

Article

Computational Modelling of Institutional Perceptions in a Civilisational Intergovernmental Organisation

Fahim Sufi ^{1,2,*} , A. K. M. Iftexharul Islam ³ and Anowara Akter ⁴¹ COEUS Institute, New Market, VA 22844, USA² Centre for Trade & Investment, University of Dhaka, Dhaka 1000, Bangladesh³ Faculty of Business, Law and Politics, University of Hull, Bloomsbury, London WC1E 6AA, UK; a.i.islam-2020@hull.ac.uk⁴ Department of Islamic History and Culture, Jagannath University, Dhaka 1100, Bangladesh; dr.anowara@ihc.jnu.ac.bd

* Correspondence: fahim.cti@du.ac.bd or research@fahimsufi.com

Abstract

Civilisational intergovernmental organisations occupy a distinctive position in global governance because their legitimacy is shaped by institutional performance, symbolic identity, collective representation, and normative expectation. This study develops a sentiment-driven computational modelling framework for analysing institutional perceptions in a 57-member civilisational intergovernmental organisation. The empirical corpus comprises 30 semi-structured interviews, including 15 expert interviews and 15 general respondent interviews. After preprocessing, removal of administrative material, and exclusion of courtesy-only responses, the final corpus contained 239 answer segments and 42,940 respondent-generated words. The methodology integrates answer-level segmentation, thematic classification, lexical sentiment analysis, transformer-based sentiment modelling, latent semantic clustering, non-parametric statistical testing, bootstrap confidence estimation, and manual validation. The results show that institutional perception is broadly positive but thematically uneven. Experts recorded a mean polarity score of 0.0803, while general respondents recorded 0.1016. However, group-level differences were not statistically significant at segment level, Mann–Whitney $U = 6171.50$, $p = 0.0704$, or interview level, $U = 71.00$, $p = 0.0888$. In contrast, sentiment differed significantly across institutional themes, Kruskal–Wallis $H = 45.8768$, $p < 0.001$, and across latent semantic clusters, with Kruskal–Wallis $H = 35.0124$ and $p < 0.001$. The strongest positive sentiment appeared in reform and future orientation, mean polarity = 0.1700, socioeconomic cooperation, 0.1330, and institutional effectiveness, 0.1325. The weakest sentiment concerned institutional weakness and constraints, 0.0546, and political and security role, 0.0693. Latent semantic clustering identified socioeconomic and economic cooperation as the most positively evaluated cluster, with mean polarity 0.1465. Overall, the findings reveal a measurable tension between symbolic-developmental legitimacy and operational scepticism.



Academic Editors: Roman Makarov and Zhiyong (Johnny) Zhang

Received: 30 May 2026

Revised: 9 July 2026

Accepted: 18 July 2026

Published: 21 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: computational social science; sentiment analysis; institutional perceptions; civilisational intergovernmental organisation; thematic modelling; latent semantic clustering; transformer based sentiment analysis; symbolic legitimacy; operational scepticism; cultural regionalism

1. Introduction

Civilisational intergovernmental organisations present a distinctive analytical problem because their perceived legitimacy is not produced solely through institutional performance, policy delivery, or formal procedural design. It is also shaped by symbolic identity, collective representation, normative expectation, and cultural affiliation. Existing scholarship on international organisational legitimacy has shown that performance, representation, procedure, and public confidence are central to how international organisations are evaluated [1–4]. At the same time, computational social science has developed mature methods for modelling sentiment, polarity, and latent thematic structure from text [5–8]. Lexicon-based models offer transparent sentiment measurement [9], transformer-based models such as BERT and RoBERTa capture contextual semantic dependencies [10,11], and topic modelling enables the extraction of latent discourse clusters [7,12], while latent semantic clustering more precisely describes TF-IDF and latent semantic analysis representations grouped through clustering rather than probabilistic topic inference. Yet, limited work has connected these computational methods with the study of institutional perception in civilisational intergovernmental organisations using interview-based evidence. This creates a clear research gap: legitimacy studies remain conceptually rich but often computationally limited, while sentiment-driven modelling studies are technically advanced but insufficiently grounded in institutional theory.

In this study, a civilisational intergovernmental organisation is understood as a formal multilateral organisation whose legitimacy is shaped not only by state-based cooperation and institutional performance, but also by shared civilisational identity, cultural affiliation, collective representation, and normative expectation. Analysing sentiment toward such an organisation is scientifically relevant because its legitimacy cannot be assessed only through formal mandates, policy outputs, or organisational structure. Perceptions of trust, symbolic representation, operational credibility, reform expectation, and institutional scepticism are expressed through evaluative language. Sentiment analysis is therefore not used merely because the technique is available; it is used because it provides a systematic way to measure how respondents linguistically evaluate different institutional functions. The study is guided by three analytical expectations: first, perceptions will vary more strongly across institutional functions than across respondent groups; second, socioeconomic, developmental, and reform-oriented themes will attract more positive sentiment than political-security and constraint-oriented themes; and third, expert respondents may express more cautious evaluations than general respondents because of their greater familiarity with institutional limitations.

To address this gap, this study develops a sentiment-driven computational modelling framework for analysing institutional perceptions in a 57-member civilisational intergovernmental organisation. The empirical corpus consists of 30 semi-structured interviews, including 15 expert interviews and 15 general respondent interviews. After preprocessing, removal of administrative material, and exclusion of courtesy-only responses, the final analytical corpus contained 239 answer segments and 42,940 respondent-generated words. The expert corpus contributed 121 segments and 24,333 words, whereas the general respondent corpus contributed 118 segments and 18,607 words. The methodological innovation lies in integrating answer-level segmentation, similarity-based thematic classification, lexical sentiment analysis, transformer-based contextual sentiment modelling, latent semantic clustering, non-parametric statistical testing, bootstrap confidence estimation, and manual validation. This design allows institutional perception to be treated not as an impressionistic judgement, but as a multidimensional computational construct shaped by polarity, subjectivity, thematic membership, latent semantic cluster membership, respondent category, and evaluative intensity.

The results indicate that institutional perception is broadly positive but thematically uneven. General respondents recorded a higher mean polarity score, 0.1016, than expert respondents, 0.0803; however, this difference was not statistically significant at either segment level, Mann–Whitney $U = 6171.50$, $p = 0.0704$, or interview level, $U = 71.00$, $p = 0.0888$. The sentiment label distribution also showed no significant group-level association, $\chi^2 = 3.0990$, $p = 0.2124$, Cramer’s $V = 0.1139$. In contrast, sentiment differed significantly across institutional themes, Kruskal–Wallis $H = 45.8768$, $p < 0.001$, and across latent semantic clusters, with Kruskal–Wallis $H = 35.0124$ and $p < 0.001$. Reform and future orientation generated the highest mean polarity score, 0.1700, followed by socioeconomic cooperation, 0.1330, and institutional effectiveness, 0.1325. The weakest sentiment was associated with institutional weakness and constraints, 0.0546, and political and security role, 0.0693. Latent semantic clustering further identified socioeconomic and economic cooperation as the most positively evaluated semantic cluster, with a mean polarity of 0.1465.

The significance of this study lies in demonstrating that perceptions of a civilisational intergovernmental organisation are better explained by the institutional function under evaluation than by respondent category alone. The findings reveal a measurable perception architecture in which symbolic-developmental legitimacy coexists with operational scepticism. Respondents express comparatively stronger positive sentiment toward reform potential, socioeconomic cooperation, and institutional effectiveness, while displaying more restrained evaluations of political-security functions, visibility gaps, and institutional constraints. The study therefore makes four principal contributions:

- It develops a reproducible computational framework for modelling institutional perceptions from interview narratives.
- It quantifies institutional perception using 239 answer segments and 42,940 respondent-generated words.
- It demonstrates that thematic and cluster-level differences are statistically stronger than expert–general group differences.
- It provides a transferable computational approach for studying legitimacy, trust, perceived effectiveness, and scepticism in international, regional, and identity-based multilateral organisations.

2. Background

Computational analysis of textual data has become an established approach for studying attitudes, perceptions, and evaluative judgments in social, political, and organisational contexts. Sentiment analysis provides one of the most widely used foundations for this task because it converts subjective language into measurable polarity and affective orientation [5,6]. Lexicon-based models, such as VADER, remain useful because they provide transparent and interpretable baseline measures of textual sentiment [9]. However, institutional discourse is often indirect, cautious, and context-dependent. For this reason, transformer-based language models, including BERT and RoBERTa, have become increasingly important because they capture contextual semantic relationships that are difficult to detect through dictionary-based methods alone [10,11].

A second stream of relevant work concerns topic modelling, semantic clustering, and automated text analysis. Early probabilistic models, particularly latent Dirichlet allocation, established topic modelling as a method for extracting latent thematic structures from large textual corpora [7]. More recent embedding-based approaches, such as BERTopic, extend this logic by combining semantic representation with interpretable topic extraction [12]. The present study does not treat its clusters as probabilistic topics; instead, it uses TF-IDF representations projected through latent semantic analysis and grouped through

k-means clustering as interpretable latent semantic clusters. In political and policy research, automated text analysis is increasingly used not as a replacement for qualitative interpretation, but as a means of scaling and formalising it [8,13]. This point is particularly relevant for interview-based research, where the central challenge is to preserve interpretive nuance while producing reproducible analytical outputs.

A third body of literature examines the legitimacy and perception of international organisations. Existing studies show that institutional legitimacy is shaped by performance, procedural fairness, representation, and public confidence [1–4]. These works are important because they indicate that perceptions of international organisations cannot be reduced to formal effectiveness alone. They are also shaped by identity, symbolic authority, and expectations regarding institutional behaviour. For civilisational or identity-based organisations, this issue becomes even more pronounced, since legitimacy may derive from collective representation as much as from policy implementation.

The present study also builds on prior computational work, where sentiment analysis, entity extraction, regression modelling, GPT-integrated classification, and media bias quantification were used to analyse large-scale textual corpora [14,15]. These studies provide a methodological precedent for treating textual narratives as structured computational evidence. The present paper extends that logic from media and cyber intelligence corpora to interview-based institutional perception modelling. Table 1 contextualizes this study with existing studies.

Table 1. Summary of relevant research streams and their relevance to the present study.

Research Stream	Representative Studies	Relevance to the Present Study
Sentiment analysis	Pang and Lee [5]; Liu [6]; Hutto and Gilbert [9]	Provides the foundation for converting subjective interview narratives into measurable polarity, neutrality, and evaluative intensity.
Transformer-based language modelling	Devlin et al. [10]; Liu et al. [11]; Barbieri et al. [16]	Supports context-sensitive sentiment modelling where political and institutional evaluations may be indirect, cautious, or mixed.
Topic modelling, latent semantic clustering, and computational text analysis	Blei et al. [7]; Grootendorst [12]; Grimmer and Stewart [8]; Isoaho et al. [13]	Justifies the extraction of latent themes and semantic clusters from qualitative textual data while retaining interpretability.
Institutional legitimacy and perception	Tallberg and Zürn [1]; Dellmuth and Tallberg [2]; Dellmuth et al. [3]; Steffek [4]	Establishes that perceptions of international organisations are shaped by performance, representation, confidence, and legitimacy.

Despite these advances, limited work has combined sentiment analysis, transformer-based modelling, latent semantic clustering, statistical testing, and manual validation to study perceptions of civilisational intergovernmental organisations using interview data. The present study addresses this gap by proposing a computational thematic sentiment modelling framework that transforms qualitative institutional narratives into measurable perception signals. In doing so, it connects computational social science with the study of international organisations, offering a replicable approach for analysing symbolic legitimacy, institutional effectiveness, and operational scepticism.

3. Methodology

This study adopts a computational thematic sentiment modelling framework to quantify institutional perceptions expressed in semi-structured interview narratives. The schematic of the conceptual architecture is summarised in Figure 1. The framework begins with raw interview transcripts, proceeds through corpus construction, segmentation, thematic classification, sentiment modelling, latent semantic clustering, statistical comparison, robustness assessment, and manual validation, and culminates in an interpretive institutional perception synthesis.

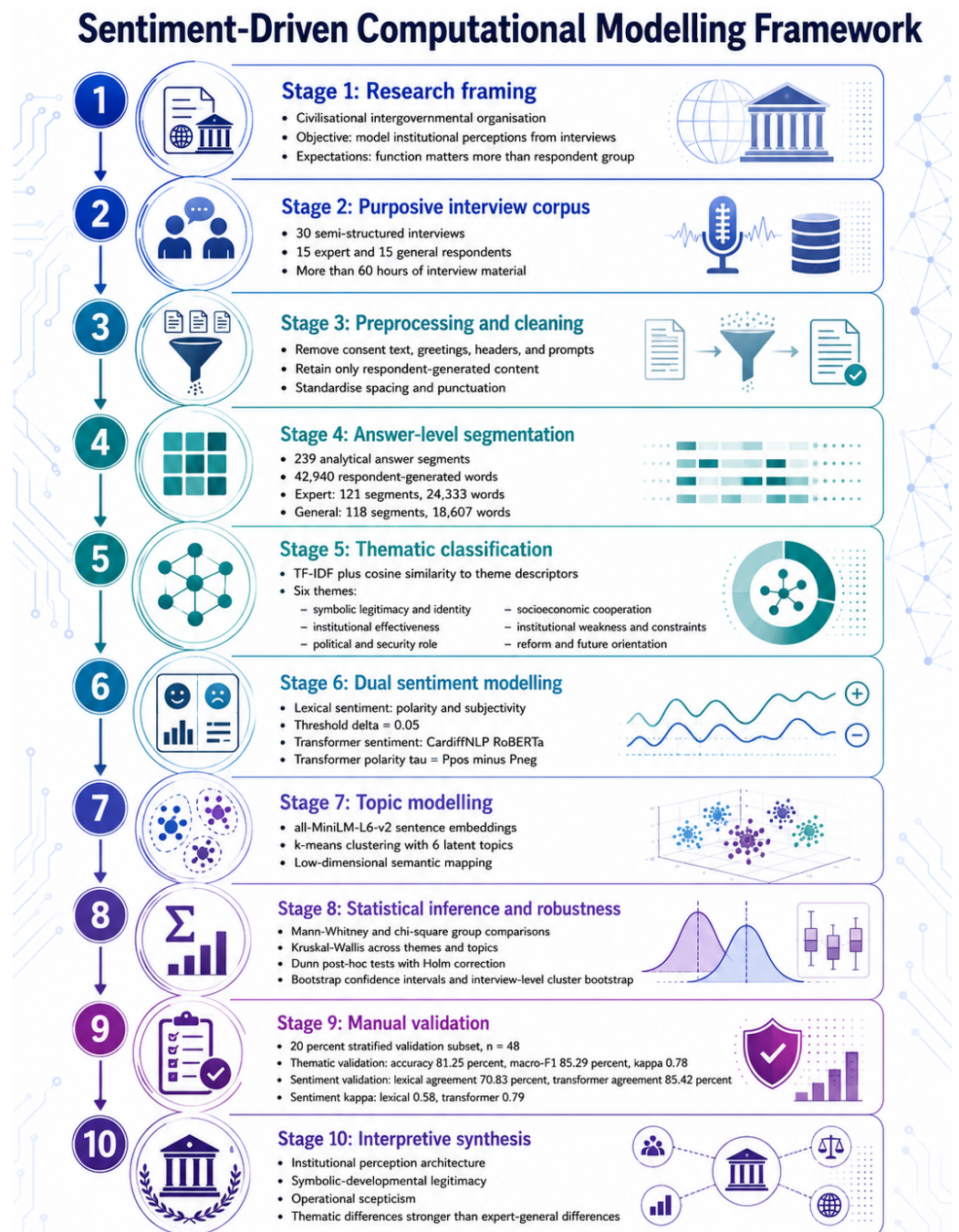


Figure 1. Analytical architecture of the computational thematic sentiment modelling framework for institutional perception analysis, showing how interview narratives are transformed into validated thematic, sentiment, topic, statistical, and interpretive outputs.

The methodological rationale follows three principles. First, interview narratives are treated as structured textual evidence rather than as simple opinion statements. Second, sentiment is modelled at multiple levels, including lexical polarity, contextual transformer-based sentiment, thematic sentiment, and cluster-level sentiment. Third, computational outputs are validated through manual coding because sentiment models may misinterpret politically nuanced discourse, indirect criticism, or mixed evaluations. Lexicon-based methods provide transparent baseline sentiment indicators [5,9], whereas transformer-based models capture contextual semantic dependencies more effectively [10,11]. Latent semantic clustering is used to identify perception clusters in the corpus. It is reported as clustering rather than conventional probabilistic topic modelling because the procedure uses TF-IDF and latent semantic analysis representations followed by k-means clustering. Non-parametric statistical tests are selected because sentiment distributions derived from qualitative interview segments cannot be assumed to follow normality [17,18]. Bootstrap confidence intervals are used to estimate uncertainty around theme-level and cluster-level mean sentiment scores [19].

3.1. Data Collection and Interview Corpus

The empirical dataset was constructed from 30 semi-structured interviews conducted with two respondent groups: 15 general respondents and 15 expert respondents. A purposive sampling strategy was adopted to identify participants with academic, professional, or institutional familiarity with international organisations and the role of the 57-member civilisational intergovernmental organisation examined in this study. General respondents primarily included PhD fellows, postgraduate researchers, and early-career scholars in international relations, political science, and related fields. Expert respondents included university academics, senior researchers, policy professionals, officials, and individuals with professional exposure to international cooperation or related institutional settings. Interviews were conducted through both face-to-face and online formats, depending on participant location and availability. Respondents were geographically dispersed across Muslim-majority contexts, including Bangladesh, Pakistan, Turkey, and Malaysia, as well as Muslim-minority contexts, including the United States, the United Kingdom, and Romania. All interviews followed a semi-structured protocol designed to maintain consistency across core questions while allowing respondents to elaborate on institutional legitimacy, governance capacity, cooperation, communication, constraints, and reform.

The use of 30 interviews is justified by the purposive and criterion-based design of the study. Since the organisation examined here is a specialised intergovernmental institution, meaningful commentary required prior academic, professional, or institutional familiarity with international organisations and related political contexts. The sample was therefore organised into two purposively differentiated tiers: 15 general respondents with postgraduate or early-career academic grounding in history, international relations, political science, or related fields, and 15 expert respondents, including researchers, faculty members, policy professionals, officials, journalists, or individuals with direct professional exposure to the organisation. This structure supports analytical comparison between theoretically informed and practitioner-oriented perspectives. In qualitative inquiry, sample sufficiency is judged less by numerical representativeness than by information richness, relevance to the research question, and thematic saturation [20–22]. Prior saturation research indicates that core themes can often be identified within relatively small (as little as only 12) purposive samples, particularly when participants are information-rich and the topic is bounded [23]. Indeed, Guest et al. (2006) demonstrated in their landmark study that saturation can often be achieved with as few as 12 interviews in a homogeneous purposive sample [23]. A sample of 30 with two distinct expert tiers substantially surpasses this threshold. In the

present study, each interview exceeded two hours, producing more than 60 h of interview material and 42,940 respondent-generated words after cleaning. The dataset is therefore treated as sufficient for exploratory and theory-guided computational interpretation, while the manuscript avoids claims of broad population-level generalisation.

3.2. Notation

Table 2 defines the principal notation used in the methodological formulation.

Table 2. Notation used in the computational thematic sentiment modelling framework.

Symbol	Definition
$\mathcal{D}$	Complete interview corpus
d_i	The i -th interview document
N	Number of interview documents, where $N = 30$
$\mathcal{D}_E$	Expert interview corpus
$\mathcal{D}_G$	General respondent interview corpus
$\mathcal{S}$	Set of answer-level segments after preprocessing
s_j	The j -th answer segment
M	Number of answer segments, where $M = 239$ after final cleaning
g_j	Respondent group label for segment s_j , where $g_j \in \{E, G\}$
w_j	Word count of segment s_j
$\mathcal{T}$	Set of institutional themes
T_k	The k -th institutional theme
θ_j	Theme assignment of segment s_j
p_j	Positive lexical sentiment score of segment s_j
n_j	Negative lexical sentiment score of segment s_j
u_j	Neutral lexical sentiment score of segment s_j
c_j	Compound lexical polarity score of segment s_j
b_j	Subjectivity score of segment s_j
τ_j	Transformer-based polarity score of segment s_j
γ_j	Transformer confidence score for segment s_j
$\mathbf{e}_j$	Latent semantic vector of segment s_j
z_j	Latent semantic cluster assignment of segment s_j
$\mathcal{V}$	Manual validation subset
m_j	Manual sentiment label assigned to segment s_j
IP_j	Composite institutional perception score for segment s_j

3.3. Corpus Construction and Preprocessing

Let the full corpus be represented as

$$\mathcal{D} = \{d_1, d_2, \dots, d_N\}, \quad (1)$$

where $N = 30$. The corpus consists of two respondent groups:

$$\mathcal{D} = \mathcal{D}_E \cup \mathcal{D}_G, \quad \mathcal{D}_E \cap \mathcal{D}_G = \emptyset, \quad (2)$$

with

$$|\mathcal{D}_E| = 15, \quad |\mathcal{D}_G| = 15. \quad (3)$$

Each raw transcript d_i^{raw} was transformed into a cleaned document d_i^{clean} through a preprocessing operator:

$$d_i^{clean} = f_{clean}(d_i^{raw}). \quad (4)$$

The cleaning function removed non-analytical textual components:

$$d_i^{clean} = d_i^{raw} \setminus \{C, I, H, A, Q\}, \quad (5)$$

where C denotes consent material, I denotes repeated instructions, H denotes headers, A denotes administrative text, and Q denotes interviewer questions. The retained content therefore consists only of respondent-generated discourse:

$$d_i^{clean} = R_i. \quad (6)$$

3.4. Answer-Level Segmentation

Each cleaned transcript was segmented into answer-level analytical units:

$$d_i^{clean} = \{s_{i1}, s_{i2}, \dots, s_{im_i}\}, \quad (7)$$

where m_i is the number of retained answer segments in document d_i . The full segment-level corpus is then

$$\mathcal{S} = \bigcup_{i=1}^N \{s_{i1}, s_{i2}, \dots, s_{im_i}\}. \quad (8)$$

After removing courtesy-only and non-substantive segments, the final segment corpus contained:

$$|\mathcal{S}| = M = 239. \quad (9)$$

The total number of respondent-generated words was

$$W = \sum_{j=1}^M w_j = 42,940. \quad (10)$$

Each segment is represented as

$$s_j = (x_j, g_j, w_j), \quad (11)$$

where x_j is the textual content, g_j is the respondent group, and w_j is the word count.

3.5. Thematic Classification

The study defines a set of institutional perception themes:

$$\mathcal{T} = \{T_1, T_2, \dots, T_K\}, \quad (12)$$

where $K = 6$. The themes are symbolic legitimacy and identity, institutional effectiveness, political and security role, socioeconomic cooperation, institutional weakness and constraints, and reform and future orientation.

Each segment was assigned to an institutional theme using a computational similarity function:

$$\theta_j = f_{\text{theme}}(s_j), \quad \theta_j \in \mathcal{T}. \quad (13)$$

To operationalise this, each segment and each theme descriptor were transformed into TF-IDF vectors. Let $\mathbf{v}_j$ denote the TF-IDF vector for segment s_j , and let $\mathbf{q}_k$ denote the TF-IDF vector for theme T_k . The cosine similarity between segment s_j and theme T_k is

$$\text{sim}(s_j, T_k) = \frac{\mathbf{v}_j \cdot \mathbf{q}_k}{\|\mathbf{v}_j\| \|\mathbf{q}_k\|}. \quad (14)$$

The assigned theme is therefore

$$\theta_j = \arg \max_{T_k \in \mathcal{T}} \text{sim}(s_j, T_k). \quad (15)$$

This procedure provides a transparent bridge between qualitative theoretical categories and computational classification. It avoids treating themes as purely emergent machine clusters and instead constrains the classification according to theoretically meaningful institutional dimensions.

3.6. Validation Design for Thematic Classification

To improve transparency in the thematic classification procedure, the six institutional theme descriptors used for TF-IDF cosine similarity matching are reported in Table 3. These descriptors were constructed from three sources: the conceptual framing of institutional legitimacy in the literature, the semi-structured interview protocol, and an initial close reading of the interview corpus. The descriptors were not used as outcome labels in a supervised model; rather, they served as theoretically anchored semantic anchors against which each answer segment was compared. This procedure was intended to preserve conceptual interpretability while allowing systematic and reproducible segment-level classification.

Table 3. Institutional theme descriptors used for TF-IDF cosine similarity classification.

Theme	Descriptor Terms Used for Classification
Symbolic legitimacy and identity	legitimacy, identity, representation, solidarity, collective identity, cultural belonging, civilisational identity, symbolic authority, shared values, community, recognition, moral voice
Institutional effectiveness	effectiveness, performance, achievement, implementation, institutional role, capacity, coordination, governance, decision making, policy delivery, organisational function, practical contribution
Political and security role	political role, diplomacy, conflict, security, mediation, international politics, rights advocacy, geopolitical influence, crisis response, peace, protection, minority rights, political coordination
Socioeconomic cooperation	cooperation, development, economic cooperation, trade, education, science, technology, poverty, health, social progress, development partnership, investment, knowledge exchange
Institutional weakness and constraints	weakness, limitation, constraint, fragmentation, lack of unity, implementation gap, bureaucracy, ineffectiveness, internal division, lack of enforcement, political rivalry, institutional failure
Reform and future orientation	reform, modernisation, future, transformation, improvement, strategic change, institutional renewal, digitalisation, youth, innovation, long-term vision, future relevance

To assess the reliability of the automated thematic assignment, a validation subset of approximately 20% of the analytical segments was selected using stratified sampling across respondent group and computationally assigned theme. Given the final corpus size of 239 answer segments, the validation subset contained 48 segments. Two independent coders manually assigned each sampled segment to one of the six institutional themes using the definitions in Table 3. Disagreements between coders were resolved through discussion, and the reconciled manual labels were used as the reference labels for evaluating the TF-IDF cosine similarity classification.

Classification performance was evaluated using accuracy, macro-precision, macro-recall, and macro-F1. Accuracy was calculated as the proportion of validation segments for which the automated theme label matched the reconciled manual label. Macro-F1 was used because the theme distribution was uneven and because smaller categories, such as reform and future orientation, should not be overwhelmed by larger categories. Inter-coder

reliability was assessed using Cohen's kappa before reconciliation. The validation metrics and confusion matrix are reported in the Results section.

3.7. Lexical Sentiment Modelling

Lexical sentiment analysis was used as the first sentiment estimation layer because it provides interpretable baseline polarity scores. For each segment s_j , the lexical sentiment model returns

$$L_j = (p_j, n_j, u_j, c_j, b_j), \quad (16)$$

where p_j is positive sentiment, n_j is negative sentiment, u_j is neutrality, c_j is compound polarity, and b_j is subjectivity.

The compound polarity score is formalised as

$$c_j = \frac{p_j - n_j}{p_j + n_j + u_j + \epsilon}, \quad (17)$$

where $\epsilon > 0$ is a small stabilising constant. The group-level mean polarity is

$$\bar{c}_g = \frac{1}{|\mathcal{S}_g|} \sum_{s_j \in \mathcal{S}_g} c_j, \quad g \in \{E, G\}. \quad (18)$$

Similarly, theme-level mean polarity is computed as

$$\bar{c}_{T_k} = \frac{1}{|\mathcal{S}_{T_k}|} \sum_{s_j \in \mathcal{S}_{T_k}} c_j, \quad (19)$$

where

$$\mathcal{S}_{T_k} = \{s_j : \theta_j = T_k\}. \quad (20)$$

3.8. Transformer-Based Sentiment Modelling

Lexical approaches are limited when applied to political and institutional discourse because sentiment may be expressed indirectly, cautiously, or in mixed form. To address this limitation, a transformer-based classifier was incorporated as a contextual sentiment layer. Transformer models are appropriate here because they learn bidirectional contextual representations and can capture phrase-level and sentence-level semantic dependencies [10,11].

For each answer segment s_j , the transformer classifier estimates a probability distribution over sentiment labels:

$$P(y_j | s_j) = f_{trans}(s_j), \quad (21)$$

where

$$y_j \in \{\text{positive, neutral, negative}\}. \quad (22)$$

The predicted sentiment label is

$$\hat{y}_j = \arg \max_{y \in Y} P(y | s_j). \quad (23)$$

The transformer confidence score is defined as

$$\gamma_j = \max\{P_{pos,j}, P_{neu,j}, P_{neg,j}\}. \quad (24)$$

A continuous transformer polarity score is calculated as

$$\tau_j = P_{pos,j} - P_{neg,j}, \quad \tau_j \in [-1, 1]. \quad (25)$$

3.9. Lexical-Transformer Agreement

To compare lexical and transformer outputs, the lexical compound score was mapped into a categorical sentiment label:

$$\hat{z}_j = \begin{cases} \text{positive,} & c_j > \delta, \\ \text{negative,} & c_j < -\delta, \\ \text{neutral,} & -\delta \leq c_j \leq \delta, \end{cases} \quad (26)$$

where δ is the neutrality threshold. Model agreement is then

$$A_{LT} = \frac{1}{M} \sum_{j=1}^M \mathbb{I}(\hat{z}_j = \hat{y}_j), \quad (27)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This measure evaluates the consistency between the transparent lexical baseline and the context-sensitive transformer model.

3.10. Latent Semantic Clustering

Latent semantic clustering was used to identify interpretable semantic groupings in the interview corpus. The procedure is not treated as conventional probabilistic topic modelling. Instead, answer segments were represented through TF-IDF features, projected into a lower-dimensional latent semantic space, and grouped using k-means clustering. Each segment was transformed into a numerical representation:

$$\mathbf{e}_j = f_{lsa}(s_j), \quad \mathbf{e}_j \in \mathbb{R}^d. \quad (28)$$

The full latent semantic matrix is

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_M \end{bmatrix}. \quad (29)$$

A clustering function then assigns each segment to a latent semantic cluster:

$$z_j = f_{cluster}(\mathbf{e}_j), \quad z_j \in \{1, 2, \dots, K_z\}. \quad (30)$$

In this study, six latent semantic clusters were retained as an interpretability-driven compromise rather than as a uniquely optimal statistical solution. Although some diagnostic criteria were stronger for neighbouring values of K , the six-cluster solution provided the clearest balance between interpretability, thematic separation, cluster-size balance, stability, and comparability with the institutional dimensions under investigation.

For each semantic cluster z , the cluster-level polarity is

$$\bar{c}_z = \frac{1}{|\mathcal{S}_z|} \sum_{s_j \in \mathcal{S}_z} c_j, \quad (31)$$

where

$$\mathcal{S}_z = \{s_j : z_j = z\}. \quad (32)$$

Representative terms were extracted from the highest-weighted TF-IDF features within each cluster, and semantic labels were assigned through joint inspection of these representative terms and exemplar answer segments. This allows the study to determine

which latent perception clusters generate the strongest positive, neutral, or restrained evaluative patterns.

3.11. Statistical Testing

The primary group-level comparison tests whether expert and general respondents differ in their segment-level sentiment distributions. The null hypothesis is:

$$H_0 : F_E(c) = F_G(c), \quad (33)$$

where $F_E(c)$ and $F_G(c)$ denote the sentiment score distributions for expert and general respondents. The alternative hypothesis is

$$H_1 : F_E(c) \neq F_G(c). \quad (34)$$

Because normality cannot be assumed for interview-derived sentiment scores, the Mann–Whitney U test was used:

$$U_E = n_E n_G + \frac{n_E(n_E + 1)}{2} - R_E, \quad (35)$$

where n_E and n_G are group sample sizes and R_E is the rank sum for the expert group.

For Mann–Whitney comparisons, effect size was reported as the rank-biserial correlation, denoted by r_{rb} . Using the Mann–Whitney statistic for the expert group, the rank-biserial correlation was calculated as:

$$r_{rb} = \frac{2U_E}{n_E n_G} - 1, \quad (36)$$

where U_E is the Mann–Whitney statistic for the expert group, and n_E and n_G are the expert and general respondent sample sizes. The statistic ranges from -1 to $+1$. A negative value indicates that the expert group tends to have lower sentiment ranks than the general respondent group, whereas a positive value indicates higher expert-group ranks. The magnitude was interpreted using the absolute value of r_{rb} , with values below 0.10 treated as negligible, values around 0.10–0.30 as small, values around 0.30–0.50 as small to moderate or moderate, and values above 0.50 as large.

Because the analytical corpus consists of multiple answer segments nested within 30 interviews, the interview rather than the segment was treated as the primary sampling unit for robustness interpretation. Segment-level tests were retained because thematic and sentiment labels were assigned at the answer-segment level; however, they were interpreted as within-corpus evidence rather than as independent population-level observations. To reduce the risk of pseudo-replication, group-level sentiment comparison was additionally repeated at the interview level by aggregating segment polarity scores within each interview. For interview i , the mean interview-level polarity was calculated as

$$\bar{c}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} c_{ij}, \quad (37)$$

where m_i is the number of retained answer segments in interview i and c_{ij} is the polarity score of segment j in interview i . The expert and general respondent groups were then compared using the resulting 30 interview-level mean polarity scores. Accordingly, segment-level results are reported as fine-grained textual evidence, while interview-level results are used as a robustness check against inflated inference due to nested segments.

To further address the nested structure of the data, an interview-level cluster bootstrap was used as a robustness check for the main group-level comparison. Rather than resampling individual answer segments, the bootstrap procedure resampled interviews with replacement within respondent groups. For each bootstrap iteration, all retained segments belonging to the selected interviews were included, and the mean polarity difference between expert and general respondents was recalculated. This procedure preserves the within-interview dependence structure and avoids treating all 239 segments as fully independent observations. The cluster bootstrap was repeated with 1000 resamples, and 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

To test whether sentiment differs across themes, the Kruskal–Wallis statistic is

$$H = \frac{12}{M(M+1)} \sum_{k=1}^K \frac{R_k^2}{n_k} - 3(M+1), \quad (38)$$

where R_k is the rank sum for theme T_k and n_k is the number of segments assigned to that theme.

For categorical sentiment label distributions, a chi-square test was applied:

$$\chi^2 = \sum_{r=1}^R \sum_{c=1}^C \frac{(O_{rc} - E_{rc})^2}{E_{rc}}, \quad (39)$$

where O_{rc} and E_{rc} are the observed and expected frequencies in row r and column c . Cramer's V was used as an effect size measure:

$$V = \sqrt{\frac{\chi^2}{M \min(R-1, C-1)}}. \quad (40)$$

Bootstrap confidence intervals were computed for theme-level and cluster-level mean sentiment:

$$CI_{95\%}(\bar{c}) = [Q_{0.025}(\bar{c}^*), Q_{0.975}(\bar{c}^*)], \quad (41)$$

where $\bar{c}^*$ denotes bootstrap-resampled mean polarity estimates.

3.12. Manual Validation

A manual validation sample was selected to assess the reliability of computational sentiment outputs. The validation subset is

$$\mathcal{V} \subset \mathcal{S}, \quad |\mathcal{V}| \approx 0.20M. \quad (42)$$

Given $M = 239$, the validation set contains approximately:

$$|\mathcal{V}| \approx 48. \quad (43)$$

Each sampled segment was manually coded as

$$m_j \in \{\text{positive, negative, neutral, mixed}\}. \quad (44)$$

Agreement between manual labels and transformer predictions is

$$A_{MT} = \frac{1}{|\mathcal{V}|} \sum_{s_j \in \mathcal{V}} \mathbb{I}(m_j = \hat{y}_j). \quad (45)$$

Agreement between manual labels and lexical predictions is

$$A_{ML} = \frac{1}{|\mathcal{V}|} \sum_{s_j \in \mathcal{V}} \mathbb{I}(m_j = \hat{z}_j). \quad (46)$$

where required, Cohen's kappa may be used to adjust for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (47)$$

where p_o is observed agreement and p_e is expected agreement by chance [24].

The same validation subset was also used to compare manually assigned sentiment labels with the lexical and transformer-based sentiment predictions, allowing manual–lexical agreement, manual–transformer agreement, Cohen's kappa, and disagreement patterns to be assessed empirically.

3.13. Composite Institutional Perception as Interpretive Formalisation

The final modelling layer does not estimate a predictive regression model. Instead, it provides an interpretive formalisation of how the different computational outputs of the study can be understood as components of institutional perception. This clarification is important because the present corpus contains 30 interviews and 239 answer-level segments, which is appropriate for exploratory computational interpretation but not for stable estimation of a multi-parameter predictive model. Accordingly, the following expression should be read as a conceptual synthesis of the analytical dimensions used in the study, rather than as an empirically calibrated equation.

For each segment s_j , institutional perception is represented as a multidimensional construct:

$$IP_j = \mathcal{F}(c_j, \tau_j, \Theta_j, Z_j, G_j), \quad (48)$$

where c_j denotes lexical polarity, τ_j denotes transformer-based polarity, Θ_j denotes thematic membership, Z_j denotes latent semantic cluster membership, and G_j denotes respondent group. The function $\mathcal{F}(\cdot)$ is not estimated in this study. Rather, it summarises the analytical logic that institutional perception is shaped by lexical sentiment, contextual sentiment, thematic framing, latent semantic structure, and respondent category.

The corpus-level institutional perception pattern can therefore be interpreted as

$$IP = \{\bar{c}_g, \bar{c}_{T_k}, \bar{c}_z, \tau_j, \Theta_j, Z_j, G_j\}, \quad (49)$$

where $\bar{c}_g$ represents group-level mean polarity, $\bar{c}_{T_k}$ represents theme-level mean polarity, and $\bar{c}_z$ represents cluster-level mean polarity. This formulation is intended to organise the empirical findings rather than to produce a single predictive score. It allows the study to interpret institutional perception as a structured configuration of symbolic legitimacy, socioeconomic and developmental confidence, political-security scepticism, institutional constraint, and reform-oriented expectation.

Future research using larger interview samples or survey-linked corpora may operationalise this formalisation as an estimated model by assigning weights to the lexical, contextual, thematic, semantic-cluster, and respondent-level components. In the present study, however, the formulation is used only as an interpretive synthesis framework for integrating the empirical results reported in the Results and Discussion sections.

3.14. Algorithmic Implementation

The full computational procedure is summarised in Algorithm 1. The algorithm presents the operational workflow used for corpus construction, thematic assignment, sentiment modelling, validation, robustness checking, and interpretive synthesis.

The computational complexity of Algorithm 1 was also analysed to clarify its scalability. Let M denote the number of answer segments, W the total number of respondent-generated words, K the number of predefined themes, K_z the number of latent semantic clusters, d the latent semantic dimension, I the number of clustering iterations, B the number of bootstrap resamples, $|V|$ the validation subset size, and T the maximum transformer token length. In this study, $M = 239$, $W = 42,940$, $K = 6$, $K_z = 6$, $d=100$, $B = 1000$, $|V| = 48$, and T was capped at 512 tokens. Corpus cleaning, segmentation, and lexical sentiment analysis scale linearly with the corpus size, approximately $\mathcal{O}(W)$. TF-IDF thematic assignment scales as $\mathcal{O}(W + MK)$ because each segment is compared with the predefined theme descriptors. Transformer-based sentiment inference represents the dominant contextual sentiment cost, approximately $\mathcal{O}(MT^2)$, due to the quadratic self-attention operation with respect to sequence length. Latent semantic clustering using TF-IDF, truncated SVD, and k-means scales approximately with the TF-IDF matrix size and the number of clustering iterations, while statistical testing and bootstrap estimation scale approximately as $\mathcal{O}(M \log M + BM)$. Manual validation scoring scales linearly with the validation subset, $\mathcal{O}(|V|)$. Overall, because the corpus contains 239 answer segments and the transformer input length was capped at 512 tokens, the full pipeline remains computationally tractable on standard research computing hardware. For larger corpora, the main scalability constraint would arise from transformer inference, while latent semantic vectorisation and clustering can be scaled through sparse matrix operations and efficient batching.

3.15. Computational Implementation Details

To ensure reproducibility, all computational procedures were implemented in Python 3.11 using `pandas`, `numpy`, `scipy`, `scikit-learn`, `textblob`, `transformers`, `torch`, and `matplotlib`. The reproducibility package additionally reports the preprocessing script, parameter settings, lexical threshold sensitivity analysis, K-sensitivity diagnostics, per-class validation metrics, pairwise cluster tests, and cluster-bootstrap outputs. The interview corpus was analysed in English. Prior to modelling, transcripts were manually inspected and cleaned to remove consent text, repeated instructions, interviewer prompts, greetings, headers, and administrative material. Respondent-generated answers were then segmented into answer-level units. Text normalisation included removal of excessive whitespace, standardisation of line breaks, correction of obvious transcription spacing errors, and preservation of sentence punctuation. No stemming or lemmatisation was applied before sentiment modelling, because these operations can alter sentiment-bearing expressions. Stop-word removal was used only for TF-IDF-based thematic and latent semantic cluster representations, not for lexical or transformer sentiment analysis.

Algorithm 1 Computational Thematic Sentiment Modelling and Validation Procedure

Require: Raw interview corpus $\mathcal{D}^{raw}$, respondent labels g_i , theme descriptors $\mathcal{Q}$, lexical model L , transformer model f_{trans} , latent semantic clustering model $f_{cluster}$, threshold $\delta = 0.05$, validation ratio $r = 0.20$, bootstrap iterations $B = 1000$

Ensure: Theme labels, sentiment scores, cluster labels, validation metrics, robustness statistics, and interpretive perception synthesis

- 1: Initialise cleaned segment corpus $\mathcal{S} \leftarrow \emptyset$
- 2: **for** each transcript $d_i^{raw} \in \mathcal{D}^{raw}$ **do**
- 3: Remove consent text, instructions, greetings, headers, administrative text, and interviewer questions
- 4: Segment respondent-generated text into answer-level units
- 5: Remove courtesy-only and non-substantive segments
- 6: Store each retained segment with respondent group g_i and interview identifier i
- 7: **end for**
- 8: **for** each segment $s_j \in \mathcal{S}$ **do**
- 9: Assign theme θ_j using TF-IDF cosine similarity to theme descriptors
- 10: Compute lexical polarity c_j , subjectivity b_j , and lexical label $\hat{z}_j$ using threshold δ
- 11: Obtain transformer probabilities $(P_{pos,j}, P_{neu,j}, P_{neg,j})$
- 12: Compute transformer polarity $\tau_j = P_{pos,j} - P_{neg,j}$ and transformer label $\hat{y}_j$
- 13: Generate latent semantic vector $\mathbf{e}_j = f_{lsa}(s_j)$
- 14: **end for**
- 15: Apply $f_{cluster}$ to $\{\mathbf{e}_j\}$ and assign semantic cluster label z_j to each segment
- 16: Compute group-level, theme-level, and cluster-level polarity summaries
- 17: Compute bootstrap confidence intervals for theme-level and cluster-level mean polarity
- 18: Conduct Mann–Whitney tests for expert–general comparison
- 19: Conduct Kruskal–Wallis tests across themes and semantic clusters
- 20: Conduct chi-square test for categorical sentiment label distributions
- 21: Aggregate segment polarity scores within each interview to obtain interview-level means
- 22: Recompute the expert–general comparison using interview-level mean polarity scores
- 23: **for** $b = 1$ to B **do**
- 24: Resample interviews with replacement within each respondent group
- 25: Retain all segments belonging to selected interviews
- 26: Compute bootstrap expert–general mean polarity difference Δ_b
- 27: **end for**
- 28: Compute cluster-bootstrap confidence interval $CI_{95\%}(\Delta)$ from $\{\Delta_b\}_{b=1}^B$
- 29: Select stratified validation subset $\mathcal{V} \subset \mathcal{S}$, where $|\mathcal{V}| \approx r|\mathcal{S}|$
- 30: **for** each segment $s_j \in \mathcal{V}$ **do**
- 31: Compare manual theme labels with automated theme labels θ_j
- 32: Compare manual sentiment labels with lexical labels $\hat{z}_j$ and transformer labels $\hat{y}_j$
- 33: **end for**
- 34: Compute thematic accuracy, macro-precision, macro-recall, macro-F1, confusion matrix, and intercoder agreement
- 35: Compute manual–lexical agreement, manual–transformer agreement, Cohen’s kappa, and disagreement patterns
- 36: Integrate themes, semantic clusters, sentiment scores, validation evidence, and robustness checks into an interpretive institutional perception synthesis
- 37: **return** Validation metrics, sentiment statistics, semantic clusters, robustness results, and interpretive perception synthesis

Lexical sentiment was computed using the TextBlob polarity and subjectivity estimator. For each answer segment s_j , the lexical polarity score $c_j \in [-1, 1]$ and subjectivity score $b_j \in [0, 1]$ were extracted. Segment-level lexical sentiment labels were assigned using a neutrality threshold of $\delta = 0.05$. Thus, a segment was classified as positive when $c_j > 0.05$, negative when $c_j < -0.05$, and neutral when $-0.05 \leq c_j \leq 0.05$. The stabilising constant used in the compound polarity expression was $\epsilon = 10^{-8}$, included only to avoid division by zero in formal notation. Sensitivity checks were conducted using alternative neutrality thresholds of $\delta = 0.03$, $\delta = 0.05$, and $\delta = 0.07$ to verify whether the principal thematic conclusions remained stable.

The transformer-based sentiment layer was implemented using the pretrained checkpoint `cardiffnlp/twitter-roberta-base-sentiment-latest` from the Hugging Face

transformers library. The model produces three sentiment probabilities corresponding to negative, neutral, and positive classes. The model was used in zero-shot transfer mode without fine-tuning, because the available corpus contained only 239 answer segments and was not sufficiently large for reliable supervised adaptation. Tokenisation was performed using the corresponding RoBERTa tokenizer, with truncation at 512 tokens and padding applied dynamically within each batch. For each segment, transformer polarity was calculated as $\tau_j = P_{pos,j} - P_{neg,j}$, and model confidence was calculated as $\gamma_j = \max(P_{pos,j}, P_{neu,j}, P_{neg,j})$.

Thematic classification was implemented using TF-IDF vectors and cosine similarity. Segment vectors were generated using TfidfVectorizer with unigram and bigram features, English stop-word removal, minimum document frequency of 2, and maximum document frequency of 0.90. Each of the six institutional themes was represented by a short descriptor containing theoretically relevant terms. A segment was assigned to the theme with the highest cosine similarity score. Latent semantic clustering was implemented using the same TF-IDF feature space projected through truncated singular value decomposition with 100 latent dimensions and normalisation before k-means clustering. K-means was run with $K = 6$, random state = 42, n init = 50, and the Lloyd algorithm. Representative cluster terms were extracted from the highest-weighted TF-IDF features within each cluster, and the robustness of the cluster number was assessed for $K = 4, 5, 6, 7$, and 8 using silhouette score, NPMI coherence, top-term diversity, adjusted Rand index stability, cluster-size balance, and interpretability. Table 4 highlights the implementation details for supporting research reproducibility. The selected $K = 6$ solution is therefore interpreted cautiously as a theoretically meaningful working solution, not as evidence that alternative cluster numbers are invalid.

Table 4. Computational implementation settings.

Component	Implementation Detail
Programming environment	Python 3.11 with pandas, numpy, scipy, scikit-learn, textblob, transformers, torch, and matplotlib.
Corpus language	English interview transcripts.
Preprocessing	Removal of consent text, interviewer prompts, greetings, administrative text, repeated instructions, and headers; answer-level segmentation; whitespace and line-break normalisation.
Lexical sentiment model	TextBlob polarity and subjectivity estimator.
Lexical label threshold	Positive if $c_j > 0.05$, negative if $c_j < -0.05$, neutral otherwise. Sensitivity checked for $\delta = 0.03$, $\delta = 0.05$, and $\delta = 0.07$.
Stabilising constant	$\epsilon = 10^{-8}$.
Transformer model	cardiffnlp/twitter-roberta-base-sentiment-latest, used without fine-tuning.
Transformer tokenisation	RoBERTa tokenizer, maximum sequence length of 512 tokens, dynamic padding, truncation for longer segments.
Transformer polarity	$\tau_j = P_{pos,j} - P_{neg,j}$.
Thematic classification	TF-IDF unigram and bigram vectors with cosine similarity to predefined institutional theme descriptors.
TF-IDF settings	English stop-word removal, min_df=2, max_df=0.90, unigram and bigram features.
Latent semantic representation	TF-IDF unigram and bigram features projected by truncated SVD with 100 latent dimensions and normalisation.
Semantic clustering	k-means clustering with $K = 6$, random state = 42, n init = 50, and Lloyd optimisation.
Cluster-number robustness	$K = 4, 5, 6, 7$, and 8 compared using silhouette score, NPMI coherence, top-term diversity, adjusted Rand index stability, cluster-size balance, and interpretability.

4. Results

After preprocessing, removal of administrative material, and exclusion of courtesy-only responses, the final analytical corpus contained 239 answer segments from 30 interviews, corresponding to 42,940 respondent-generated words. The expert group contributed 121 segments and 24,333 words, whereas the general respondent group contributed 118 segments and 18,607 words. The expert responses were longer on average, with 201.10 words per segment compared with 157.69 words per segment for the general group, indicating a more elaborative discourse pattern among expert participants (as seen from Table 5).

Table 5. Corpus profile after preprocessing.

Group	Interviews	Segments	Total Words	Mean Words per Segment	Mean Segments per Interview
Expert	15	121	24,333	201.10	8.07
General	15	118	18,607	157.69	7.87
Overall	30	239	42,940	179.67	7.97

The thematic classification validation results indicate that the TF-IDF cosine similarity procedure achieved acceptable agreement with manually reconciled labels. As shown in Table 6, the automated classification achieved an accuracy of 81.25%, macro-precision of 82.02%, macro-recall of 82.92%, and a macro-F1 of 81.96 percent. Inter-coder agreement before reconciliation was substantial, with Cohen's kappa of 0.78. The confusion matrix in Table 7 shows that most misclassifications occurred between conceptually adjacent categories, particularly symbolic legitimacy, institutional effectiveness, political-security role, and institutional weakness. The corresponding class specific precision, recall, and F1 scores are reported in Table 8. Most themes achieved F1 scores of at least 0.80, whereas institutional effectiveness recorded the lowest F1 score of 0.60, reflecting its greater semantic overlap with adjacent institutional categories. This indicates that the thematic classification procedure is sufficiently reliable for theme-level analysis, while also confirming that some semantic overlap is expected because institutional perception categories are related rather than fully independent.

Table 6. Validation metrics for thematic classification.

Metric	Value	Interpretation
Validation sample size	48 segments	Approximately 20% of the final analytical corpus.
Inter-coder agreement	0.78	Substantial agreement between the two independent manual coders before reconciliation.
Classification accuracy	81.25%	Proportion of automated theme labels matching reconciled manual labels.
Macro-precision	82.02%	Average precision across the six themes.
Macro-recall	82.92%	Average recall across the six themes.
Macro-F1	81.96 percent	Unweighted mean of the six class-level F1 scores.

Table 7. Confusion matrix for TF-IDF cosine similarity thematic classification against reconciled manual labels.

Manual Label/Automated Label	SL	IE	PS	SC	IW	RF
SL	12	1	1	0	0	0
IE	0	3	0	1	0	0
PS	1	0	9	0	1	0
SC	1	1	0	8	0	0
IW	0	1	1	0	6	0
RF	0	0	0	0	0	1

SL = Symbolic legitimacy and identity; IE = Institutional effectiveness; PS = Political and security role; SC = Socioeconomic cooperation; IW = Institutional weakness and constraints; RF = Reform and future orientation.

Table 8. Per-class thematic validation metrics.

Theme	Manual Support	Automated Count	Precision	Recall	F1
SL	14	14	0.8571	0.8571	0.8571
IE	4	6	0.5000	0.7500	0.6000
PS	11	11	0.8182	0.8182	0.8182
SC	10	9	0.8889	0.8000	0.8421
IW	8	7	0.8571	0.7500	0.8000
RF	1	1	1.0000	1.0000	1.0000

Before interpreting the substantive sentiment results, manual validation was also conducted for the lexical and transformer-based sentiment outputs. The validation subset contained 48 answer segments, corresponding to approximately 20% of the final analytical corpus. Each segment was manually labelled as positive, negative, neutral, or mixed, and the reconciled manual labels were compared with the lexical and transformer sentiment labels. As shown in Table 9, the transformer model achieved stronger agreement with manual interpretation than the lexical model. The lexical model achieved manual agreement of 70.83% and Cohen’s kappa of 0.58, whereas the transformer model achieved manual agreement of 85.42% and Cohen’s kappa of 0.79. The disagreement patterns indicate that lexical errors were more frequent in mixed-valence statements, while transformer errors were more likely where criticism was subtle, diplomatic, or context-dependent.

Table 9. Manual validation results for lexical and transformer-based sentiment classification.

Validation Measure	Lexical Model	Transformer Model	Interpretation
Validation sample size	48 segments	48 segments	Approximately 20% of the final analytical corpus.
Manual agreement	70.83%	85.42%	Percentage of model labels matching reconciled manual sentiment labels.
Cohen’s kappa	0.58	0.79	Chance-adjusted agreement between model predictions and manual labels.
Most frequent disagreement pattern	Mixed-valence segments misclassified as purely neutral or negative.	Subtle diplomatic critiques misclassified as positive or neutral.	Disagreements mainly involved mixed, indirect, or context-dependent institutional evaluations.
Main source of error	Rigid token-matching failing to parse co-occurring positive and negative terms.	Latent context dependence and implicit institutional scepticism.	Lexical errors were more common where positive and negative words co-occurred; transformer errors were more common where institutional criticism was implied rather than explicit.

The overall sentiment profile was cautiously positive. Expert respondents recorded a mean polarity score of 0.0803, while general respondents recorded a higher mean polarity score of 0.1016, as shown in Table 10. Median polarity followed the same pattern, with 0.0764 for experts and 0.1014 for general respondents. Mean subjectivity was almost identical across groups, 0.3422 for experts and 0.3456 for general respondents, suggesting that both groups expressed comparable levels of evaluative intensity. The distribution shown in Figure 2 confirms that both groups clustered in the mildly positive range, with only a small number of negative outliers.

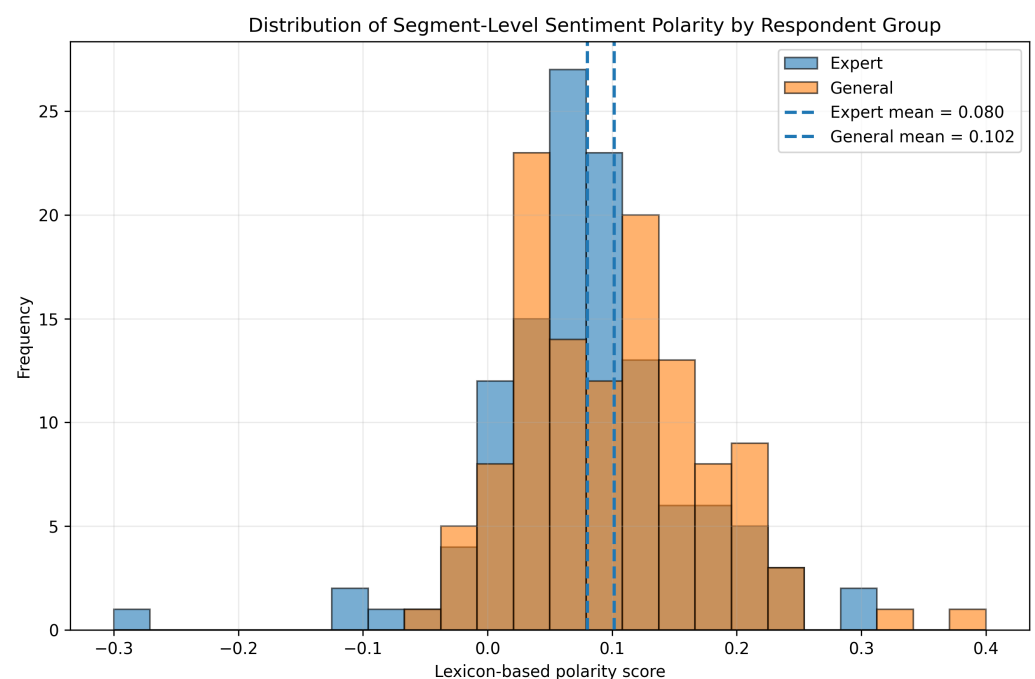
Table 10. Descriptive sentiment statistics by respondent group.

Group	Mean Polarity	Median Polarity	SD Polarity	Mean Subjectivity	Median Subjectivity	Mean Word Count
Expert	0.0803	0.0764	0.0805	0.3422	0.3354	201.10
General	0.1016	0.1014	0.0771	0.3456	0.3344	157.69

The categorical sentiment distribution further supports this pattern as evident from Table 11. In the expert corpus, 85 of 121 segments were positive, representing 70.2%, while 31 segments were neutral, 25.6%, and 5 were negative, 4.1%. In the general corpus, 81 of 118 segments were positive, representing 68.6%, 36 were neutral, 30.5%, and only 1 was negative, 0.8%. Thus, the general group produced fewer explicitly negative segments, although the overall label distribution did not differ significantly between groups.

Table 11. Distribution of sentiment labels by respondent group.

Group	Negative	Neutral	Positive
Expert	5 (4.1%)	31 (25.6%)	85 (70.2%)
General	1 (0.8%)	36 (30.5%)	81 (68.6%)

**Figure 2.** Distribution of segment-level sentiment polarity by respondent group.

Sensitivity analysis confirmed that the categorical sentiment results were not dependent on the selected lexical neutrality threshold. Table 12 shows that the expert–general label association remained non-significant under all three examined thresholds, while the continuous Mann–Whitney comparison was unchanged because the threshold affects only categorical labels and not the underlying polarity scores.

Table 12. Lexical neutrality threshold sensitivity analysis.

δ	Overall N/Neu/P	Expert N/Neu/P	General N/Neu/P	χ^2	p	Cramer's V
0.03	7/43/189	6/21/94	1/22/95	3.5629	0.1684	0.1221
0.05	6/67/166	5/31/85	1/36/81	3.0990	0.2124	0.1139
0.07	4/100/135	4/50/67	0/50/68	3.9704	0.1374	0.1289

As depicted in Table 13, Theme-level results revealed a clear asymmetry in institutional perception. The highest mean polarity was observed for reform and future orientation, 0.1700, although this theme contained only seven segments. Among the more substantively represented categories, socioeconomic cooperation produced the strongest positive sentiment, with 49 segments and a mean polarity of 0.1330. Institutional effectiveness followed closely, with 19 segments and a mean polarity of 0.1325. By contrast, institutional weakness and constraints produced the lowest mean polarity, 0.0546, across 43 segments. Political and security role also remained relatively restrained, with a mean polarity of 0.0693 across 53 segments. This pattern indicates that respondents evaluated developmental and reform-oriented functions more positively than coercive, political, or implementation-related functions.

Figure 3 shows that general respondents were more positive than experts in four of the six thematic categories. The largest group gap appeared in symbolic legitimacy and identity, where general respondents recorded a mean polarity of 0.099 compared with 0.057 among experts. Institutional weakness and constraints showed a similar directional difference, 0.075 for general respondents compared with 0.037 for experts. This suggests that expert discourse was more cautious when discussing identity, legitimacy, and structural limitations.

Table 13. Theme-level sentiment summary with bootstrap confidence intervals.

Theme	Segments	Expert	General	Mean Polarity	Mean Subjectivity	95% CI
Reform and future orientation	7	5	2	0.1700	0.4072	[0.1423, 0.1904]
Socioeconomic cooperation	49	25	24	0.1330	0.3380	[0.1155, 0.1494]
Institutional effectiveness	19	8	11	0.1325	0.3469	[0.1004, 0.1605]
Symbolic legitimacy and identity	68	30	38	0.0803	0.3664	[0.0598, 0.1050]
Political and security role	53	30	23	0.0693	0.3288	[0.0566, 0.0818]
Institutional weakness and constraints	43	23	20	0.0546	0.3221	[0.0291, 0.0757]

As seen from Table 14, latent semantic clustering identified six interpretable perception clusters. The most positive cluster was socioeconomic and economic cooperation, with a mean polarity of 0.1465 across 28 segments. Institutional capacity, member states, and constraints followed with a mean polarity of 0.1122 across 53 segments. Political role, rights, and member-state coordination produced a more restrained mean polarity of 0.0745, while formation, ummatic feelings, and solidarity recorded 0.0638. This pattern indicates that identity-based discourse was not uniformly celebratory, but was often embedded within reflective or historically qualified statements.

To test whether the six-cluster solution was an arbitrary modelling choice, sensitivity analysis compared $K = 4$ through $K = 8$ using silhouette score, NPMI coherence, top-term diversity, adjusted Rand index stability, cluster-size balance, and interpretability. Table 15 shows that no single value of K dominates on every criterion. The six-cluster solution was retained because it provided a defensible compromise between interpretability, separation,

and comparability with the six institutional perception dimensions, while avoiding the more fragmented solutions obtained at $K = 7$ and $K = 8$.

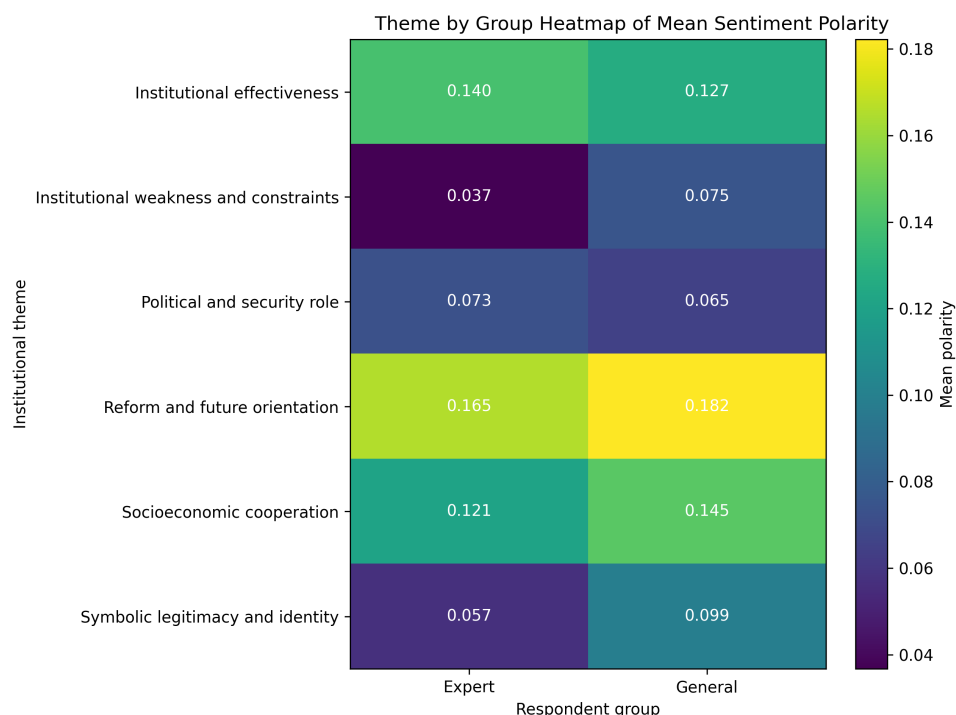


Figure 3. Heatmap of mean sentiment polarity across institutional themes and respondent groups.

Table 14. Cluster-level sentiment summary with representative terms.

Semantic Cluster	Segments	Mean Polarity	Mean Subjectivity	Representative Terms
Socioeconomic and economic cooperation	28	0.1465	0.3525	economic, OIC countries, development, socio economic, trade
Institutional capacity, member states, and constraints	53	0.1122	0.3407	member, member states, organisation, ability, hinder
Visibility, public knowledge, and media coverage	28	0.0787	0.3439	media, lack, role, public, coverage
Regional identity and organisational uniqueness	32	0.0760	0.3763	regional, identity, organisation, uniqueness, cooperation
Political role, rights, and member-state coordination	72	0.0745	0.3132	muslim, countries, member, political, rights
Formation, ummatic feelings, and solidarity	26	0.0638	0.3864	ummatic, feelings, formation, muslim, sense

Table 15. Robustness analysis for alternative numbers of latent semantic clusters.

K	Silhouette	NPMI	Diversity	Stability ARI	Size Range	Interpretability
4	0.0604	0.2381	0.8750	0.4875	29–128	Interpretable but less stable
5	0.0708	0.3133	0.9000	0.6209	24–130	Interpretable but imbalanced
6	0.0805	0.2589	0.8333	0.6889	26–72	Balanced and substantively aligned
7	0.0878	0.2818	0.8000	0.7186	26–70	More fragmented
8	0.0949	0.2868	0.8250	0.8354	26–40	Stable but analytically finer-grained

The two-dimensional semantic cluster map in Figure 4 is used as an interpretive visualisation rather than as an additional statistical model. It shows a visibly structured

semantic space in which socioeconomic and economic cooperation appears as a relatively coherent cluster, while political, rights-related, and identity-oriented discourse occupies more restrained semantic regions. The three-dimensional sentiment surface in Figure 5 is likewise illustrative and supports interpretation by showing that sentiment intensity is not uniformly distributed across the theme-cluster matrix. Peaks are concentrated around developmental and institutional performance regions, whereas lower surfaces correspond to weakness, political restraint, and visibility-related concerns.

As seen from Table 16, inferential testing confirms that the main differences lie across institutional dimensions rather than between respondent groups. The segment-level Mann–Whitney test comparing expert and general polarity scores produced $U = 6171.50$ and $p = 0.0704$, which is not significant at the 0.05 level. At the interview level, the corresponding test produced $U = 71.00$ and $p = 0.0888$. The chi-square test for sentiment label distribution by group was also not significant, $\chi^2 = 3.0990$, $p = 0.2124$, with a weak association, Cramer’s $V = 0.1139$. In contrast, polarity varied significantly across themes, Kruskal–Wallis $H = 45.8768$, $p < 0.001$, and across latent semantic clusters, Kruskal–Wallis $H = 35.0124$, $p < 0.001$.

Table 16. Summary of inferential statistical tests.

Test	Statistic	<i>p</i> -Value	Effect Size	Interpretation
Mann–Whitney, segment-level group comparison	6171.50	0.0704	−0.1355	Small group difference
Mann–Whitney, interview-level group comparison	71.00	0.0888	−0.3689	Small to moderate group difference
Chi-square, label distribution by group	3.0990	0.2124	0.1139	Weak association
Kruskal–Wallis, polarity across themes	45.8768	<0.001	–	Significant thematic variation
Kruskal–Wallis, polarity across semantic clusters	35.0124	<0.001	–	Significant cluster variation

Note: Mann–Whitney effect sizes are rank-biserial correlations, $r_{rb} = (2U_E / (n_E n_G)) - 1$. Negative values indicate lower expert-group sentiment ranks relative to the general respondent group. The chi-square effect size is Cramer’s V .

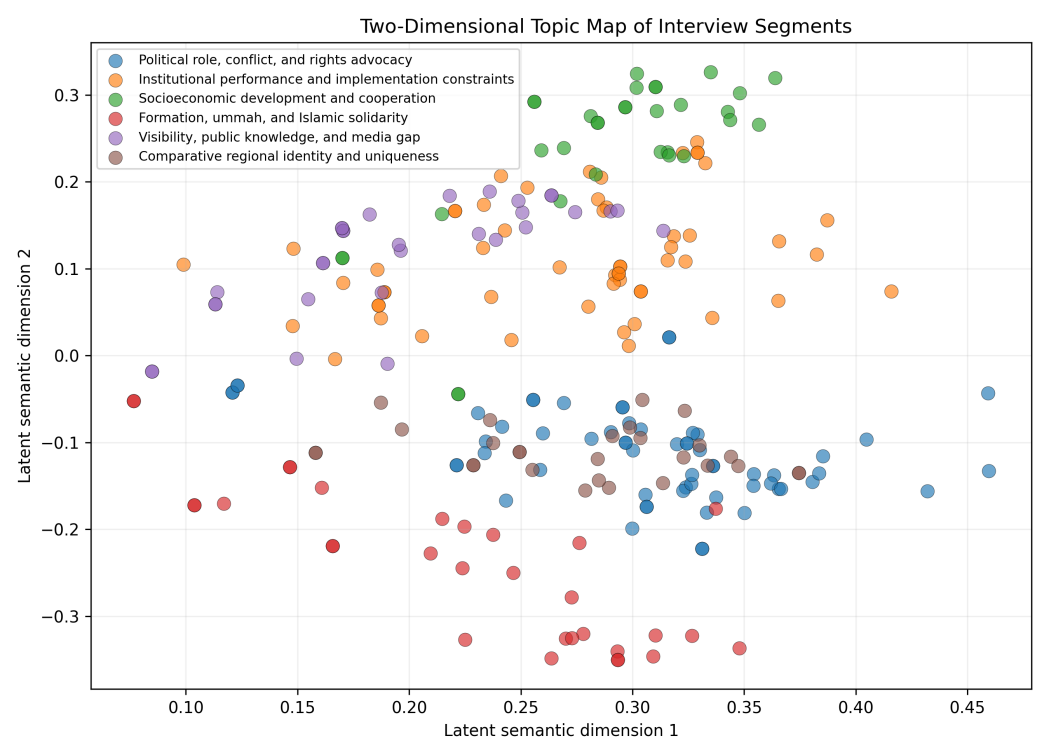


Figure 4. Two-dimensional latent semantic cluster map of interview segments.

Three-Dimensional Theme–Topic Sentiment Surface

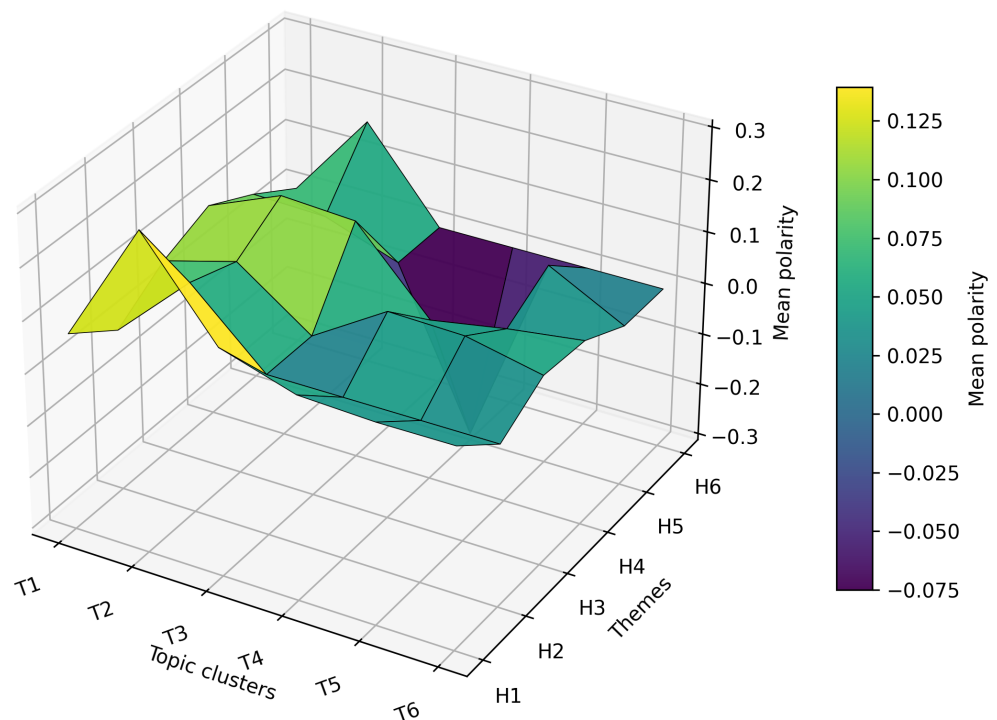


Figure 5. Three-dimensional sentiment surface across themes and latent semantic clusters.

To isolate the specific loci of these significant variations and establish an inferential foundation for the coexistence of symbolic-developmental optimism and operational skepticism, post hoc pairwise comparisons were conducted using Dunn’s test with a Holm–Bonferroni correction. While all 15 pairwise theme combinations were evaluated, Table 17 reports the primary theoretically salient contrasts that define these perceptual cleavages, along with their standardized effect sizes ($r = z / \sqrt{N}$). These theme-level pairwise contrasts should be interpreted as exploratory post hoc evidence rather than as definitive interview-level inferential estimates, because additional interview-level pairwise bootstrapping across all theme contrasts was limited by small category sizes, particularly the reform and future orientation category.

The post hoc results reveal that sentiment towards *Socioeconomic cooperation* is significantly higher than both *Institutional weakness and constraints* ($z = 4.82, p_{\text{Holm}} < 0.001$) and the *Political and security role* ($z = 3.91, p_{\text{Holm}} < 0.001$), carrying moderate effect sizes. Similarly, *reform and future orientation* exhibits a significant positive separation from *Institutional weakness* ($z = 2.94, p_{\text{Holm}} = 0.013$). These robust pairwise variations provide clear inferential evidence that respondents systematically decouple the organisation’s high normative/developmental utility from its operational limitations and structural constraints.

Table 17. Selected Dunn’s post hoc pairwise comparisons for theme-level sentiment polarity with Holm–Bonferroni correction.

Comparison Pair (Theme A vs. Theme B)	Mean Diff.	Dunn’s z	p_{unadj}	p_{Holm}	Effect Size (r)
Socioeconomic coop. vs. Inst. weakness	+0.0784	4.82	<0.001	<0.001	0.312 (Moderate)
Socioeconomic coop. vs. Political/Security	+0.0637	3.91	<0.001	<0.001	0.253 (Small-Mod)
Reform/Future vs. Inst. weakness	+0.1154	2.94	0.003	0.013	0.190 (Small)
Inst. effectiveness vs. Inst. weakness	+0.0779	2.71	0.007	0.027	0.175 (Small)
Symbolic legitimacy vs. Inst. weakness	+0.0257	1.84	0.066	0.198	0.119 (Negligible)

Because latent semantic clusters also showed significant polarity variation, pairwise cluster contrasts were additionally examined using Mann–Whitney tests with Holm correction. Table 18 reports the theoretically most relevant corrected contrasts, using rank-biserial correlation as the effect-size measure. These results show that the socioeconomic and economic cooperation cluster is significantly more positive than the political, identity, visibility, and formation-related clusters.

Table 18. Selected post hoc pairwise comparisons for latent semantic cluster sentiment polarity with Holm correction.

Comparison Pair	Mean Diff.	<i>U</i>	<i>p</i> _{unadj}	<i>p</i> _{Holm}	<i>r</i> _{rb}
Socioeconomic/economic coop. vs. Political/rights	+0.0720	1660.00	less than 0.001	less than 0.001	0.6468
Socioeconomic/economic coop. vs. Regional identity	+0.0705	738.00	less than 0.001	less than 0.001	0.6473
Socioeconomic/economic coop. vs. Visibility/media gap	+0.0678	644.00	less than 0.001	less than 0.001	0.6429
Socioeconomic/economic coop. vs. Formation/solidarity	+0.0827	566.00	less than 0.001	0.0058	0.5549
Institutional capacity vs. Political/rights	+0.0378	2550.00	0.0014	0.0149	0.3365

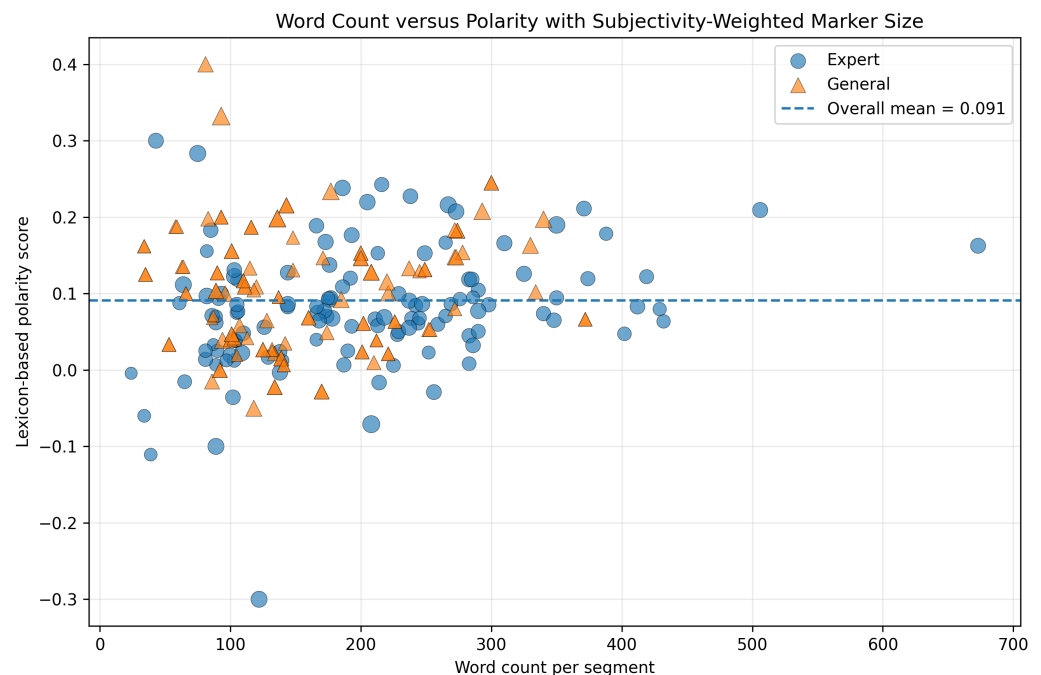
The inferential results were also checked at the interview level to address the nesting of answer segments within interviews. Although the segment-level Mann–Whitney test indicated no statistically significant expert–general difference, $U = 6171.50$, $p = 0.0704$, the same conclusion was retained when polarity scores were aggregated at the interview level. The interview-level Mann–Whitney test produced $U = 71.00$, $p = 0.0888$, indicating that the group-level difference remained non-significant when the interview, rather than the segment, was treated as the unit of comparison. This reduces concern that the expert–general comparison was driven by pseudo-replication across multiple answer segments from the same respondent. Nevertheless, the segment-level thematic and cluster analyses are interpreted as structured within-corpus evidence rather than as estimates supporting broad population-level generalisation.

An interview-level cluster bootstrap was also conducted to assess whether the main expert–general comparison was sensitive to the nesting of segments within interviews. In this procedure, interviews rather than individual segments were resampled, and all segments belonging to each selected interview were retained within each bootstrap sample. The resulting 95% cluster-bootstrap confidence interval for the expert–general mean polarity difference was $[-0.0454, 0.0028]$. Because this interval included zero, the result supports the conclusion that the expert–general sentiment difference should not be interpreted as statistically robust. This finding is consistent with both the segment-level Mann–Whitney test, ($U = 6171.50$), ($p = 0.0704$), and the interview-level Mann–Whitney test, ($U = 71.00$), ($p = 0.0888$). The same cluster-bootstrap logic was extended to the main latent semantic cluster claim by resampling interviews and recalculating mean polarity differences between the highest-polarity socioeconomic/economic cooperation cluster and each remaining cluster. As shown in Table 19, all intervals exclude zero, supporting the robustness of the main cluster-level interpretation. Theme-level bootstrap confidence intervals are reported in Table 13; additional pairwise theme bootstrapping was not emphasised because the reform category contained only seven segments and would yield unstable interview-level resamples.

Table 19. Interview-level cluster-bootstrap robustness for the highest-polarity latent semantic cluster.

Comparison	Observed Difference	95% CI	Includes Zero
Socioeconomic/economic coop. minus formation/solidarity	0.0827	[0.0313, 0.1351]	No
Socioeconomic/economic coop. minus institutional capacity	0.0343	[0.0104, 0.0581]	No
Socioeconomic/economic coop. minus regional identity	0.0705	[0.0424, 0.0952]	No
Socioeconomic/economic coop. minus visibility/media gap	0.0678	[0.0352, 0.0974]	No
Socioeconomic/economic coop. minus political/rights	0.0720	[0.0451, 0.0988]	No

Figure 6 shows that longer responses did not necessarily produce more positive sentiment. Instead, longer expert responses tended to spread across a broader polarity range, indicating more elaborated and qualified evaluation. Figure 7 presents the interview-level profile using deidentified numeric labels. Bubble size represents total word volume per interview. The plot shows that general interviews occupy several of the highest polarity and subjectivity positions, whereas expert interviews are more densely concentrated in the moderate polarity range. This supports the interpretation that expert respondents were more analytically cautious, while general respondents expressed somewhat stronger affective positivity.

**Figure 6.** Segment word count and sentiment polarity with subjectivity-weighted marker size.

Overall, the results show that perceptions of the organisation are not reducible to a simple positive or negative judgement. The dominant empirical pattern is a dual structure: respondents attach comparatively stronger positive sentiment to reform, socioeconomic cooperation, and institutional effectiveness, while expressing more restrained sentiment toward political-security roles, visibility gaps, and institutional constraints. The statistically significant differences across themes and semantic clusters, combined with non-significant differences between respondent groups, indicate that the object of evaluation matters more than the respondent category. The study therefore identifies a measurable perception architecture in which symbolic and developmental legitimacy coexist with operational scepticism.

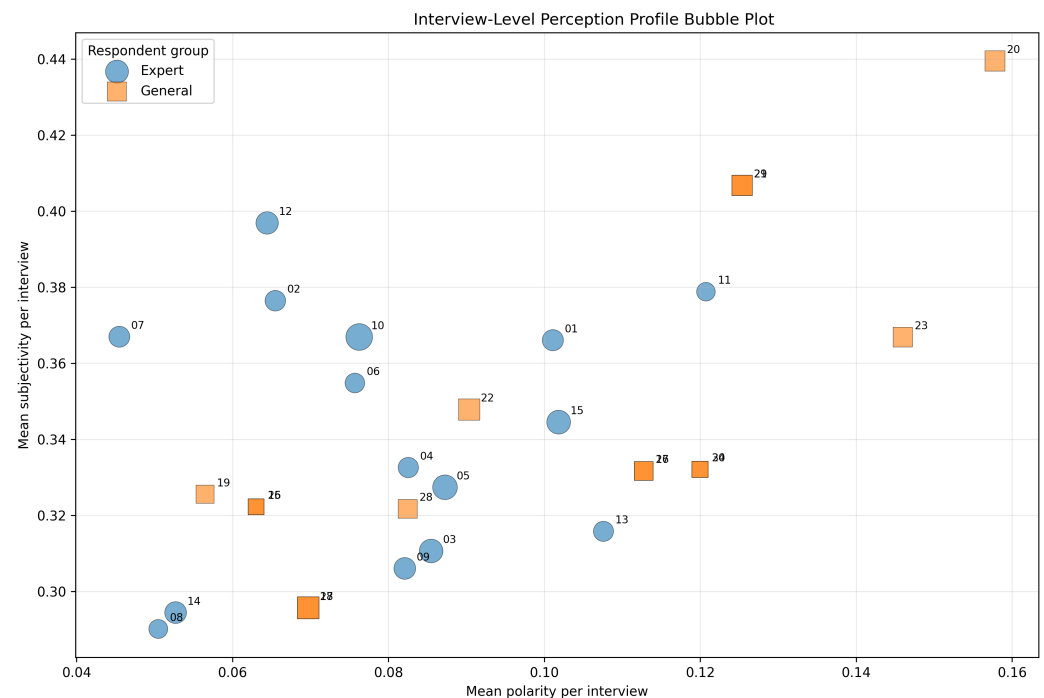


Figure 7. Interview-level perception profile based on mean polarity, mean subjectivity, and total word volume.

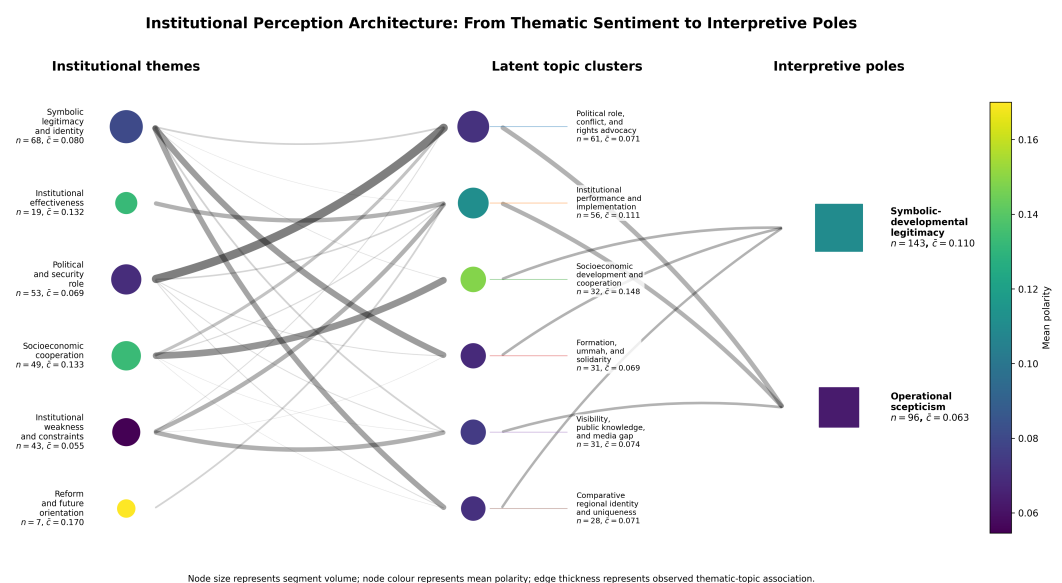
5. Discussion

Before interpreting the empirical findings, it is useful to position the present study against the main bodies of literature on which it builds. Table 20 summarises how existing work on international organisational legitimacy, sentiment analysis, transformer-based language modelling, latent semantic clustering, and prior computational studies relates to the contribution of this paper. The comparison shows that the present study does not merely apply sentiment analysis to interview data; rather, it integrates institutional theory with a multi-layered computational framework for measuring perception through sentiment, theme, semantic cluster, and statistical validation.

To move from literature positioning to empirical interpretation, Figure 8 provides a synthesis of the institutional perception architecture identified in this study. The figure integrates the thematic classification, latent semantic clustering, and sentiment polarity results into a single three-layer representation. The left layer shows the six institutional themes, the middle layer shows the six latent semantic clusters, and the right layer condenses the analysis into two higher-order interpretive poles: symbolic-developmental legitimacy and operational scepticism. Node size represents segment volume, node colour represents mean polarity, and edge thickness represents the observed thematic-cluster association. This synthesis demonstrates that institutional perception is not organised around a simple positive-negative divide, but around a structured tension between reform-oriented, socioeconomic, and symbolic legitimacy on one side, and political-security, visibility, and institutional constraint-related scepticism on the other.

Table 20. Positioning of the present study relative to relevant literature.

Research Direction	Main Contribution of Prior Work	Extension Provided by the Present Study
Institutional legitimacy of international organisations	Prior studies show that legitimacy is shaped by institutional performance, representation, procedure, public confidence, and discursive legitimization [1–4].	This study converts institutional perception into measurable sentiment, theme, and semantic-cluster level indicators using 239 answer segments and 42,940 respondent-generated words.
Sentiment analysis and opinion mining	Sentiment studies provide methods for measuring polarity, subjectivity, and evaluative orientation in textual data [5,6,9].	This study applies sentiment modelling to interview-based institutional perception rather than to standard opinion corpora, news, or social media text.
Transformer-based language modelling	Transformer models improve contextual interpretation of language by capturing semantic dependencies beyond dictionary-based word matching [10,11].	This study incorporates transformer-based sentiment modelling as a contextual robustness layer for politically nuanced and institutionally cautious discourse.
Latent semantic clustering and computational text analysis	Topic modelling and clustering methods extract latent discourse structures and support the formal analysis of political and policy texts [7,8,12,13].	This study links semantic clusters with sentiment scores and institutional themes, showing significant variation across clusters, Kruskal–Wallis $H = 35.0124$, $p < 0.001$.
Prior computational studies by the author	Previous work used sentiment analysis, entity extraction, regression, GPT-based classification, and media bias quantification for large-scale textual corpora [14,15].	This study extends that computational logic from media and cyber intelligence data to interview-based institutional perception modelling.

**Figure 8.** Institutional perception architecture linking themes, latent semantic clusters, and the two interpretive poles of symbolic-developmental legitimacy and operational scepticism. Node size denotes segment volume, node colour denotes mean polarity, and edge thickness denotes thematic-cluster association.

Building on Table 20 and Figure 8, the findings extend existing work on institutional legitimacy by showing that perception of a civilisational intergovernmental organisation can be quantified without reducing it to a single approval or disapproval score. Earlier studies have established that international organisational legitimacy depends on performance, representation, procedure, and confidence [1–4]. The present study adds a computational layer to that literature by demonstrating that institutional perception varies more sharply across themes and semantic clusters than across respondent categories. Expert and general respondents differed only modestly in overall sentiment, with mean polarity scores of 0.0803

and 0.1016, respectively. These group-level differences were not statistically significant at either segment level, Mann–Whitney $U = 6171.50$, $p = 0.0704$, or interview level, $U = 71.00$, $p = 0.0888$. By contrast, the variation across institutional themes was highly significant, Kruskal–Wallis $H = 45.8768$, $p < 0.001$, indicating that the object of evaluation matters more than respondent type.

The results reveal a dual perception structure. On one side, respondents expressed stronger positive sentiment toward reform and future orientation, mean polarity = 0.1700, socioeconomic cooperation, 0.1330, and institutional effectiveness, 0.1325. On the other side, more restrained sentiment appeared in relation to institutional weakness and constraints, 0.0546, and political and security role, 0.0693. This suggests that the organisation is not perceived as simply successful or unsuccessful. Rather, it is evaluated through a differentiated structure in which symbolic and developmental legitimacy coexist with operational scepticism. The latent semantic clustering results strengthen this interpretation: socioeconomic and economic cooperation emerged as the most positively evaluated semantic cluster, with a mean polarity of 0.1465, whereas political, identity, visibility, and formation-related clusters generated more cautious sentiment. This interpretation is explicitly corroborated by our post hoc pairwise testing (Tables 17 and 18), which demonstrates that the positive valuation of developmental and reformist efforts stands in statistically significant opposition to the systemic skepticism directed at institutional constraints and geopolitical execution. This pattern is consistent with the idea that civilisational organisations may retain legitimacy through representation, solidarity, and development-oriented aspirations, even when their political and implementation capacities are questioned.

Theoretically, this pattern can be interpreted through the distinction between different sources of institutional legitimacy. Developmental and socioeconomic functions are more likely to generate positive evaluations because they are associated with practical cooperation, shared benefit, low sovereignty cost, and future-oriented institutional promise. In these areas, respondents can recognise value even when implementation remains incomplete, because the organisation is perceived as providing a platform for cooperation, visibility, and collective aspiration. Political-security functions, by contrast, operate in a more contested domain [25]. They require coordination among sovereign member states, credible enforcement capacity, diplomatic unity, and the ability to respond to conflict, minority rights, and geopolitical crises. These are high-expectation and high-risk areas in which institutional limitations become more visible. The weaker sentiment toward political-security roles and institutional constraints should therefore not be read simply as rejection of the organisation. Rather, it reflects a legitimacy gap between what the organisation symbolically represents and what it can operationally deliver in politically sensitive domains [26]. This distinction helps explain why respondents can simultaneously value the organisation's developmental, identity-based, and reform-oriented functions while remaining sceptical about its capacity for political coordination and security-related action.

Methodologically, the study contributes to computational social science by integrating multiple layers of analysis into a single reproducible framework. The pipeline combines answer-level segmentation, similarity-based thematic classification, lexical sentiment analysis, transformer-based contextual sentiment modelling, latent semantic clustering, non-parametric testing, bootstrap confidence estimation, and manual validation. This combination is important because qualitative interview data are often too context-dependent for simple automated scoring, while purely interpretive methods may not produce comparable indicators across groups and themes. The framework therefore provides a middle path: it preserves interpretive sensitivity while producing measurable outputs that can be compared statistically. The finding that sentiment differs significantly across themes and

semantic clusters, but not across respondent groups, also illustrates the value of modelling perception at multiple analytical levels.

Several limitations should be acknowledged. The study is based on 30 semi-structured interviews and 239 answer-level segments, which is appropriate for exploratory and theory-guided computational analysis of interview narratives but does not support strong population-level generalisation. Although segmentation increases analytical granularity, the segments are nested within interviews and should not be interpreted as fully independent observations; therefore, both segment-level and interview-level comparisons were reported, with the main expert–general result remaining non-significant under interview-level aggregation. The purposive sampling strategy may also introduce respondent selection bias, since participants were selected for their academic, professional, or institutional familiarity with the organisation. In addition, the analysis depends partly on automated lexical and transformer-based sentiment tools, which may misinterpret diplomatic language, indirect criticism, cautious praise, culturally embedded meanings, or mixed institutional evaluations, even though manual validation was used to reduce this risk. The transformer checkpoint used in this study was originally designed for social-media sentiment classification rather than interview-based institutional and political discourse. Its outputs are therefore treated as a contextual robustness layer rather than as a domain-fine-tuned gold standard, and future work should compare or fine-tune general-domain and political-domain sentiment models on larger institution-specific corpora. Finally, the findings should be interpreted as perception patterns within the analysed corpus and organisation, rather than as definitive estimates of wider public or expert opinion or as automatically generalisable to all civilisational, regional, or intergovernmental organisations. Future research should extend the framework using larger interview samples, broader respondent categories, additional country contexts, multilingual corpora, and hierarchical or mixed-effects modelling.

6. Conclusions

This study developed and demonstrated a sentiment-driven computational framework for modelling institutional perceptions in a 57-member civilisational intergovernmental organisation. Using 30 semi-structured interviews, the study produced a final analytical corpus of 239 answer segments and 42,940 respondent-generated words. The evidence shows that institutional perception is broadly positive but unevenly distributed. General respondents were slightly more positive than experts, but the difference was not statistically significant. The stronger and more meaningful differences appeared across institutional themes and latent semantic clusters, where reform, socioeconomic cooperation, and institutional effectiveness attracted more positive sentiment, while institutional weakness, political role, and security-related functions generated more cautious evaluations.

The central conclusion is that institutional perception in civilisational intergovernmental organisations is best understood as a multidimensional construct rather than as a simple sentiment outcome. The organisation examined in this study is perceived through a measurable tension between symbolic developmental legitimacy and operational scepticism. This has implications for both computational methodology and international organisation research. Computationally, it shows that interview narratives can be transformed into valid and interpretable sentiment, theme, and semantic-cluster level indicators. Conceptually, it shows that legitimacy is not only a matter of institutional performance, but also of how different organisational functions are affectively and cognitively evaluated by stakeholders.

A limitation concerns sample size and inferential generalisability. The study is based on 30 semi-structured interviews and 239 answer-level segments. This design is appropriate for exploratory and theory-guided computational analysis of interview narratives,

but it does not support strong population-level generalisation. The segment-level corpus increases analytical granularity, but the segments are nested within interviews and therefore should not be interpreted as fully independent observations. To reduce this concern, the study reports both segment-level and interview-level group comparisons, and the main expert-general result remains non-significant under interview-level aggregation. Even so, the findings should be interpreted as evidence of perception patterns within the analysed corpus rather than as definitive estimates of wider public or expert opinion. Future research should extend the design using larger interview samples, additional country contexts, and hierarchical or mixed-effects modelling to estimate respondent-level and segment-level variation more formally. Further extensions can proceed in three directions. First, larger multilingual corpora could be used to test whether similar perception structures appear across member states, regions, and language communities. Second, general-domain or political-domain transformer models can be compared and fine-tuned on institution-specific corpora to improve sensitivity to diplomatic, cultural, and politically cautious language. Third, the proposed framework can be applied comparatively to other international, regional, or identity-based multilateral organisations. Such extensions would allow researchers to examine whether the coexistence of symbolic legitimacy and operational scepticism is a broader feature of complex intergovernmental organisations.

Author Contributions: Conceptualisation, F.S., A.K.M.I.I. and A.A.; methodology, F.S.; software, F.S.; validation, A.A., F.S. and A.K.M.I.I.; formal analysis, F.S.; investigation, F.S.; resources, A.K.M.I.I.; data curation, F.S.; writing—original draft preparation, F.S.; writing—review and editing, F.S., A.K.M.I.I. and A.A.; visualization, F.S.; supervision, A.K.M.I.I.; project administration, A.K.M.I.I.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved on the 20th July 2023 by the Faculty of Business, Law and Politics Research Ethics Committee at the University of Hull (HU67RX United Kingdom). The original research title approved by the Ethics Committee was “Regionalism: OIC- A case study”. (Ethics approval letter attached as a supplementary document).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: Due to platform terms, privacy considerations, and the potential sensitivity of conflict-related discourse, the full tweet text and user-level metadata are not publicly released. Aggregated statistics, derived construct counts, and non-identifying analytical outputs may be made available upon reasonable request. Any shared data will exclude direct user identifiers and will follow privacy-preserving procedures.

Acknowledgments: The authors express their sincere gratitude to the Faculty of Business, Law and Politics, University of Hull, for its oversight and ethical approval of this research (HU67RX, United Kingdom). Special appreciation is extended to the Department of Management, University of Dhaka, for facilitating access to respondents and for logistical support during the survey administration phase. The authors acknowledge the invaluable contributions of participating students whose informed and voluntary responses enabled the empirical foundation of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
CI	Confidence Interval
IP	Institutional Perception

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
OIC	Organisation of Islamic Cooperation
RoBERTa	Robustly Optimized BERT Pretraining Approach
TF-IDF	Term Frequency–Inverse Document Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner

References

1. Tallberg, J.; Zürn, M. The Legitimacy and Legitimation of International Organizations: Introduction and Framework. *Rev. Int. Organ.* **2019**, *14*, 581–606. <https://doi.org/10.1007/s11558-018-9330-7>.
2. Dellmuth, L.M.; Tallberg, J. The Social Legitimacy of International Organisations: Interest Representation, Institutional Performance, and Confidence Extrapolation in the United Nations. *Rev. Int. Stud.* **2015**, *41*, 451–475. <https://doi.org/10.1017/S0260210514000230>.
3. Dellmuth, L.M.; Scholte, J.A.; Tallberg, J. Institutional Sources of Legitimacy for International Organisations: Beyond Procedure versus Performance. *Rev. Int. Stud.* **2019**, *45*, 627–646. <https://doi.org/10.1017/S026021051900007X>.
4. Steffek, J. Triangulating the Legitimacy of International Organizations: Beliefs, Discourses, and Actions. *Int. Stud. Rev.* **2023**, *25*, viad054. <https://doi.org/10.1093/isr/viad054>.
5. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135.
6. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>.
7. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
8. Grimmer, J.; Stewart, B.M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Anal.* **2013**, *21*, 267–297. <https://doi.org/10.1093/pan/mps028>.
9. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 3–5 June 2019; pp. 4171–4186.
11. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
12. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794.
13. Isoaho, K.; Gritsenko, D.; Mäkelä, E. Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Stud. J.* **2021**, *49*, 300–324. <https://doi.org/10.1111/psj.12343>.
14. Sufi, F.K. Identifying the Drivers of Negative News with Sentiment, Entity and Regression Analysis. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100074. <https://doi.org/10.1016/j.jjime.2022.100074>.
15. Sufi, F.K. A New Computational Method for Quantification and Analysis of Media Bias in Cybersecurity Reporting. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 4561–4570. <https://doi.org/10.1109/TCSS.2025.3560199>.
16. Barbieri, F.; Camacho-Collados, J.; Espinosa-Anke, L.; Neves, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 16–20 November 2020; pp. 1644–1650. <https://doi.org/10.18653/v1/2020.findings-emnlp.148>.
17. Mann, H.B.; Whitney, D.R. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60.
18. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.
19. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall: London, UK, 1993.
20. Patton, M.Q. *Qualitative Research and Evaluation Methods*, 3rd ed.; Sage: Thousand Oaks, CA, USA, 2002.
21. Creswell, J.W.; Poth, C.N. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, 4th ed.; Sage: Thousand Oaks, CA, USA, 2018.
22. Morse, J.M. The Significance of Saturation. *Qual. Health Res.* **1995**, *5*, 147–149.
23. Guest, G.; Bunce, A.; Johnson, L. How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods* **2006**, *18*, 59–82.
24. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.

25. Dellmuth, L.; Tallberg, J. Public Opinion and International Organizations. *Rev. Int. Organ.* **2026**. <https://doi.org/10.1007/s11558-025-09614-2>.
26. Tørstad, V. Can Transparency Strengthen the Legitimacy of International Institutions? Evidence from the UN Security Council. *J. Peace Res.* **2024**, *61*, 228–245 <https://doi.org/10.1177/00223433221123996>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.