermining construct validity. To address this, we introduce an additional category, *Other*, which captures segments that do not align with the OCS framework or do not reflect organizational culture as defined in this study. Importantly, *Other* is treated as a mutually exclusive category: documents assigned to *Other* are not assigned to any of the six OCS dimensions. This design ensures a clear separation between theoretically grounded cultural signals and out-of-scope or non-conforming content, while preserving the integrity of the original framework. Each document x_i is represented by a binary label vector $y_i \in \{0,1\}^7$, where $y_{ic} = 1$ indicates the presence of category c.

We compare two transformer-based classifiers, the fully dutch model called BERTje (de Vries et al., 2019) and a multilingual XLM-RoBERTa (Conneau et al., 2019), with a instruction-tuned LLM, GPT-5-mini. For the transformer-based models, we fine-tune a sigmoid-activated classification head to enable independent probability estimation per label. Training minimizes binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{7}\left[y_{ic}\log(\hat{y}_{ic}) + (1-y_{ic})\log(1-\hat{y}_{ic})\right], \tag{1}$$

where $\hat{y}_{ic} \in [0,1]$ denotes the predicted probability for class c in document x_i.

Hyperparameters are optimized using grid search on the validation set to ensure controlled and reproducible model selection. For transformer-based classifiers, performance is particularly sensitive to optimization dynamics and regularization. We therefore tune hyperparameters that directly influence convergence behaviour, generalization capacity, and overfitting risk. The learning rate determines the step size of gradient updates and is known to critically affect fine-tuning stability in pretrained transformers. Batch size influences gradient variance and training stability. The number of training epochs controls the degree of adaptation to the task-specific data and must balance underfitting and overfitting. Finally, dropout probability regulates model capacity by introducing stochastic regularization in the classification head. The explored search space is summarized in Table 6.

Table 6: Hyperparameter search space for BERT-based classifiers.

Hyperparameter	Values searched
Learning rate	$\{2e^{-5}, 3e^{-5}, 5e^{-5}\}$
Batch size	$\{2, 4, 8, \}$
Number of epochs	$\{3, 5, 7\}$
Dropout probability	$\{0.1, 0.2, 0.3\}$

Prior research on multi-label classification similarly finds that optimising per-label thresholds can substantially improve F1-based performance without retraining the underlying model (Pillai et al., 2013). Given the substantial label imbalance, as shown in Section 3.2, we also experiment with this threshold tuning. After model training, each document in the validation set x_i receives a probability score $\hat{y}_{ic}$ for every class c. To convert these probabilities into binary predictions, we determine a separate threshold τ_c for each class using the validation set. Concretely, for each class c, we evaluate a range of candidate threshold values in the interval $[0,1]$, in increments of 0.01. For each candidate threshold, predicted probabilities on the validation set are converted into binary labels, and the corresponding $F1_c$ score, shown in Equation 2, is computed. The threshold τ_c is then selected as the value that maximizes $F1_c$ on the validation data. This procedure yields one optimized decision boundary per class. This threshold calibration is performed exclusively on the validation set, and the selected τ_c values are fixed before evaluating final performance on the held-out test set. This approach allows the model to adjust its operating point to class imbalance while maintaining strict separation between model selection and final evaluation.

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \ \text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \ \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

Binary predictions after training are subsequently defined as:

$$\hat{y}_{ic}^{binary} = \begin{cases} 1 & \text{if } \hat{y}_{ic} \geq \tau_c, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For GPT-5-mini, we evaluate two prompting regimes. In the zero-shot condition, the model receives task instructions, formal category definitions (following Appendix A), and structured output constraints. In the few-shot condition, the prompt additionally includes three random, labelled demonstrations drawn exclusively from the training set. These examples clarify category boundaries and illustrate possible co-occurrence patterns, allowing assessment of minimal in-context adaptation without extensive prompt engineering. We chose not to include a full-dataset prompting condition, as excessively long prompts increase token costs, reduce inference efficiency, and may dilute relevant signals within the finite context window. Prior work also suggests that adding large numbers of demonstrations does not necessarily improve performance and can even degrade results (Huang et al., 2023).

All models are evaluated on the metrics discussed in Section 4.2.2, using a fixed train/validation/test split of 80% train, 10% validation and 10% test. Hyperparameter tuning, threshold calibration, and selection of few-shot examples rely solely on training and validation data. Final performance is reported on the held-out test set.

4.2.2 Metrics

Model performance is evaluated using complementary multi-label classification metrics. Because label frequencies are imbalanced and documents may contain multiple simultaneous categories, no single metric sufficiently captures performance characteristics. We therefore report Hamming Loss and macro-averaged F1.

Hamming Loss Hamming Loss measures the proportion of incorrectly predicted label assignments across all documents and classes:

$$\text{HammingLoss} = \frac{1}{N \cdot 7} \sum_{i=1}^{N} \sum_{c=1}^{7} \mathbf{1}\left(y_{ic} \neq \hat{y}_{ic}^{binary}\right). \tag{4}$$

This metric evaluates per-label error independently, penalizing both false positives and false negatives equally. Its value ranges between 0 and 1, where lower values indicate better performance. Hamming Loss does not require all labels of a document to be correct simultaneously, making it less sensitive to multi-label strictness. It therefore provides a fine-grained measure of label-wise prediction quality.

Macro-averaged F1 Macro-F1 computes the F1 score independently for each class and then averages across classes:

$$F1_{macro} = \frac{1}{7} \sum_{c=1}^{7} F1_c. \tag{5}$$

In contrast to micro-F1, macro-F1 assigns equal weight to each class, irrespective of frequency. This makes it particularly informative under label imbalance, as it highlights performance on minority categories that may otherwise be obscured. Its values ranges from 0 to 1, with higher macro-F1 indicating more balanced classification performance across all cultural dimensions.

Together, these metrics capture complementary aspects of model performance: Hamming Loss reflects per-label error rates, and macro-F1 assesses robustness to class imbalance Reporting both metrics provides a comprehensive evaluation under multi-label and imbalanced conditions.

4.2.3 Inference

At inference time, previously unseen answers are classified using the best-performing approach identified in Section 4.2.1. For transformer-based models, class-specific thresholds selected during training are applied to convert probabilities into binary label assignments, whereas GPT-based models directly generate category labels through prompting.

4.3 Topic Modelling

Topic modelling complements the classification stage by uncovering latent themes in the interview data. Whereas classification assigns responses to predefined categories, topic modelling follows a data-driven approach that surfaces patterns specific to each organization. This enables the analysis to capture context-dependent issues that may not align with existing theoretical constructs, addressing earlier concerns regarding overly rigid measurement frameworks (Druckman et al., 1997).

4.3.1 Experimental Setup

Model Selection We compare three topic modelling approaches: LDA, BERTopic, and GPT5-mini. LDA and BERTopic are included because they are among the most commonly used topic modelling approaches in the literature. For BERTopic we use the Dutch pre-trained BERT model from de Vries et al. (2019) called BERTje. The inclusion of GPT-based models serves to evaluate whether recent generative models can produce more coherent and interpretable OC topics without relying on explicit topic modelling architectures.

Table 7: Hyperparameter search space for topic modeling experiments

Model	Component	Parameters and Values
BERTopic	UMAP	$n\_neighbors \in \{5, 10\}$, $n\_components \in \{5, 10\}$
	HDBSCAN	$min\_cluster\_size \in \{3, 5, 10\}$, $min\_samples \in \{1, 3, 5, 10\}$, $cluster\_selection\_method \in \{\text{eom}, \text{leaf}\}$
LDA	–	$num\_topics \in \{10, 25, 50, 100, 150\}$, $doc\_topic\_prior \in \{\text{None}, 0.01, 0.05, 0.1, 1.0\}$, $topic\_word\_prior \in \{\text{None}, 0.01, 0.05, 0.1, 1.0\}$

Hyperparameter Optimisation For both LDA and BERTopic, the first step consists of hyperparameter optimisation, as model configuration has a substantial impact on topic quality. A grid search is performed over the parameter ranges shown in Table 7, resulting in 96 configurations for BERTopic, and 125 for LDA. For BERTopic, tuning focuses on the UMAP and HDBSCAN components, as these directly influence embedding compression and cluster formation, and therefore the resulting topics. For LDA, tuning focuses on the number of topics and the document–topic and topic–word priors, which control sparsity and distributional smoothness. All model configurations are evaluated using the quantitative metrics described in Section 4.3.2. Based on these metrics, six configurations are selected for a subsequent qualitative assessment by a human on a small, fixed subset of documents, with the aim of evaluating quality and practical topic usefulness. This step is necessary as the literature shows quantitative metrics do not always allign with humand evaluation (Wu et al., 2024).

Configuration Selection Next we select 3 representative LDA and 3 BERTopic configurations for further analysis. The selection of the six model configurations proceeds in multiple stages. First, a hard exclusion step removes unsuitable models. Configurations with fewer than ten topics are discarded, as topic modelling is intended to provide greater granularity than the preceding classification step with six bins. Second, for each evaluation metric, the lowest-performing 25% of configurations are excluded. This step enforces a minimum quality threshold on each metric, ensuring that configurations with very weak performance on any single quantitative dimension are not forwarded to subsequent analysis, even if they are not dominated[6] in a pareto sense. Third, configurations that are strictly dominated by others are excluded. Dominance is assessed separately for LDA and BERTopic to ensure that a diverse set of competitive model configurations is retained within each modelling approach. For BERTopic we also include relative anomaly size as a measure to be minimised. Relative anomaly size is described in Equation 6, where N is the amount of total documents and K the amount of produced topics. This anomaly size reflects the proportion of documents assigned to the outlier or "other" cluster relative to an approximation of other topic sizes. Finally, three configurations per model family are selected from the resulting configuration set. During this selection, near-duplicate configurations are avoided to ensure diversity among the candidates forwarded to the human evaluation.

$$anomaly\_rel = \frac{anomaly\_size}{N/K} \tag{6}$$

Model Comparison In the third stage, the selected LDA and BERTopic configurations are benchmarked against GPT5-mini. The GPT model includes two parameters: (1) zero-shot vs. few-shot prompting and (2) the use of a topic list. In the zero-shot setting, the prompt contains only instructions describing the topic modelling task. In the few-shot setting, the prompt additionally includes a small number of labelled examples to guide the model. When the topic list parameter is enabled, the prompt is augmented with a list of previously generated topic labels. This list aims to reduce the creation of redundant topics that differ only in minor wording variations, encouraging the model to reuse existing labels when appropriate.

[6]A configuration is dominated if all metric values are $\leq$ those of another configuration and at least one metric value is strictly worse.

The comparison proceeds along two dimensions. First, all models and configurations are evaluated using the quantitative metrics defined in Section 4.3.2. Second, a structured human evaluation is conducted. For a subset of 4% of documents, the top 20 topic words generated by each model are assessed in relation to the corresponding document. We adopt a topic-rating setup inspired by Hoyle et al. (2021). The human evaluator assess each topic based on two criteria. The first is **topic quality**, defined as the degree to which the topic words reflect the content of the associated document, indicating how well the topic represents the document's main theme. The second is **business relevance**, defined as the extent to which the topic words capture the underlying cultural value, meaning, or practice expressed in the document, indicating whether the topic contributes meaningfully to the OC analysis task. Both criteria are scored on a 7-point Likert scale, following evidence that seven response categories optimize reliability and discriminative capacity in subjective evaluations (Cai et al., 2016).

For the LLM-based approaches, each document is processed independently, with the model instructed to generate a single topic label summarizing the OC theme present in the text. Because LLMs do not natively produce topic–word distributions, class-based TF–IDF (c-TF-IDF) is applied post hoc to derive representative topic terms from the generated labels. This harmonizes the output format across methods, ensuring that all approaches yield comparable topic representations and enabling consistent computation of the evaluation metrics.

4.3.2 Metrics

In order to evaluate we use three metrics. Two commonly used metrics to asses the quality of topic models are topic coherence and topic diversity. We also use semantic aware topic diversity as suggested by Wu et al. (2024).

Cv-coherence Coherence quantifies the sense of thematic alignment within topics. In this paper we use C_V coherence, as Röder et al. (2015) show that it achieves the highest correlation with human judgments of topic quality compared to other coherence variants. We compute C_V using the implementation provided by the `gensim` python package. The measure is an indirect coherence metric that combines normalized pointwise mutual information (NPMI) with cosine similarity over word context vectors. Given a topic represented by its top-N words $W = \{w_1, \ldots, w_N\}$, a context vector $\mathbf{v}_i \in \mathbb{R}^N$ is constructed for each word w_i as

$$\mathbf{v}_i = \big(\mathrm{NPMI}(w_i, w_1), \ldots, \mathrm{NPMI}(w_i, w_N)\big), \tag{7}$$

where

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}. \tag{8}$$

Word probabilities are estimated using a sliding window over a large reference corpus. Topic coherence is then computed as the average cosine similarity between each word vector and the centroid vector $\mathbf{v}_c = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$:

$$C_V(W) = \frac{1}{N}\sum_{i=1}^{N} \cos(\mathbf{v}_i, \mathbf{v}_c). \tag{9}$$

$C_V \in [0, 1]$, where higher values indicate more semantically coherent and interpretable topics.

Topic diversity To assess redundancy across topics, we compute topic diversity following a commonly used definition in topic model evaluation. Let a topic model produce K topics, each represented by its top-N words, and let $\mathcal{W} = \{w_{k,n} \mid k = 1, \ldots, K;\ n = 1, \ldots, N\}$ denote the multiset of all top words across topics. Topic diversity is defined as the proportion of unique words across all topic representations:

$$\mathrm{TD} = \frac{|\mathrm{unique}(\mathcal{W})|}{K \cdot N}. \tag{10}$$

$TD \in [0, 1]$, where higher values indicate less lexical overlap between topics and thus greater topic distinctiveness.

Semantic-aware topic diversity In addition to lexical topic diversity, we compute semantic-aware topic diversity (TSD) following Wu et al. (2024). TSD measures redundancy at the level of word *pairs* rather than individual words. Let a topic model produce K topics, each represented by its top-T words. For a topic k, let $t^{(k)}$ denote the set of all unordered word pairs (x_i, x_j) with $i < j$. Across all topics, let $\#(x_i, x_j)$ denote the total number of occurrences of a given unordered word pair. TSD is then defined as

$$\mathrm{TSD} = \frac{2}{K \cdot T \cdot (T-1)} \sum_{k=1}^{K} \sum_{(x_i, x_j) \in t^{(k)}} \mathbb{I}\big(\#(x_i, x_j)\big), \tag{11}$$

where $\mathbb{I}(\cdot)$ refers to an indicator function that equals to 1 if $\#$(xi, xj) $= 1$ and equals 0 otherwise. $TSD \in [0, 1]$, with higher values indicating greater semantic diversity across topics due to fewer repeated word-pair structures.

4.3.3 Inference

At inference time, we employ the best-performing model identified in the experiment described in Section 4.3.1. This model discovers topics in an unsupervised manner using data aggregated across all companies, resulting in a more robust and stable topic representation. The trained model is subsequently used to assign the learned topics to previously unseen documents. This setup enables effective modelling of recurring themes even when individual companies contribute relatively little to the dataset, while still allowing for the extraction of company-specific insights through differential topic prevalence. Importantly, this approach preserves an information barrier at inference time, such that each company is exclusively exposed to its own data and results, while De Hofnar benefits from learning a shared topic structure across all participating organizations.

4.4 Sentiment Analysis

The primary analytical focus of this thesis is on what is being said and by how many employees, making classification and topic modelling the core components of the pipeline. Sentiment analysis is therefore treated as a supporting signal that provides contextual information about identified themes, rather than as an independent modelling objective. Training a domain-specific sentiment model would require substantial annotated data and expert labelling effort, which is not proportionate to its auxiliary role in the framework. Therefore, sentiment analysis is implemented using pre-trained transformer-based models and a GPT-based large language model, without domain-specific fine-tuning.

4.4.1 Experiment Setup

For the transformer-based approach, the multilingual sentiment model trained by NLP Town (2020) is employed. This model predicts sentiment on a discrete ordinal scale ranging from 1 to 5 stars. For visualisation purposes, these ratings are linearly mapped to a scale between -1.0 and $+1.0$, as explained in Section 3.3. The model supports six languages, including Dutch, and is trained on a large corpus of product reviews, of which approximately 12% (80k samples) are Dutch. Reported performance on Dutch data indicates an accuracy of 57% for exact predictions and 93% accuracy within a one-star margin. This model was selected due to its multilingual capability, availability of Dutch training data, and its formulation of sentiment as an ordinal prediction task. To the best of our knowledge, no publicly available sentiment models are specifically trained on Dutch OC data. Consequently, a model trained on product reviews is adopted as a pragmatic baseline, acknowledging that while it is not ideally aligned with the domain, it represents the most suitable publicly available alternative.

As a second approach, sentiment is inferred using a GPT-based LLM through structured prompting. The prompt explicitly constrains the model to assign one of five ordinal sentiment labels [-1.0, 1.0] based on the input text. This enables zero-shot classification, leveraging the generalization capabilities of LLMs without additional training. This setup allows for a comparison between a fixed, domain-independent classifier and a flexible, context-aware model, highlighting the trade-off between computational efficiency and contextual reasoning. Both of these models are evaluated on the dataset detailed in Section 3.3.

4.4.2 Metrics

Model performance is evaluated using both classification-based and distance-aware metrics. While sentiment prediction can be framed as a classification task, the ordinal nature of sentiment labels (e.g., 1–5) implies that not all errors are equally severe. Therefore, in addition to standard classification metrics, regression-style metrics are included to account for the magnitude of prediction errors.

Accuracy Accuracy measures the overall proportion of correctly predicted labels:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i = \hat{y}_i) \tag{12}$$

where y_i is the true label, $\hat{y}_i$ is the predicted label, and $\mathbb{1}(\cdot)$ is the indicator function that returns 1 when the prediction was correct. Accuracy ranges from 0 to 1, where higher values indicate a larger proportion of exact matches. However, it does not account for the severity of misclassification.

Precision Precision measures the reliability of positive predictions for a given class:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

where TP denotes true positives and FP false positives. Precision ranges from 0 to 1, with higher values indicating fewer false positive predictions for that class.

Recall Recall measures the model's ability to identify all relevant instances of a given class:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

where FN denotes false negatives. Recall ranges from 0 to 1, with higher values indicating that fewer true instances are missed.

F1-score The F1-score provides a balanced measure between precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

The F1-score ranges from 0 to 1, where higher values indicate a better trade-off between false positives and false negatives.

Mean Absolute Error (MAE) To explicitly account for the ordinal nature of sentiment labels, MAE is included:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{16}$$

MAE measures the average absolute distance between predicted and true labels. Lower values indicate better performance, with a value of 0 representing perfect predictions. Unlike classification metrics, MAE penalizes predictions proportionally to their distance from the true label, making it particularly suitable for ordinal sentiment tasks.

4.4.3 Inference

At inference time, the best-performing sentiment model identified in Section 4.4.1 is applied to all unseen documents. Each document is assigned a sentiment catagory ranging from -1.0 to 1.0. These sentiment scores are subsequently aggregated at the level of OC categories and topics. This enables the identification of patterns such as consistently negative sentiment within specific themes or departments. As such, sentiment acts as a contextual signal that enriches the interpretation of classification and topic modelling outputs, supporting the generation of actionable organizational insights.

Chapter 5 Results

In this section, we present the results of the previously described experiments. We begin with the classification results, followed by the four stages of the topic modeling analysis, and conclude with the sentiment analysis results. Each experiment is discussed in a dedicated subsection to ensure clarity and structured interpretation. We additionally demonstrate how these extracted cultural signals can be integrated into a dashboard to support actionable insights.

Table 8: Summary of grid search results across model families. Each value represents the minimum, mean, and maximum performance over all evaluated configurations within the family.

Model Family	HL_{min}	$HL_{\textbf{mean}}$	HL_{max}	$F1_{\text{min}}$	$F1_{\textbf{mean}}$	$F1_{\text{max}}$
BERTje	0.0994	**0.1129**	**0.1377**	0.2341	0.4467	0.5707
BERTje (Thresholds Tuned)	**0.0890**	0.1286	0.2091	**0.4625**	**0.5771**	**0.6769**
XLM-RoBERTa	0.1014	0.1492	0.1781	0.0000	0.1466	0.4148
XLM-RoBERTa (Thresholds Tuned)	0.1201	0.5431	0.8292	0.2636	0.3495	0.5766
GPT	0.2547	0.2619	0.2692	0.3765	0.3930	0.4096

Note: Hamming loss measures the fraction of incorrectly predicted labels; lower values indicate better performance.

5.1 Classification Results

Table 8 summarizes the grid search outcomes across all evaluated model families, including both base transformer models and their threshold-tuned counterparts. Overall, the threshold-tuned *BERTje* variant demonstrates the strongest performance, achieving the lowest observed Hamming loss ($HL_{\text{min}} = 0.0890$) alongside the highest mean and maximum macro-F1 scores (0.5771 and 0.6769, respectively). In contrast, the untuned *BERTje* model attains the lowest mean and maximum Hamming loss values (0.1129 and 0.1377), indicating more stable error rates across configurations but at the cost of reduced peak F1 performance.

For *XLM-RoBERTa*, the untuned variant achieves a competitive best-case Hamming loss (0.1014), whereas threshold tuning substantially improves its maximum macro-F1 score (from 0.4148 to 0.5766). This gain, however, is accompanied by a marked degradation in Hamming loss, with the tuned variant exhibiting significantly higher mean and maximum values (0.5431 and 0.8292), suggesting increased prediction noise or overcompensation in threshold calibration.

The GPT-based approach performs consistently worse across all metrics, with a minimum Hamming loss of 0.2547 and macro-F1 scores confined to a relatively narrow range (0.3765–0.4096). Compared to the transformer-based families, GPT models show both higher error rates and limited variability in F1 performance, indicating weaker sensitivity to configuration changes and overall lower effectiveness for this classification task.

Table 9 reports, for each model family, the configurations achieving the best Hamming loss and the best macro-F1 score. For *BERTje*, the optimal configurations differ across metrics: the lowest Hamming loss is obtained with a batch size of 4, dropout of 0.1, learning rate of 3×10^{-5}, and 5 epochs, whereas the highest macro-F1 score is achieved with a smaller batch size of 2, higher dropout (0.3), a lower learning rate of 2×10^{-5}, and 7 epochs. The optimal threshold-tuned *BERTje* configuration outperforms both, achieving the best results on both metrics with a batch size of 8, dropout of 0.3, learning rate of 3×10^{-5}, and 7 epochs.

For *XLM-RoBERTa*, the configurations yielding the best Hamming loss and macro-F1 score are nearly identical, both using a batch size of 8, learning rate of 2×10^{-5}, and 7 epochs, and

Table 9: Best configuration per model family and selection criterion

Family	**Selected by**	**HL**	$F1_{\textbf{macro}}$	**Parameters**
BERTje	best_hamming_loss	0.0994	0.4947	{"batch_size": 4, "dropout": 0.1, "lr": 3e-05, "epochs": 5}
BERTje	best_f1_macro	0.1035	0.5707	{"batch_size": 2, "dropout": 0.3, "lr": 2e-05, "epochs": 7}
BERTje (Thresholds Tuned)	joint_best	**0.0890**	**0.6769**	{"batch_size": 8, "dropout": 0.3, "lr": 3e-05, "epochs": 7}
XLM-RoBERTa	best_hamming_loss	0.1014	0.3993	{"batch_size": 8, "dropout": 0.3, "lr": 2e-05, "epochs": 7}
XLM-RoBERTa	best_f1_macro	0.1170	0.4148	{"batch_size": 8, "dropout": 0.2, "lr": 2e-05, "epochs": 7}
XLM-RoBERTa (Thresholds Tuned)	joint_best	0.1201	0.5766	{"batch_size": 4, "dropout": 0.3, "lr": 2e-05, "epochs": 7}
GPT	joint_best	0.2547	0.4096	{"few-shot": 3}

Note: HL denotes Hamming loss (lower is better). Only key hyperparameters are shown for readability.

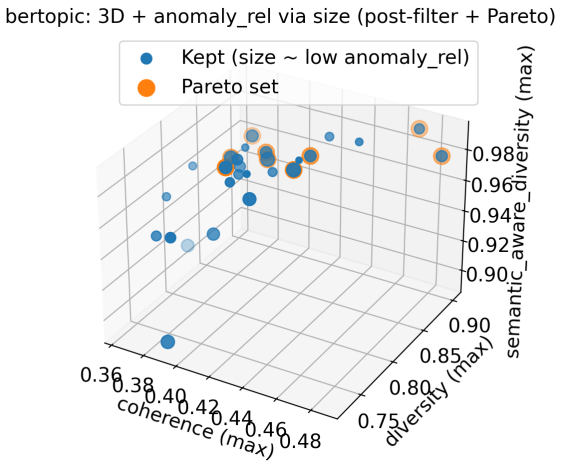
differing only in dropout (0.3 for Hamming loss versus 0.2 for macro-F1). The threshold-tuned variant achieves $HL = 0.1201$ and $F1_{\text{macro}} = 0.5766$ with a smaller batch size of 4 and higher dropout (0.3), while maintaining the same learning rate and number of epochs.

For GPT, the few-shot configuration yields the best performance, achieving $HL = 0.2547$ and $F1_{\text{macro}} = 0.4096$, outperforming the zero-shot setup across both metrics.

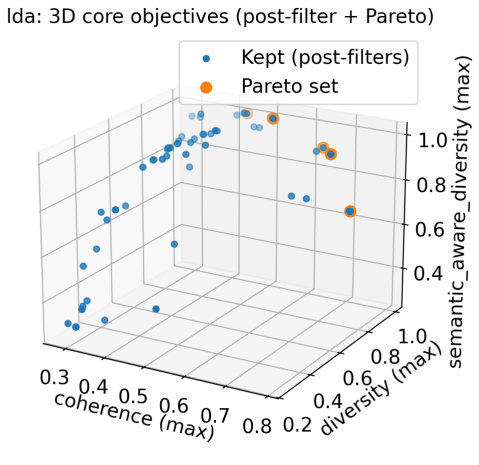
Table 10: Per-class F1 scores for best configurations

Model	Selection	C0	C1	C2	C3	C4	C5	C6
BERTje	best_hamming_loss	0.7941	0.6506	0.8182	0.0000	0.2000	0.5000	0.5000
BERTje	best_f1_macro	0.7717	0.6364	0.7234	0.0000	0.3636	1.0000	0.5000
BERTje (Threshold Tuned)	joint_best	0.7801	0.7073	0.8511	0.3333	0.4000	1.0000	0.6667
XLM-RoBERTa	best_hamming_loss	0.7321	0.6882	0.7500	0.0000	0.0000	0.0000	0.6250
XLM-RoBERTa	best_f1_macro	0.7328	0.6053	0.6038	0.0000	0.4615	0.0000	0.5000
XLM-RoBERTa (Thresholds Tuned)	joint_bes	0.8000	0.6408	0.8163	0.1667	0.4000	0.6667	0.5455
GPT	joint_bes	0.6772	0.6349	0.4946	0.1333	0.2917	0.4615	0.1739
Average	**all models**	**0.7517**	**0.6488**	**0.7076**	**0.1037**	**0.3121**	**0.5281**	**0.4701**

Table 10 reports per-class F1 scores across the best-performing configurations detailed above. Averaged across configurations, the highest predictive performance was obtained for Class 0 (Teamwork and Conflict) with an average F1 score of 0.7517, followed by Class 2 (Information Flow) with 0.7076 and Class 1 (Climate and Morale) with 0.6488. Intermediate performance was observed for Class 5 (Meetings) with 0.5281 and Class 6 (Other) with 0.4701. Lower predictive



(a) BERTopic, with the relative anomaly size being the size of the dots.



(b) LDA

Figure 6: Three-dimensional pareto fronts comparing topic-model configurations across model families. The figures show in orange the non-dominated trade-offs between coherence, topic diversity, and semantic-aware diversity for BERTopic and LDA.

performance was found for Class 4 (Supervision) with 0.3121, while Class 3 (Involvement) was the most difficult category with an average F1 score of 0.1037.

Across individual configurations, the strongest single-class scores were achieved for Class 5, where multiple threshold-tuned models reached an F1 score of 1.0000. For the more frequent categories, high peak scores were observed for Class 2 (0.8511), Class 0 (0.8000), and Class 1 (0.7073). Class 6 reached a maximum F1 score of 0.6667, while Class 4 peaked at 0.4615. For Class 3, the highest observed F1 score was 0.3333.

Threshold tuning substantially affected several minority classes. For BERTje, the threshold-tuned configuration increased Class 3 performance from 0.0000 to 0.3333 and Class 5 from 0.5000 or 1.0000 depending on selection criteria to 1.0000 in the joint-best configuration. For XLM-RoBERTa, threshold tuning increased Class 3 from 0.0000 to 0.1667 and Class 5 from 0.0000 to 0.6667. In contrast, performance for the majority classes (Classes 0, 1, and 2) remained comparatively stable before and after threshold adjustment.

5.2 Topic Modelling Results

5.2.1 Grid Search Results

After completion of the grid search, 35 BERTopic configurations were excluded because they produced fewer than the minimum required number of topics[7]. Subsequently, an additional 34 BERTopic configurations and 64 LDA configurations were removed for falling within the bottom 25% on at least one evaluation metric. The remaining configurations are visualized in Figure 6.

[7]This constraint does not apply to LDA, where the number of topics is an explicit hyperparameter.

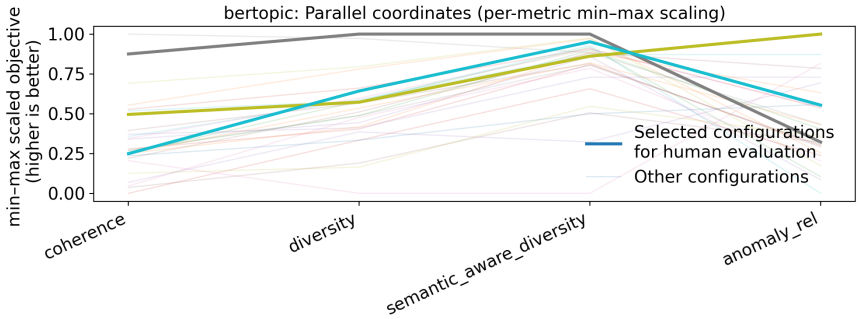


Figure 7: Parallel coordinates plot of BERTopic hyperparameter configurations after filtering. Each line represents a single configuration and each axis corresponds to an evaluation metric. Metrics are independently min–max scaled, such that axis values indicate relative performance within the plotted set rather than absolute scores. Thicker lines highlight the configurations selected for subsequent human evaluation, illustrating the trade-offs between coherence, topic diversity, semantic-aware diversity and relative anomaly size.

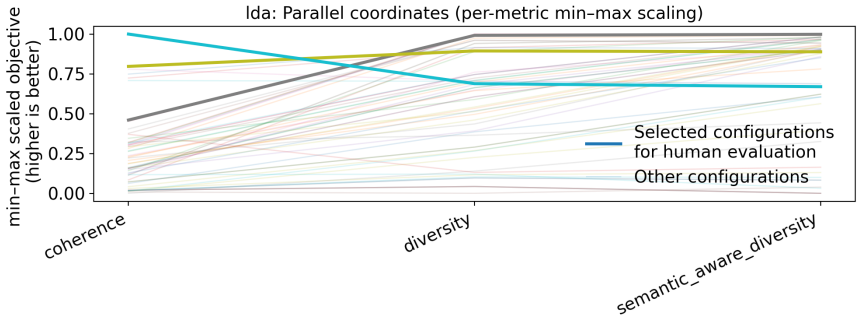


Figure 8: Parallel coordinates plot of LDA hyperparameter configurations after filtering. Each line represents a single configuration and each axis corresponds to an evaluation metric. Metrics are independently min–max scaled, emphasizing relative differences across coherence, topic diversity, and semantic-aware diversity. Thicker lines indicate the configurations selected for human evaluation.

From this reduced set, pareto-optimal configurations were identified using coherence, topic diversity, semantic-aware diversity, and, for BERTopic only, relative anomaly size. This resulted in nine pareto-optimal configurations for BERTopic and seven for LDA, depicted in orange.

5.2.2 Configuration Selection Results

From each pareto front, three representative configurations were selected for qualitative inspection, yielding six configurations in total. Selection was guided by coverage of different trade-offs rather than maximal performance on a single metric. For BERTopic, the selected configurations include models with topic counts ranging from 115 to 210 and anomaly sizes between 214 and 307 documents. For LDA, the selected configurations correspond to topic counts of 50, 100, and 150. These trade-offs are visualized in Figures 7 and 8, which show the performance of all remaining configurations after filtering, normalized using min–max scaling to emphasize relative differences across metrics.

Table 11: Topic model evaluation results.

Model	Coherence	Diversity	Sem.-Aware Div.	Runtime (s)	#Topics
$BERTopic_1$	0.414	0.835	0.984	31.26	125
$BERTopic_2$	0.415	0.815	0.969	31.73	115
$BERTopic_3$	0.466	0.899	0.992	53.60	210
LDA_1	0.684	0.915	0.919	12.67	100
LDA_2	**0.788**	0.762	0.760	16.41	150
LDA_3	0.509	**0.990**	**1.000**	**7.35**	50
GPT_1	0.672	0.728	0.926	10950.76	1042
GPT_2	0.669	0.735	0.932	9517.04	1085
GPT_3	0.360	0.595	0.755	8842.72	45
GPT_4	0.367	0.638	0.806	7778.81	59

5.2.3 Model Comparison Results

Table 11 presents the quantitative evaluation of the selected BERTopic, LDA, and GPT-based topic modelling configurations[8]. In terms of topic coherence, LDA_2 achieves the highest score (0.788). BERTopic configurations yield lower coherence values, ranging from 0.414 to 0.466, while GPT-based configurations show a wider spread between 0.360 and 0.672.

For topic diversity and semantic-aware diversity, LDA_3 consistently attains the highest scores, with 0.990 and 1.000, respectively. It is followed by LDA_1 (0.915 diversity) and $BERTopic_3$ (0.899 diversity; 0.992 semantic-aware diversity), with $BERTopic_1$ also achieving a high semantic-aware diversity score (0.984). GPT-based approaches obtain lower diversity scores overall, ranging from 0.595 to 0.735, and semantic-aware diversity values between 0.755 and 0.932.

Substantial differences are observed in computational efficiency. LDA configurations are the fastest, with runtimes between 7.35 and 16.41 seconds, followed by BERTopic (31.26 to 53.60 seconds). In contrast, GPT-based approaches require significantly longer runtimes, ranging from 7778.81 to 10950.76 seconds.

The number of generated topics varies considerably across methods. BERTopic produces between 115 and 210 topics, while LDA generates a fixed number determined by the hyperparameter k (50, 100, and 150). The GPT configurations exhibit the largest variation, with GPT_1 and GPT_2 producing 1042 and 1085 topics, respectively, and GPT_3 and GPT_4 producing substantially fewer topics (45 and 59).

5.2.4 Human Evaluation Results

Table 12 presents the human evaluation results for the topic modelling approaches. Across all configurations, GPT_1 achieves the highest mean Topic Quality ($TQ\mu = 3.12$) and Business Relevance

[8]Each model configuration is indicated by a subscript number (e.g., $BERTopic_1$). The corresponding hyperparameter settings are provided in Appendix C.

Table 12: Human Evaluation Results per Configuration

Model	n	TQ_{min}	TQ_{μ}	TQ_{med}	TQ_{max}	BR_{min}	BR_{μ}	BR_{med}	BR_{max}
$BERTopic_1$	56	1	1.55	1.0	4	1	1.50	1.0	4
$BERTopic_2$	56	1	1.62	1.0	4	1	1.55	1.0	4
$BERTopic_3$	56	1	1.80	1.0	5	1	1.70	1.0	5
LDA_1	56	1	1.29	1.0	5	1	1.21	1.0	5
LDA_2	56	1	1.23	1.0	5	1	1.18	1.0	5
LDA_3	56	1	1.23	1.0	5	1	1.18	1.0	5
GPT_1	56	1	**3.12**	**3.0**	**7**	1	**2.91**	**3.0**	**7**
GPT_2	56	1	3.00	**3.0**	6	1	2.71	2.0	6
GPT_3	55	1	1.95	2.0	5	1	1.93	2.0	5
GPT_4	56	1	2.04	2.0	4	1	1.98	2.0	4

($BR\mu = 2.91$) scores, followed by GPT2 ($TQ\mu = 3.00$, $BR\mu = 2.71$), indicating consistently stronger human-perceived performance for the GPT-based approaches.

Within the BERTopic family, $BERTopic_3$ yields the highest mean scores ($TQ\mu = 1.80$, $BR_{\mu} = 1.70$), while BERTopic1 and $BERTopic_2$ perform slightly lower, with $TQ\mu$ ranging from 1.55 to 1.62 and $BR\mu$ from 1.50 to 1.55.

The LDA configurations obtain the lowest human evaluation scores overall. LDA1 achieves $TQ\mu = 1.29$ and $BR_{\mu} = 1.21$, while LDA_2 and LDA_3 both reach $TQ\mu = 1.23$ and $BR\mu = 1.18$. Across all LDA configurations, the median Topic Quality and Business Relevance scores are 1.0, indicating consistently low ratings across evaluated topics.

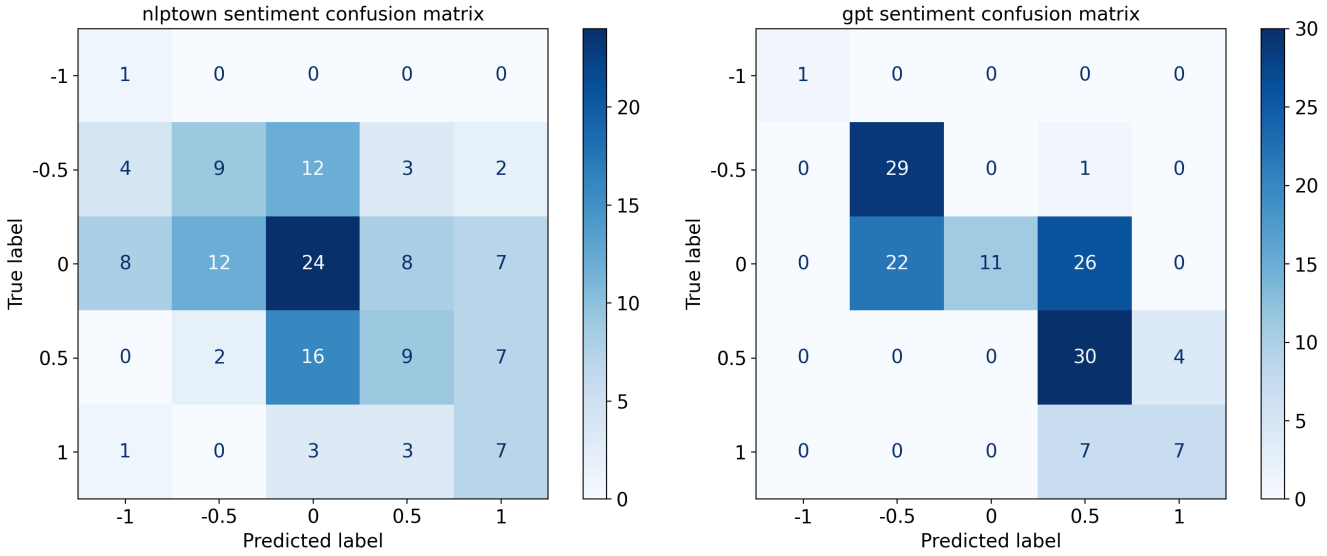
5.3 Sentiment Analysis Results

Table 13 presents the performance of the transformer-based multilingual sentiment model and the GPT-based approach on the normalized sentiment scale ranging from -1.0 to 1.0. GPT achieves higher scores than the NLP town baseline on all predictive performance metrics. Specifically, GPT attains an accuracy of 0.5725, precision of 0.5491, recall of 0.5164, and F1-score of 0.4527, compared with 0.3623, 0.3189, 0.4943, and 0.3176 respectively for NLP town. GPT also achieves a lower mean absolute error (MAE), with a value of 0.2174 compared with 0.4275 for NLP town. In contrast, NLP town completes inference faster, with a runtime of 74.05 seconds compared with 873.81 seconds for GPT.

Table 13: Sentiment evaluation summary

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**MAE**	**Runtime (s)**
NLP town	0.3623	0.3189	0.4943	0.3176	0.4275	**74.05**
GPT	**0.5725**	**0.5491**	**0.5164**	**0.4527**	**0.2174**	873.81

Figure 9 further illustrates the prediction patterns of both sentiment models. NLP town displays substantial dispersion across neighbouring and non-neighbouring sentiment classes, with predictions distributed across nearly all cells of the matrix. Correct classifications are most concentrated around the neutral label (0), where 24 instances are predicted correctly, while positive (0.5) and strongly positive (1.0) classes are frequently underestimated toward neutral or mildly positive categories. GPT exhibits a more concentrated prediction structure, with most observations located on or directly adjacent to the diagonal. Correct classifications are highest for the positive class (0.5) with 30 correct predictions and the negative class (−0.5) with 29 correct predictions. However, GPT shows no correct predictions for the strongly negative class (−1.0), and several neutral (0) and strongly positive (1.0) instances are shifted toward neighbouring classes.



(a) NLP town

(b) GPT

Figure 9: Confusion matrices for sentiment classification showing predictions versus actual sentiment labels.

Chapter 6 Discussion

6.1 Interpretation of Findings

This section interprets the empirical findings beyond model performance alone. Specifically, it considers what the results imply for the practical assessment of organizational culture, how reliable the extracted insights are for managerial use, and where caution remains necessary. In doing so, the findings are connected back to the broader literature on organizational culture measurement and applied NLP.

6.1.1 Classification: Fine-Tuned Transformers Provide the Most Reliable Foundation for Cultural Classification

The most important takeaway from the classification experiment is that fine-tuned transformer-based models, particularly BERTje with threshold calibration, provide the strongest and most reliable performance for detecting OCS dimensions in Dutch organizational text. This indicates that predefined cultural constructs can be operationalized with sufficient accuracy to support automated large-scale analysis. The results show that we are able to detect the dominant OCS dimensions, while minority categories remain more challenging under standard decision thresholds. This pattern is common in imbalanced multi-label classification, where frequent labels contribute more often to the training objective and therefore receive stronger optimisation signals. The strong gains obtained through threshold tuning, especially for *BERTje*, show that part of the performance gap is not caused by representation learning alone, but by suboptimal label decision boundaries in imbalanced multi-label settings. In other words, the model may already encode useful semantic distinctions internally, but the final step of converting probabilities into binary labels can still be poorly calibrated. This is practically relevant because infrequent categories such as Involvement and Meetings may still carry important organizational signals despite their low prevalence. For this research, the results support the use of post-hoc calibration methods for transformer models when applying automated OCS analysis, as they improve coverage of under-represented dimensions without materially reducing performance on the dominant classes.

The transformer models differed not only in predictive performance, but also in robustness. The *BERTje* family consistently outperforms *XLM-RoBERTa* on this Dutch-language task, likely because monolingual pre-training allows greater representational capacity to be devoted to Dutch-specific vocabulary, syntax, and discourse patterns, which is consistent with the findings of de Vries et al. (2019). By contrast, the threshold-tuned *XLM-RoBERTa* results show substantial increases in Hamming loss, indicating a less stable response to calibration. This might be due to *XLM-RoBERTa* having less stable or more overlapping probability distributions. Lowering thresholds then helps recover minority classes, so F1 scores improved, especially because Macro F1 gives equal weight to rare classes. However, those same threshold changes may also have activated many additional incorrect labels across the full label matrix. Hamming loss penalizes every wrong binary decision equally, including extra false positives, so it worsens. This is relevant from an applied perspective, as models intended for repeated use in organizational diagnostics should not only achieve strong best-case results, but also behave reliably across clients and time. Robust models are preferable because organizational datasets are often small, noisy, and heterogeneous, meaning unstable models may produce conclusions that vary depending on minor sampling differences rather than genuine cultural signals. Consistent model behaviour is particularly important in consultancy settings, where clients expect stable conclusions across repeated measurements and comparable projects.

Finally, the GPT-based approach shows a different performance profile from the transformer-based models. It's best results were obtained in the few-shot setup, indicating that in-context examples improve classification quality. This likely occurs because labelled examples make the task expectations more concrete: they clarify category boundaries, output format, and possible label co-occurrences. Rather than relying only on general instructions, the model can infer the intended decision rule from demonstrations, which is a well-established strength of in-context

learning (Mosbach et al., 2023). However, the GPT-based approaches remain consistently below the best fine-tuned transformer models on both Hamming loss and macro-F1. A likely explanation is that fine-tuned classifiers are directly optimized on the target dataset, allowing their parameters to adapt specifically to the language patterns, label distributions, and domain nuances of the OC task. This suggests that, for supervised OC classification tasks with annotated data available, task-specific transformer models remain the stronger choice when compared with a few-shot setting LLM. From a business perspective, this also supports the use of smaller local models for the classification component, as they combine stronger predictive performance in this use case with greater independence from external API-based inference.

These findings are broadly consistent with prior NLP literature. Minaee et al. (2022) identify fine-tuned transformer models as strong and reliable baselines for text classification, while Koch and Pasch (2023) similarly report that transformer architectures outperform simpler approaches in automated culture measurement tasks. Our results also align with Ginting et al. (2023), who demonstrate that organizational culture can be meaningfully operationalized through supervised text classification. In contrast, the GPT-based few-shot approach remained below the best fine-tuned models. This partially contrasts with recent studies reporting that large instruction-tuned models can perform competitively in zero-shot and few-shot text classification tasks (Chae and Davidson, 2026). An explanation could be that such gains are strongest in English, single-label, or clearly specified benchmark tasks, whereas the present setting involves Dutch multi-label organizational text with domain-specific category boundaries.

For De Hofnar, these findings suggest that supervised classification is currently the most deployment-ready component of the pipeline and can serve as the structural backbone of automated cultural diagnostics. Rather than manually coding hundreds of interview fragments, organizations can use the classifier to consistently map employee statements onto validated cultural dimensions. This enables faster reporting, more standardized analysis across projects, and longitudinal comparison over time. For example, repeated negative statements classified under Information Flow may indicate communication issues during organizational change, while recurring signals under Supervision may justify targeted leadership development or management coaching. In this sense, the classifier can help direct attention toward the areas most likely to require intervention. More broadly, this demonstrates that validated organizational culture frameworks such as the OCS can be translated into computable constructs without losing practical interpretability.

At the same time, caution is required when interpreting these outputs. Classification results should be treated as aggregated indicators of recurring organizational patterns rather than as precise judgments about individuals, teams, or causal mechanisms. Low-frequency categories such as Involvement remain more difficult to detect and should therefore be interpreted alongside qualitative evidence rather than in isolation. Similarly, a high prevalence of conflict-related statements does not by itself prove dysfunctional collaboration, but rather signals an area worthy of further investigation.

6.1.2 Topic Modelling: GPT Generates the Most Human-Relevant Themes, but Topic Representation Remains a Key Limitation

The most important takeaway from the topic modelling experiment is that GPT-based approaches generate the most human-relevant and practically meaningful themes, while traditional topic models achieve stronger automated scores but weaker interpretability. This distinction is critical because the primary objective of topic modelling within this thesis is not merely statistical optimisation, but the discovery of themes that consultants and managers can recognize and act upon. Although LDA performs best on several quantitative metrics, particularly coherence and diversity, it performs lowest on both human evaluation dimensions. One explanation is that LDA appears to converge toward highly dominant generic topics assigned to many documents. This suggests a structural mismatch between LDA's bag-of-words assumptions and the nuanced, context-dependent nature of organizational culture text. Broad clusters of common terms may therefore appear statistically coherent even when humans perceive them as vague, repetitive, or insufficiently informative. For this research, the findings indicate that strong automated scores alone are insufficient for selecting topic models intended to support organizational diagnosis.

The GPT-based approaches show a different performance profile. GPT_1 and GPT_2 achieve the strongest Topic Quality and Business Relevance scores, indicating that generative models are better able to summarize document meaning in ways that humans consider useful. A likely explanation is that LLMs can leverage broader semantic and contextual knowledge to infer the latent issue expressed in a document rather than relying solely on lexical co-occurrence. However, these models also generate extremely large numbers of topics. Because each document receives a generated label and is grouped only with documents assigned the exact same label, semantically similar texts may receive slightly different labels and become fragmented across many small clusters. As a result, the outputs move partly from topic modelling toward document-level summarization or keyword extraction. This implies a practical trade-off between specificity and consolidation: GPT produces more meaningful labels, but less stable shared topic structures.

The BERTopic configurations occupy an intermediate position between LDA and GPT. Relative to LDA, BERTopic yields stronger human evaluation scores, suggesting that contextual embeddings and clustering capture more document-level variation than bag-of-words methods. This likely occurs because semantically similar responses can be grouped even when employees use different vocabulary to describe the same issue. In organizational settings, where concerns such as mistrust, unclear leadership, or change fatigue may be phrased in multiple ways, this is an important advantage. However, BERTopic remains substantially below GPT on both Topic Quality and Business Relevance, indicating that its generated topic representations still only partially capture the intended meaning of the underlying text. BERTopic therefore appears to be a more balanced but still imperfect compromise between interpretability and scalable clustering.

More broadly, the findings highlight an important limitation of representing topics through top-n words alone, as indicated by the low human evaluation scores. In the domain of organizational culture, salient meanings are often embedded in narratives, implicit tensions, and contextual relationships rather than in isolated word co-occurrence patterns. For example, a generated topic may consist of terms such as ["making a distinction", "distinction", "between rental", "pull functions", "each

other later", "distinction again"], while a more meaningful abstraction would be a label such as "Function Separation". In such cases, keyword-based representations fail to capture the underlying organizational issue expressed in the text. This is particularly problematic in business contexts, where the objective is not merely to cluster documents, but to surface interpretable and actionable cultural patterns.

These findings are broadly consistent with prior literature. Wu et al. (2024) report that commonly used topic metrics often correlate weakly with human judgment, which aligns strongly with the present results. Gupta et al. (2022) similarly show that text mining can uncover organizational themes beyond traditional survey frameworks, supporting the value of exploratory topic discovery. Our findings also partially align with recent LLM-based topic modelling studies such as Pham et al. (2024), who report improved interpretability from generative approaches. However, the present study additionally shows that higher interpretability may come at the cost of excessive topic fragmentation. A plausible explanation is that organizational interview data often contain subtle and overlapping narratives, causing the model to generate highly specific document-level labels when prompted independently. In the absence of previously generated labels within the prompt, the model has no mechanism to reuse existing themes, which reduces label consistency and limits the emergence of stable clusters.

For De Hofnar, these results suggest that topic modelling should currently be positioned as an exploratory support tool rather than the primary measurement layer of the pipeline. Its greatest value lies in surfacing company-specific issues that predefined frameworks may miss, or not fully capture. Used correctly, topic modelling can help consultants identify recurring narratives faster and enrich dashboard outputs with organization-specific themes. More broadly, this demonstrates that data-driven methods can complement validated culture frameworks by capturing local context without replacing theoretical structure.

At the same time, caution is required when interpreting these outputs. Topic labels should not automatically be treated as objective organizational truths, causal explanations, or direct recommendations for intervention. Because topic quality remains sensitive to modelling choices and representation format, consultant review remains necessary before communicating findings to clients. More broadly, the results suggest that topic modelling remains the least robust component of the current pipeline. While recurring themes can already be extracted automatically, topic representation requires further development before it can reliably support production-level cultural diagnostics.

6.1.3 Sentiment Analysis: GPT Better Capture Nuanced Sentiment, but at Higher Cost

The most important takeaway from the sentiment analysis experiment is that GPT-based sentiment analysis better captures nuanced sentiment in organizational text than the multilingual transformer baseline, but at substantially higher computational cost. GPT outperforms the transformer-based model across all evaluation metrics, indicating stronger predictive quality in this setting. A plausible explanation is that large language models acquire broad semantic and pragmatic knowledge during large-scale pre-training, enabling them to interpret indirect wording,

contextual cues, mixed signals, and subtle evaluative language in a zero-shot setting. By contrast, encoder-based transformer classifiers often rely more heavily on supervised fine-tuning to adapt general language representations to the specific sentiment expressions of a domain. For this research, the results suggest that broad pre-training can partly compensate for the absence of domain-specific sentiment training data.

The confusion matrices further show that GPT not only improves aggregate performance, but also produces more calibrated ordinal predictions. Prediction errors are more often limited to adjacent sentiment categories rather than extreme misclassifications, which is particularly desirable on ordered scales ranging from strongly negative to strongly positive. This indicates that GPT better preserves sentiment intensity, rather than merely distinguishing positive from negative language. In organizational contexts, this is practically relevant because the difference between neutral feedback, mild dissatisfaction, and severe frustration can materially influence how findings should be interpreted and prioritized by management.

These findings are broadly consistent with prior literature. Zhang et al. (2024) report that LLMs perform strongly in zero-shot and low-resource sentiment classification settings, particularly when labelled training data are limited. The results also align with Shafikuzzaman et al. (2024), who similarly find that LLM-based approaches can achieve strong performance under zero-shot evaluation conditions. Taken together with the present findings, this suggests that GPT is already a strong practical option for sentiment analysis in settings where annotated organizational data are unavailable. At the same time, prior literature indicates that supervised transformer-based approaches remain highly competitive when task-specific training data are available, suggesting that the creation of labelled organizational sentiment data could be a valuable direction for future improvement.

For De Hofnar, these results suggest that sentiment analysis can already be applied effectively in a zero-shot setting using a GPT model, while further gains may be achievable through supervised fine-tuning of transformer-based models on domain-specific data. In practice, sentiment analysis is likely most valuable as a prioritization layer rather than as a stand-alone diagnostic tool. When negative sentiment consistently clusters around themes such as supervision, communication, or workload, this may indicate areas requiring managerial attention. Conversely, strongly positive sentiment around autonomy, teamwork, or leadership may reveal cultural strengths worth preserving. Used in combination with classification and topic modelling, sentiment can therefore help distinguish which recurring themes represent urgent friction points and which reflect healthy organizational dynamics.

At the same time, caution is required when interpreting sentiment outputs. Sentiment scores should not be treated as direct measures of employee wellbeing, engagement, or psychological safety, nor should they be used to evaluate individual managers or teams on the basis of limited text samples. Organizational communication often contains irony, restraint, mixed emotions, or politically cautious language that can complicate automated interpretation. More broadly, sentiment analysis is a useful supporting component of the pipeline, but its strongest value emerges when

interpreted relationally and in combination with richer thematic evidence rather than as an isolated score.

6.2 Dashboard

To illustrate how the extracted signals could be translated into organizational insights, an interactive dashboard was developed that combines OCS dimensions, topic modelling outputs, and sentiment scores within a single analytical interface. This dashboard should be viewed as a design proposal rather than a validated contribution, as comprehensive user evaluation, usability testing, and assessment of its effectiveness for managerial decision-making were beyond the scope of this thesis. The dashboard follows a layered evidence approach in which each component offers a complementary perspective. OCS dimension classification provides an initial structure by organizing textual data into consistent organizational categories, enabling aggregation across interviews. Topic modelling adds interpretive depth by surfacing recurring themes within each category, while sentiment analysis contributes an evaluative layer by estimating the polarity of employee expressions. Combined, these elements may support a more interpretable overview in which structural context, semantic content, and emotional tone are considered jointly.

This concept was translated into five interconnected dashboard components. Screenshots of the implemented dashboard are provided in Appendix D. First, a bubble visualization represents each OCS category according to its relative frequency and average sentiment, offering a rapid overview of potentially relevant areas. Second, an alternative graph-based representation of the same information emphasizes comparative interpretation between categories. Third, a drill-down view for each category displays associated topics, sentiment distributions, and representative text excerpts, linking aggregated indicators to underlying qualitative evidence. Fourth, a temporal sentiment chart tracks changes in average sentiment over time, which may help identify trends or shifts following organizational events or interventions. Finally, a search functionality enables targeted exploration of the dataset by retrieving employee statements related to specific topics, projects, or keywords. Taken together, these components demonstrate one possible way in which NLP outputs could be presented in a more accessible and decision-support oriented format.

6.3 Limitations

This thesis adopts an exploratory approach aimed at reducing the time and effort required for qualitative analysis of OC through NLP. Consequently, the proposed pipeline should be interpreted as a proof of concept rather than a fully optimized or production-ready system. While the results demonstrate the feasibility of extracting meaningful cultural signals from textual data, each component of the pipeline presents opportunities for further refinement. A key limitation of this study lies in the annotation process, where all labelled data was produced by a single annotator. This was a deliberate design decision for two reasons. First, organizational constraints limited the availability of additional trained annotators. Second, the interview data contain highly sensitive information, making outsourcing annotation infeasible. However, this does introduce the risk of subjective bias and constrains the reliability of the ground truth used for evaluation. Future work should therefore incorporate multiple annotators, enabling the computation of inter-annotator agreement metrics such as Cohen's kappa or Krippendorff's alpha. This would allow for a more

rigorous assessment of label consistency and support aggregation strategies such as consensus labelling or probabilistic label modelling, thereby improving the robustness and credibility of the evaluation framework.

A second limitation concerns the model exploration procedure. While multiple models were evaluated, the search process over architectures and hyperparameters is a simple grid search algorithm. Future work could benefit from more systematic and efficient search strategies across more models and a broader configuration space. Methods that could be explored are *Hyperband optimization*, which enables resource-efficient exploration by combining early stopping with adaptive allocation to promising configurations, and *Bayesian optimization*, which provides a principled approach to navigating the hyperparameter space based on prior evaluations. Additionally, evolutionary algorithms and neural architecture search (NAS) offer the potential to automatically discover high-performing model structures beyond manually defined configurations. Expanding the search space to include a wider variety of model families, including both classical machine learning methods and transformer-based architectures, would further strengthen the comparative analysis.

Another important area for improvement concerns the granularity and structural consistency of the textual inputs to the NLP pipeline. In the current pipeline, documents exhibit substantial differences in format and length, ranging from detailed narratives to short phrases or bullet points. Furthermore, individual documents often contain multiple topics, which dilutes the clarity of extracted signals and complicates downstream analysis. A promising direction is to introduce topic-level segmentation prior to modelling. LLMs could be used to decompose responses into coherent thematic units, for example by restructuring input into paragraphs where each paragraph represents a single topic. This would standardize input representations, reduce noise caused by stylistic variation, and improve the signal-to-noise ratio for both classification and topic modelling tasks. In addition, incorporating aspect-based or feature-level sentiment analysis would allow for more precise attribution of sentiment to specific themes, enhancing both interpretability and diagnostic value.

Finally, the current approach is limited to a single data source, namely employee interviews. While this provides rich qualitative insight, it restricts the scope of analysis to a single perspective within the organization. An important extension would be to integrate multiple textual data sources, such as management communications, internal policy documents, and external platforms like employee reviews. This would enable comparative analyses between intended culture, as communicated by leadership, and perceived culture, as experienced by employees. Detecting and quantifying discrepancies between these perspectives would offer a more comprehensive view of organizational dynamics and provide stronger indicators of cultural misalignment and friction. Such multi-source analysis would further enhance the practical applicability of the proposed framework as a tool for organizational diagnosis and intervention.

Chapter 7 Conclusion

This study set out to examine how NLP can be leveraged to quantify OC from textual data and translate these measurements into actionable insights. The first sub-question concerned which

cultural model is best suited to measure OC. The Organizational Culture Survey framework proposed by Glaser et al. (1987) is identified as the most suitable conceptual foundation. Its dimensional structure provides a well-established and interpretable lens through which cultural signals can be operationalised, while remaining sufficiently flexible to accommodate data-driven extensions.

The second sub-question addressed which NLP methods are most effective for detecting signals of culture in employee interviews. The findings indicate that three complementary NLP components are particularly effective for capturing cultural signals in employee-generated text: classification, topic modelling, and sentiment analysis. Classification enables the structured mapping of text to predefined cultural dimensions, ensuring alignment with theoretical constructs. Topic modelling adds semantic granularity by uncovering recurring themes within these dimensions, while sentiment analysis captures the evaluative tone associated with these themes. In combination, these methods form a coherent analytical pipeline that balances theoretical grounding with empirical sensitivity to organization-specific dynamics.

The third sub-question addressed how traditional language models and LLMs compare in terms of performance for OC analysis. The comparative experiments reveal a clear trade-off between LLMs and smaller, task-specific transformer-based models. In this study, the evaluated GPT-based approach performed better in the low-resource sentiment and topic-labelling settings. In contrast, when sufficient annotated data is present, transformer-based models such as BERT variants achieve competitive and in some cases superior performance, benefiting from task-specific fine-tuning compared to the few-shot setting LLM. This pattern aligns with prior empirical findings (Shafikuzzaman et al., 2024) and suggests that model selection should be guided by data availability and task requirements rather than a one-size-fits-all preference for model scale.

The fourth sub-question addressed how model outputs can be aggregated into interpretable and actionable organizational insights. To address this, a structured dashboard design is proposed, which integrates and presents model outputs in a coherent and accessible manner. By organizing results in a layered design, the system allows users to move from high-level cultural dimensions to underlying topics and associated sentiment patterns. This hierarchical approach supports both exploratory analysis and targeted diagnosis, enabling stakeholders to identify not only where cultural friction occurs, or where cultural strengths are concentrated, but also what specific issues drive it and how they are experienced by employees.

Overall, this thesis demonstrates that NLP can provide a scalable and theoretically grounded approach to analysing organizational culture from qualitative employee data. By combining validated cultural dimensions with data-driven theme discovery and sentiment analysis, the proposed pipeline offers both a methodological contribution to OC research and a practical foundation for software-supported cultural diagnostics in organizational settings.

7.1 Scientific Contribution

This thesis makes three primary contributions to the study of OC using NLP. First, it provides a structured and empirically grounded pipeline for extracting cultural signals from qualitative employee text data. By operationalizing the OCS framework as a multi-label classification task, the study translates abstract cultural dimensions into measurable constructs that can be detected automatically in interview data. This contributes a way to translate abstract culture theory into computable constructs, which helps bridge the gap between qualitative OC research and scalable automated analysis.

Second, the thesis presents a systematic comparison between transformer-based models and LLMs across three complementary analytical tasks: classification, topic modelling, and sentiment analysis. The results provide empirical insight into when LLMs are effective and when fine-tuned transformer-based models remain preferable, thereby contributing to the growing literature on model selection trade-offs in applied NLP.

Third, the study proposes an integrated analytical framework in which classification, topic modelling, and sentiment analysis are combined into a coherent pipeline for OC analysis. Methodologically, this framework brings together theory-driven and data-driven components within a single interpretive structure. Classification operationalizes established cultural dimensions, topic modelling captures emergent and organization-specific themes, and sentiment analysis adds an evaluative layer that reflects how these themes are experienced by employees. In doing so, the framework enables the simultaneous analysis of *where* cultural patterns occur, *what* is being discussed, and *how* these issues are perceived. More broadly, this contribution helps bridge the methodological gap (Hansen and Świderska, 2024; Druckman et al., 1997; Jung et al., 2009) between scalable but shallow measurement approaches, and richer qualitative approaches that are often difficult to standardize and scale.

7.2 Practical Relevance

In addition to its scientific contribution, this research has clear practical relevance. First, the proposed NLP pipeline offers a strong value proposition in terms of efficiency, scalability, and consistency. De Hofnar previously required approximately one week of full-time equivalent manual work to analyse interview data and derive insights for a single project, the proposed pipeline can process and synthesise the interview data of the three participating companies into a dashboard in approximately 30 minutes. This substantially reduces the time and labour required for cultural diagnostics, while also standardising the analysis process across projects. Unlike manual interpretation, which may vary between analysts or over time, the pipeline applies the same analytical procedure to every text segment, for every client, in a consistent and reproducible manner. This consistency further supports longitudinal use, as results can be compared over time to assess whether interventions have led to measurable improvements in areas such as morale, communication, or supervision.

Second, the pipeline is grounded in a theoretically and empirically validated cultural framework. The use of the OCS model provides a structured and academically established backbone for the

analysis, ensuring that the classification of cultural signals is aligned with recognized dimensions of organizational culture. This grounding enhances the validity of the overall approach, particularly in relation to the more data-driven components such as topic modelling. While topic modelling enables the discovery of emergent themes in an unsupervised manner, its outputs can be inherently variable and harder to interpret in isolation. By anchoring these findings within the OCS framework, the pipeline combines exploratory insights with theoretically grounded structure, resulting in outputs that are both data-driven and methodologically robust.

Finally, this thesis demonstrates that automated cultural diagnostics are technically feasible and already capable of producing structured and interpretable insights from qualitative employee data. At the same time, the results identify specific components, most notably topic representation, that require further refinement before deployment in a fully production-ready setting. The practical contribution of this work therefore lies not only in the prototype itself, but also in showing how such a system can be designed, where its current strengths lie, and which elements offer the greatest potential for future development.

Appendix

Chapter A Annotation Guidelines Classification

Your task is to assign one or more culture labels to each text segment. A segment may reflect multiple cultural dimensions. Read the text carefully and select every dimension that is meaningfully present. Below are the six dimensions you can choose from. Each description explains what the dimension means in simple terms, what kinds of behaviours or signals belong to it, and what to look for in the text.

1. Teamwork and Conflict This dimension describes how well people work together and how they handle disagreements. Select this label when a text describes the way things work, cooperation, helping each other, working as a team, or openly talking about problems. It also applies when the text mentions tension, mistrust, power struggles, or difficulty confronting issues. Anything about being honest with colleagues, giving feedback, solving conflicts, or lacking these things belongs here.

2. Climate and Morale This dimension reflects how people feel about the general atmosphere at work. Choose this label when the text talks about motivation, trust, fairness, respect, positivity, burnout, or low morale. It includes comments about feeling valued, feeling supported, or (the opposite) feeling unappreciated or discouraged. This dimension captures the emotional "vibe" of the organization.

3. Information Flow This dimension is about how well information moves through the organization. Select this label when the text mentions communication clarity, being informed (or not), transparency, reasons behind decisions, or receiving updates. It also applies when people say they don't know what's happening, feel out of the loop, or struggle to get the information needed to do their job. This is about communication channels, not emotions.

4. Involvement This dimension captures how much employees feel included in decision-making and whether their ideas matter. Choose this label when the text mentions giving input, being listened to, being asked for suggestions, or having influence over one's work. It also applies when the text describes the opposite, like feeling ignored, excluded, or not taken seriously. This is about participation and ownership.

5. Supervision This dimension focuses on the management, manager or supervisor. Select this label when the text talks about feedback, guidance, clarity of expectations, listening, taking criticism, or delegating work. Positive or negative references to one's supervisor, manager or management here.

6. Meetings This dimension refers to how meetings are run and whether they are effective. Choose this label when a text mentions the quality of meetings, sticking to the agenda, participation, creativity in discussions, or whether decisions actually lead to action. Complaints about meetings being unproductive, or praise about meetings being useful, both fit this category.

7. Other This category is mutually exclusive and should only be used if the text clearly reflects a cultural aspect that does not fit any of the six predefined dimensions. Use it sparingly. Also use this dimension if it does not talk about the current organization.

General Annotation Rules
Multi-label: Assign all labels that apply. A text can fit multiple dimensions.
Content over wording: Focus on the meaning, not exact keywords.
Local context: Use only the information in the given text segment, not assumptions.
Neutrality: Do not judge the organization; only select relevant dimensions.

Chapter B OCS Operational Definitions and Questionnaire Items

This appendix presents the full OCS (Glaser et al., 1987) framework used in this thesis. It contains the original operational definitions of the six culture dimensions, followed by the corresponding questionnaire items used to measure each dimension.

Teamwork and Conflict

Reported coordination of effort, interpersonal cooperation, rapport or antagonism, resentment, jealousy, mistrust, power struggle within sections or divisions; people talk directly and candidly about problems they have with each other.

1. People I work with are direct and honest with each other.
2. People I work with accept criticism without becoming defensive.
3. People I work with function as a team.
4. People I work with constructively confront problems.
5. People I work with are good listeners.
6. Labor and management have a productive working relationship.

Climate and Morale

Reported feelings about work conditions, motivation, general atmosphere, organizational character.

7. This organization motivates me to put out my best efforts.
8. This organization respects its workers.
9. This organization treats people in a consistent and fair manner.
10. There is an atmosphere of trust in this organization.
11. This organization motivates people to be efficient and productive.

Information Flow

Links, channels, contact, flow of communication to pertinent people or groups in the organization; reported feelings of isolation or being out of touch.

12. I get enough information to understand the big picture here.
13. When changes are made, the reasons why are made clear.
14. I know what is happening in work sections outside of my own.
15. I get the information I need to do my job well.

Involvement

Reported input and participation in decision making; respondents feel that their thoughts and ideas count and are encouraged by top management to offer opinions and suggestions.

16. I have a say in decisions that affect my work.
17. I am asked to make suggestions about how to do my job better.
18. This organization values the ideas of workers at every level.
19. My opinions count in this organization.

Supervision

Reported information by employees on their immediate supervisor; the extent to which they are given positive and negative feedback on work performance; the extent to which job expectations are clear.

20. Job requirements are made clear by my supervisor.
21. When I do a good job my supervisor tells me.
22. My supervisor takes criticism well.
23. My supervisor delegates responsibility.
24. My supervisor gives me criticism in a positive manner.
25. My supervisor is a good listener.
26. My supervisor tells me how I am doing.

Meetings

Reported information on whether meetings occur and how productive they are.

27. Decisions made at meetings get put into action.
28. Everyone takes part in discussions at meetings.
29. Our discussions in meetings stay on track.
30. Time in meetings is time well spent.
31. Meetings tap the creative potential of people present.

Chapter C Topic Model Configurations

This appendix provides the hyperparameter configurations corresponding to the topic model identifiers used in Table 11 and 12.

BERTopic Configurations

- **BERTopic**$_1$: UMAP ($n\_neighbors = 5, n\_components = 10$), HDBSCAN ($min\_cluster\_size = 5$, $min\_samples = 1$, $cluster\_selection\_method =$ leaf)
- **BERTopic**$_2$: UMAP ($n\_neighbors = 5, n\_components = 5$), HDBSCAN ($min\_cluster\_size = 5$, $min\_samples = 1$, $cluster\_selection\_method =$ eom)
- **BERTopic**$_3$: UMAP ($n\_neighbors = 5, n\_components = 5$), HDBSCAN ($min\_cluster\_size = 3$, $min\_samples = 1$, $cluster\_selection\_method =$ eom)

LDA Configurations

- **LDA**$_1$: $k = 100$, $\alpha =$ None, $\beta = 1.0$
- **LDA**$_2$: $k = 150$, $\alpha =$ None, $\beta = 1.0$
- **LDA**$_3$: $k = 50$, $\alpha = 0.01$, $\beta = 1.0$

GPT Configurations

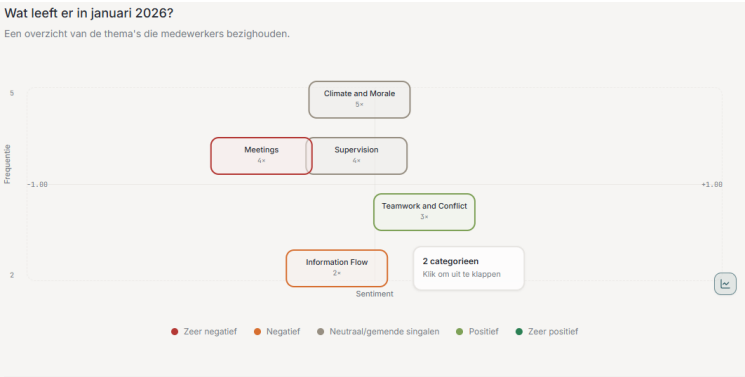
- **GPT**$_1$: GPT-5-mini, few-shot prompting ($zero\_shot = False$), list of used topics ($topic\_list = False$)
- **GPT**$_2$: GPT-5-mini, zero-shot prompting ($zero\_shot = True$), list of used topics ($topic\_list = False$)
- **GPT**$_3$: GPT-5-mini, few-shot prompting ($zero\_shot = False$), list of used topics ($topic\_list = True$)
- **GPT**$_4$: GPT-5-mini, zero-shot prompting ($zero\_shot = True$), list of used topics ($topic\_list = True$)

Chapter D Dashboard

This appendix presents the five dashboard components described in Section 6.2. Two points should be noted when interpreting the figures. First, the dashboard interface is developed in Dutch, reflecting the language of the original source data. Second, all data displayed in the figures are artificially generated using ChatGPT to protect the privacy of participating organizations. Consequently, the example topics and representative quotes might not fit together logically, and are presented in English rather than Dutch.

Each bubble represents one Organizational Culture Survey (OCS) category. Bubble size reflects relative frequency, while colour and position indicate average sentiment. This offers a rapid overview of prominent cultural themes and areas that may warrant further attention.

Figure 10: Bubble overview of OCS categories by prevalence and sentiment.



This alternative representation visualizes the same category-level information in a graph-based format. It is intended to support comparison between dimensions by emphasizing relative differences in prevalence and sentiment outcomes.

Figure 11: Comparative graph of OCS dimensions and sentiment.

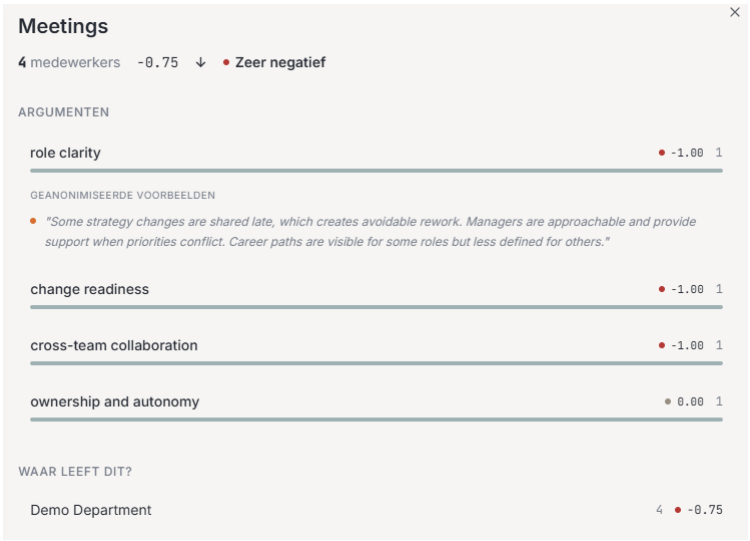


Figure 12: Category drill-down with topics, sentiment, and example statements.

Selecting a category opens a detailed view containing associated topics, sentiment distributions, and representative employee statements. This links aggregated indicators to the qualitative evidence from which they were derived.

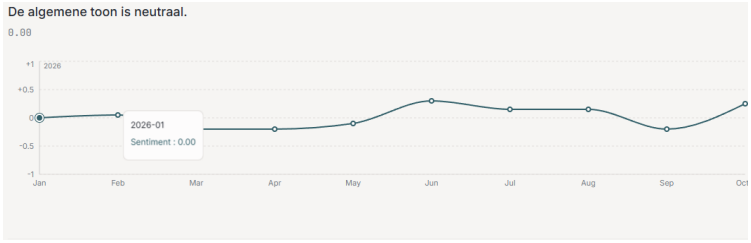


Figure 13: Temporal sentiment development across measurement periods.

This chart tracks average sentiment over time. Monitoring changes may help identify cultural shifts, assess the effects of interventions, or detect periods associated with organizational tension or improvement.

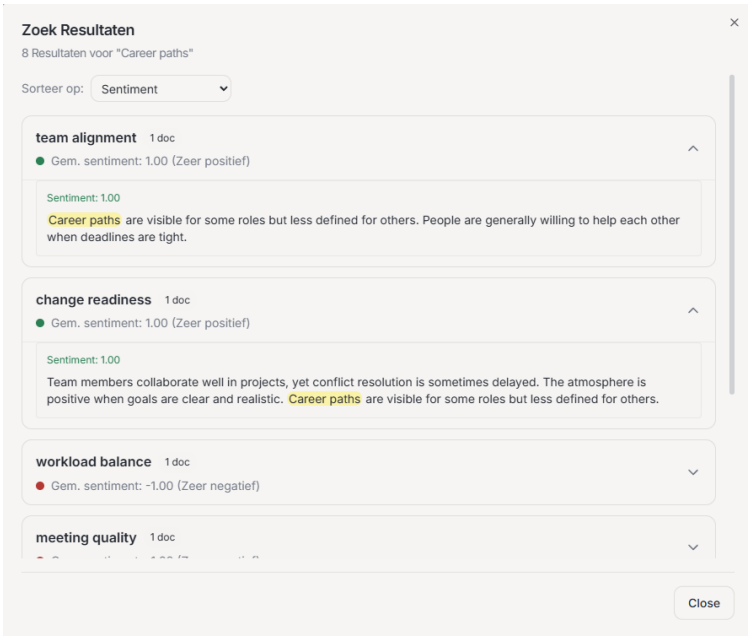


Figure 14: Keyword search interface for targeted statement retrieval.

The search functionality enables targeted exploration of the dataset by retrieving employee statements related to specific topics, projects, or keywords. This may support deeper investigation of concrete organizational issues.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2022). Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Aggarwal, S. (2024). Impact of dimensions of organisational culture on employee satisfaction and performance level in select organisations. *IIMB Management Review*, 36(3):230–238.

Akhavan, P., Ebrahim Sanjaghi, M., Rezaeenour, J., and Ojaghi, H. (2014). Examining the relationships between organizational culture, knowledge management and environmental responsiveness capability. *VINE*, 44(2):228–248.

Aldunate, A., Maldonado, S., Vairetti, C., and Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *Expert Systems with Applications*, 209:118309.

Bashiri, H. and Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowledge and Information Systems*, 66(12):7305–7361.

Bellot, J. (2011). Defining and Assessing Organizational Culture. *Nursing Forum*, 46(1):29–37.

Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:601–608.

Cai, M. Y., Lin, Y., and Zhang, W. J. (2016). Study of the optimal number of rating bars in the likert scale. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, pages 193–198, Singapore Singapore. ACM.

Cammel, S. A., De Vos, M. S., van Soest, D., Hettne, K. M., Boer, F., Steyerberg, E. W., and Boosman, H. (2020). How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC medical informatics and decision making*, 20(1):97.

Chae, Y. and Davidson, T. (2026). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. *Sociological Methods & Research*, 55(2):501–567.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. Version Number: 2.

Cooke, R. A. and Lafferty, J. (1983). Level V: Organizational cultureal invenory - form I.

Cooke, R. A. and Szumal, J. L. (1993). Measuring Normative Beliefs and Shared Behavioral Expectations in Organizations: The Reliability and Validity of the Organizational Culture Inventory. *Psychological Reports*, 72(3_suppl):1299–1330.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT Model. Version Number: 1.

Denison, D., Nieminen, L., and Kotrba, L. (2014). Diagnosing organizational cultures: A conceptual and empirical review of culture effectiveness surveys. *European Journal of Work and Organizational Psychology*, 23(1):145–161.

Denison, D. R. and Mishra, A. K. (1995). Toward a Theory of Organizational Culture and Effectiveness. *Organization Science*, 6(2):204–223.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Version Number: 2.

Druckman, D., Cott, H. V., and Singer, J. E., editors (1997). *Enhancing Organizational Performance*. National Academies Press, Washington, D.C. Pages: 5128.

Evans, M. S. (2014). A Computational Approach to Qualitative Analysis in Large Textual Datasets. *PLoS ONE*, 9(2):e87908.

Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. *Information*, 13(2):83.

Ginting, H., Fahmi, T. A., Tjakraatmadja, J. H., and Adzhani, I. A. (2023). Developing Interpretive Text Analysis for Corporate Culture Measurement. *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–6.

Glaser, S. R., Zamanou, S., and Hacker, K. (1987). Measuring and Interpreting Organizational Culture. *Management Communication Quarterly*, 1(2):173–198.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Version Number: 1.

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., and Min, H. (2024). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal*, 34(1):39–54.

Gupta, M., Gupta, A., and Cousins, K. (2022). Toward the understanding of the constituents of organizational culture: The embedded topic modeling analysis of publicly available employee-generated reviews of two major U.S.-based retailers. *Production and Operations Management*, 31(10):3668–3686.

Hansen, K. and Świderska, A. (2024). Integrating open- and closed-ended questions on attitudes towards outgroups with different methods of text analysis. *Behavior Research Methods*, 56(5):4802–4822.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., and Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, 4(1):1.

Hofstede, G., Neuijen, B., Ohayv, D. D., and Sanders, G. (1990). Measuring Organizational Cultures: A Qualitative and Quantitative Study Across Twenty Cases. *Administrative Science Quarterly*, 35(2):286.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., and Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Huang, X., Zhang, L. L., Cheng, K.-T., Yang, F., and Yang, M. (2023). Fewer is More: Boosting LLM Reasoning with Reinforced Context Pruning. Version Number: 3.

Hunt, G., Moloney, M., and Fazio, A. (2011). Embarking on large-scale qualitative research: reaping the benefits of mixed methods in studying youth, clubs and drugs. *Nordisk alkohol- & narkotikatidskrift: NAT*, 28(5-6):433–452.

Jung, T., Scott, T., Davies, H. T. O., Bower, P., Whalley, D., McNally, R., and Mannion, R. (2009). Instruments for Exploring Organizational Culture: A Review of the Literature. *Public Administration Review*, 69(6):1087–1096.

Koch, S. and Pasch, S. (2023). Culturebert: Measuring corporate culture with transformer-based language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3176–3184. IEEE.

Li, K., Mai, F., Shen, R., and Yan, X. (2018). Measuring Corporate Culture Using Machine Learning. *SSRN Electronic Journal*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Version Number: 1.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Version Number: 3.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2022). Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, 54(3):1–40.

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., and Elazar, Y. (2023). Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. Version Number: 2.

NLP Town (2020). bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. Hugging Face model repository.

O'Reilly, C. A., Chatman, J., and Caldwell, D. F. (1991). PEOPLE AND ORGANIZATIONAL CULTURE: A PROFILE COMPARISON APPROACH TO ASSESSING PERSON-ORGANIZATION FIT. *Academy of Management Journal*, 34(3):487–516.

Parker, M. J., Anderson, C., Stone, C., and Oh, Y. (2025). A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education*, 35(2):444–481.

Peters, T. and Waterman, R. H. (1982). *In search of excellence: lessons from America's best-run companies*. Harper and Row, New York ; Cambridge ; Mexico City.

Pham, C., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. (2024). TopicGPT: A Prompt-based Topic Modeling Framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Pillai, I., Fumera, G., and Roli, F. (2013). Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065.

Quinn, R. E. and Rohrbaugh, J. (1981). A Competing Values Approach to Organizational Effectiveness. *Public Productivity Review*, 5(2):122.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Schachner, M., Ardag, M. M., Holtz, P., Großer, J., Hartz, C., Van Herk, H., Bender, M., Boehnke, K., and Dobewall, H. (2024). Extracting organizational culture from text: the development and validation of a theory-driven tool for digital data. *European Journal of Work and Organizational Psychology*, 33(5):571–582.

Schein, E. H. (1984). Coming to a New Awareness of organizational culture. *Sloan Management Review*, 25:3–16.

Shafikuzzaman, M., Islam, M. R., Rolli, A. C., Akhter, S., and Seliya, N. (2024). An Empirical Evaluation of the Zero-Shot, Few-Shot, and Traditional Fine-Tuning Based Pretrained Language Models for Sentiment Analysis in Software Engineering. *IEEE Access*, 12:109714–109734.

Sull, D., Sull, C., and Chamberlain, A. (2019). Measuring Culture in Leading Companies.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., and Wang, G. (2023). Text Classification via Large Language Models. Version Number: 3.

Tadesse Bogale, A. and Debela, K. L. (2024). Organizational culture: a systematic review. *Cogent Business & Management*, 11(1):2340129.

Van Der Post, W. Z., De Coning, T. J., and Smit, E. V. (1997). An instrument to measure organizational culture. *South African Journal of Business Management*, 28(4):147–168.

Wang, H., Prakash, N., Hoang, N. K., Hee, M. S., Naseem, U., and Lee, R. K.-W. (2023). Prompting Large Language Models for Topic Modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241, Sorrento, Italy. IEEE.

Wu, X., Nguyen, T., and Luu, A. T. (2024). A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Yilmaz, C. and Ergun, E. (2008). Organizational culture and firm effectiveness: An examination of relative effects of culture traits and the balanced culture hypothesis in an emerging economy. *Journal of World Business*, 43(3):290–306.

Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. (2024). Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.