POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

Uncovering latent themes and emotional dynamics in youth precariousness

Laurea Magistrale in Mathematical Engineering - Statistical Learning

Author: Francesco Tomasi

Advisor: Alessandra Guglielmi

Co-advisor: Lara Maestripieri

Academic year: 2024-2025

---

1. Introduction

Precariousness has increasingly been recognized as a multidimensional phenomenon that extends beyond the labour market to significantly impact the living conditions and psychological well-being of younger generations (Vosko et al., 2009). This dynamic is particularly evident in Spain, where a dual labour market characterized by high temporary youth employment creates a cumulative disadvantage (Bolíbar et al., 2019). For these young individuals, job insecurity is not merely an economic status, it is a profound subjective experience defined by uncertainty, vulnerability, and a loss of control.

To investigate this subjective dimension, this thesis analyzes open-ended responses from the VulnYouth project (IGOP, 2023), focusing on how Spanish youth aged 20 to 34 answer the question: "What does precariousness mean to you?" Unlike predefined survey choices, this unstructured text allows for a nuanced understanding of the phenomenon, capturing the emotional and cognitive burden of instability that standard metrics often miss.

This thesis aims to identify the latent themes and emotions in youths' descriptions of precariousness. To interpret these narratives, the study adopts a dual computational strategy. First, Structural Topic Modeling (STM) is used to uncover the latent thematic structures of the discourse—such as material deprivation, institutional exclusion, or future uncertainty. Second, Emotion Detection is employed to capture the underlying affective states, shifting the focus from what respondents discuss to how they feel about it. This is achieved using transformer-based architectures: a fine-grained categorical classification to detect 28 discrete emotions using RoBERTa fine-tuned on the GoEmotions dataset (Demszky et al., 2020), and a continuous dimensional assessment of Valence, Arousal, and Dominance using a BERT model trained on the EmoBank dataset (Buechel & Hahn, 2017).

2. State of the art

A substantial body of research seeks to establish a clear link between occupational instability and negative mental health outcomes. The VulnYouth project, from which the data for this study are drawn, was specifically designed to explore these interconnected issues among young Spaniards. Standard survey metrics, however, often fail to capture the full emotional and cognitive burden of these experiences. Therefore, this study relies on open-ended survey responses to allow for a broader and more nuanced understanding of how individuals subjectively perceive precariousness.

Analyzing these open-ended responses presents two primary computational challenges. First, the responses consist of very short texts, which represents a known difficulty in Natural Language Processing (NLP) because traditional models often lack sufficient context to perform reliably. Second, the corpus is almost exclusively focused on negative experiences. Unlike general sentiment analysis tasks, the analytical challenge here is to make fine-grained distinctions between various negative states rather than simply classifying text as positive or negative.

To address these challenges, this study first utilizes Structural Topic Modeling (STM) to detect latent thematic structures. As an extension of Latent Dirichlet Allocation (LDA), STM incorporates document-level metadata directly into the model's prior distributions. This allows the model to formally estimate how the prevalence of specific topics varies according to the respondents' socio-economic attributes.

Second, a dimensional emotion model is applied using a Bidirectional Encoder Representations from Transformers (BERT) architecture. This approach maps emotions into a continuous vector space utilizing the VAD framework. It defines emotional states through three coordinates: Valence (the degree of pleasure or displeasure), Arousal (the physiological and psychological intensity, from calm to excited), and Dominance (the perceived level of control or agency over a situation).

Finally, a categorical emotion classification is conducted using a Robustly Optimized BERT Pretraining Approach (RoBERTa). By fine-tuning the model on the GoEmotions dataset, it is possible to resolve ambiguities in short texts and identify distinct emotional states. This provides a highly granular taxonomy of 28 distinct categories: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, and a neutral category.

3. Data description and pre-processing

The dataset used in this study is drawn from the VulnYouth project (Studying Intersectional Vulnerability of Precarious Youth in Spain). It consists of 3012 interviews conducted in February 2023 with Spanish youth aged 20 to 34. The sample is a non-probabilistic quota sample, stratified by demographic factors, and contains indicators regarding economic status, employment, education, and mental health.

The primary focus of this analysis is the open-ended survey question: "What does precariousness mean to you?". The original Spanish responses were translated into English to leverage advanced natural language models. A total of 2487 valid, non-empty text answers were extracted. The text data is notably brief, averaging 19.25 words per response with a standard deviation of 22.06. This brevity, combined with a corpus heavily skewed toward negative sentiment, poses a significant analytical challenge.

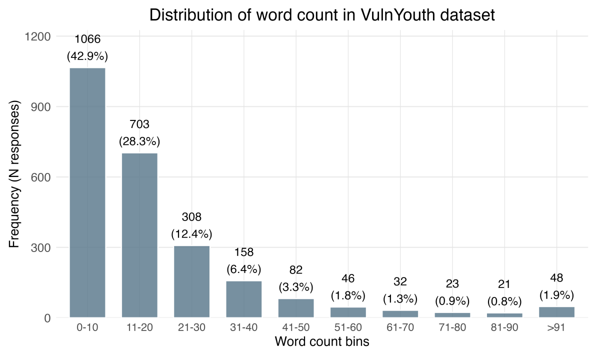

Figure 1: Distribution of word count in $V54\_PRECARIOUSNESS$ (VulnYouth). Data refers to the $N = 2487$ valid answers, translated from Spanish to English.

3.1. Text processing pipeline

Two distinct pre-processing pipelines were implemented using different programming environments to optimize for the chosen algorithms.

R was utilized for the Structural Topic Model (STM) to leverage the native `stm` package. This pipeline included five distinct pre-processing steps:

1. Preserving specific punctuation: Hyphens were retained to treat complex concepts as single units.

2. Stop word removal: Common English stop words and custom terms ("precariousness", "precarious") were filtered out.
3. Stemming: Words were reduced to their root forms.
4. Removal of punctuation and numbers: All other punctuation and numeric characters were stripped.
5. Lowercasing: All text was converted to lowercase.

Conversely, Python was selected for the BERT and RoBERTa classifiers to exploit deep learning libraries (PyTorch/TensorFlow) and GPU acceleration. Because Transformer models rely on full linguistic structure and context, a less destructive strategy was applied:

- Lowercasing: Converting all text to lowercase.
- Noise removal: Stripping HTML tags, URLs, emojis, and special symbols.
- Whitespace normalization: Consolidating multiple spaces while preserving essential punctuation.

3.2. Covariates used in the study

The Structural Topic Model incorporated document-level metadata to estimate how topic prevalence varied according to respondents' attributes. For all of these variables, following Maestripieri et al. (2024), a transformation was applied to convert them into binary indicators in order to isolate cases of acute vulnerability from the rest of the sample. The binary socio-economic covariates included in the model are economic independence from the family of origin (ecoind), perception of high-intensity job insecurity (jobinsec), acute economic strain, measured by difficulty in making ends meet (endmeet), perception of social pressure and acute perceived precarity (normepres), individual financial capacity to handle unexpected expenses (unexpmix).

Additional socio-demographic variables, such as gender, age group, and migration background, were utilized later in the study for logistic regression analysis of emotional expression.

3.3. Methodology

This study integrates unsupervised statistical learning with supervised deep learning to provide a comprehensive analysis of the unstructured text.

3.3.1 Structural Topic Modeling (STM)

To identify what respondents were discussing, the study employed Structural Topic Modeling. STM is a probabilistic generative model of word counts that extends Latent Dirichlet Allocation (LDA). While standard LDA assumes documents are independent, STM allows document-level covariates to influence the prior distributions of topic prevalence.

Generative Process The model assumes a corpus of D documents with a fixed number of topics K. The core component used in this analysis is:

- Topic Prevalence (θ_d): The proportion of document d allocated to each topic. In STM, this vector is drawn from a Logistic Normal distribution conditional on the document covariates X_d:

$$\theta_d | X_d\gamma, \Sigma \sim \text{LogisticNormal}(\mu = X_d\gamma, \Sigma) \quad (1)$$

  Here, X_d is a vector of document-level metadata (e.g., Gender, Economic Independence, Job Insecurity). This formulation allows the model to estimate not just the content of the topics, but how their prevalence varies across the groups defined by X_d.

Model Selection The optimal number of topics was fixed at $K = 20$. This decision ensures methodological consistency and allows for a direct comparison with the baseline topics obtained by Di Liberatore (2024). This specific configuration was identified by evaluating diagnostic metrics to maximize Semantic Coherence (the tendency of top words to co-occur within the same documents) while balancing Exclusivity (how distinct the topics are from one another).

3.3.2 Transformer-based emotion detection

To capture how respondents felt, the research utilized a Bidirectional Encoder Representations from Transformers (BERT), a deep learning model based on the Transformer architecture. This model has become popular due to its ease of

application in various fields, and this flexibility is achieved through so-called transfer learning: instead of starting from scratch, which would be expensive and require too much data, it uses parameters obtained during a pre-training phase, where the model learns the English language from massive corpora. An output layer is then added to adapt the new model to specialize in a specific task.
Moreover, as all the Transformer Architecture, it relies on a Self-Attention mechanism, which analyzes the entire sequence of words simultaneously rather than sequentially. This allows the model to calculate the relationship of every word to every other word, effectively resolving ambiguities in short, unstructured texts by leveraging global context.

Categorical analysis: RoBERTa For discrete emotion classification, a RoBERTa-base (Robustly optimized BERT approach) model was fine-tuned. RoBERTa was selected over standard BERT due to its superior empirical performance on multi-label classification benchmarks.

- Dataset: The model was fine-tuned on the GoEmotions dataset, which provides a high-granularity taxonomy of 27 distinct emotional labels—*admiration, amusement, approval, caring, curiosity, desire, excitement, gratitude, joy, love, optimism, pride, realization, relief, surprise, anger, annoyance, confusion, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness*— plus a *neutral* category.
- Output: The final output is a probability vector across all 28 classes, enabling the simultaneous detection of complex, mixed emotional states (e.g., a response expressing both sadness and disappointment).

Dimensional analysis: BERT For continuous emotional mapping, a standard BERT-base model was fine-tuned as a multi-output regression task using the EmoBank dataset. This approach maps emotions into a continuous 3-dimensional vector space, with each dimension ranging from 1 to 5:

- Valence: The degree of pleasantness or unpleasantness (Positive vs. Negative).
- Arousal: The physiological and psychological intensity, ranging from high activation (e.g., anger) to low activation (e.g., calm or resignation).
- Dominance: The perceived degree of control or agency the speaker feels over the situation.

3.4. Model configuration

To analyze the corpus of 2487 valid open-ended responses, the study required specific configuration strategies for both the topic modeling and the deep learning architectures.

3.4.1 STM model configuration

To identify the optimal latent structure, two distinct Structural Topic Models were estimated:

- Base Model: An unsupervised model with no covariates, replicating the exact configuration adopted in previous research (Di Liberatore, 2024) to serve as a semantic baseline.
- Covariate Model: A model incorporating document-level metadata to estimate topic prevalence. This model included five binary socio-economic covariates: economic independence (`dum_ecoind`), perception of job insecurity (`dum_jobinsec`), difficulty in making ends meet (`dum_endmeet`), perception of social pressure (`dum_normepres`), and the ability to handle unexpected expenses (`dum_unexpmix`).

The number of topics was fixed at 20. This decision ensured methodological consistency with the baseline model, allowing for direct comparison. Furthermore, this parameter was selected based on diagnostic metrics to maximize Semantic Coherence (how often top words co-occur) while balancing Exclusivity (how distinct the topics are from one another).

3.4.2 Transformer fine-tuning strategy

To adapt the pre-trained Transformer models for emotion detection, a fine-tuning approach was implemented using the Hugging Face Transformers library.

- Categorical analysis (RoBERTa on GoEmotions): The model was trained for 8 epochs with a batch size of 32 and a learning rate of 2e-5. To address class imbalance and prevent the model from defaulting to the majority "neutral" class, a MultiLabel Focal Loss was implemented. Following the training phase, a Dynamic Thresholding strategy was applied to calculate and select the specific probability cut-off that maximized the F1-score independently for each of the 28 emotions, thereby converting the probability output into a binary classification (1 if the emotion is present, 0 otherwise).

- Dimensional analysis (BERT on EmoBank): For the continuous mapping of emotions, the model was trained for 8 epochs (batch size 32, learning rate 2e-5) as a multi-output regression task. The network minimized the Concordance Correlation Coefficient (CCC) Loss, which penalizes predictions that deviate from original observations in both variance and mean. The optimal model was selected by monitoring the average CCC score across the Valence, Arousal, and Dominance dimensions on the validation set.

4. Results

The application of the computational methods to the 2487 valid open-ended responses reveals clear linguistic, emotional, and thematic patterns regarding how Spanish youth experience precariousness.

Descriptive analysis The text data is highly brief, averaging just 9 tokens per document after processing. The most frequent stems in the corpus are *work* (1,083 occurrences), *abl* (734) (being able or ability), *live* (637), *job* (501), *salari* (480) (salary or salaries) , and *condit* (461)(condition). The dominance of *work*, *job*, and *conditions* signals that the labour market is the primary aspect of precariousness. Moreover, the high frequency of *able* - stemmed in *abl* - and *live* points out that precarity is defined by the inability to sustain a life independent of family support.
Responses are generally brief, with an average of 9 tokens per document after processing. This brevity leads to a highly sparse *Document-Term Matrix*. We have found an average cosine similarity of 0.07 between responses.

Dimensional emotion analysis Mapping the texts into a continuous Valence-Arousal-Dominance (VAD) we find out that the discourse is conceptualized almost exclusively through negative sentiment, with a mean Valence of 2.56. The low intensity score (Arousal mean of 2.95) places the discourse near the neutral threshold, reflecting a muted, resigned affect rather than active agitation. Furthermore, the low Dominance score (2.73 mean) indicates respondents perceive precarity as a state controlled by external forces, lacking personal agency. Notably, female respondents narrate their experience with a statistically significant lower level of dominance (less agency) and slightly higher intensity (arousal) compared to men.

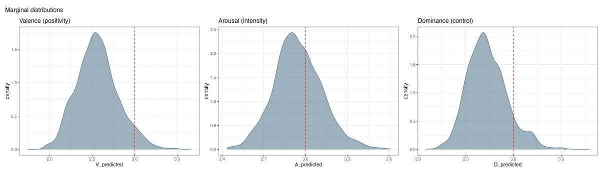

Figure 2: Marginal distributions of emotional dimensions. The red dashed line indicates the neutral score of 3.

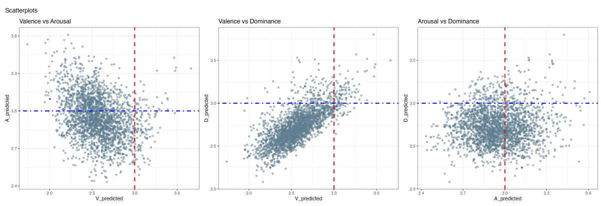

Figure 3: Pairwise scatterplots of the predicted emotional dimensions (Valence, Arousal, and Dominance) for the VulnYouth corpus. The red and blue dashed lines represent the neutral states

Categorical emotion analysis The RoBERTa model successfully categorized the specific emotional labels, showing a landscape dominated by resignation rather than active conflict. Disappointment (51.2%) and Sadness (48.8%) are the most prevalent emotions, frequently co-occurring to form a complex state of "sad resignation" over failed expectations. High-arousal negativity is practically absent: active states like Anger (0.5%) and Fear (1.6%) appear with negligible frequency, reinforcing the finding that the prevailing mood is one of powerlessness.

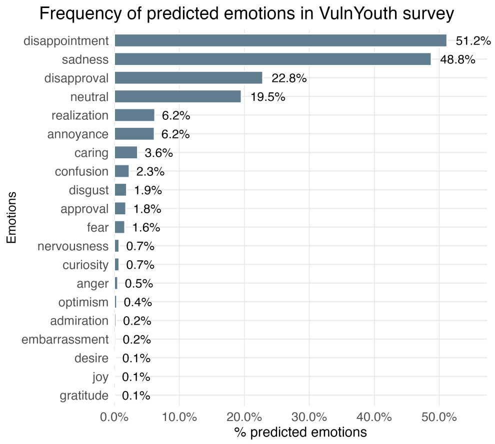

Figure 4: Frequency of predicted emotions in the open-ended questions of precariousness in the VulnYouth survey.

Dividing the 28 emotions in 4 sentimental groups To synthesize the complex emotional landscape, the 28 categorical emotional labels were aggregated into four high-level sentiment groups based on their affective valence and composition:

- Positive group: composed of 15 emotions associated with favorable or constructive states (*admiration, amusement, approval, caring, curiosity, desire, excitement, gratitude, joy, love, optimism, pride, realization, relief, surprise*).
- Negative group: composed of 12 emotions associated with distress or aversion (*anger, annoyance, confusion, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness*).
- Ambivalent group: containing responses that simultaneously exhibit at least one positive and one negative emotion, reflecting a complex or mixed emotional state.
- Neutral group: containing responses classified exclusively as "neutral" or lacking any specific emotional marker.

The Negative group clearly dominates the affective narratives identified in the open-ended responses, confirming that the subjective conceptualization of precariousness is overwhelmingly rooted in distress.

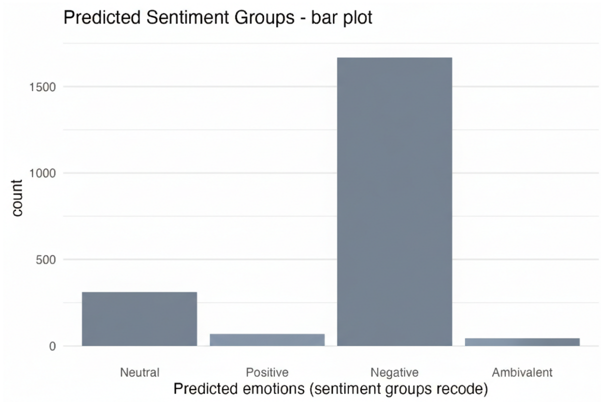

Figure 5: Distribution of aggregated sentiment groups. Note the dominance of negative sentiment and the significant minority of positive and neutral responses.

The neutrality paradox In the previous framework, to formally test whether sociodemographic and material vulnerability factors are associated with emotional expression, we estimated two binary logistic regression models. Respondents were coded as emotional (1 = assigned to Positive, Negative, or Ambivalent sentiment groups) versus neutral (0 = exclusively Neutral category).

The first model includes gender, age group, and migration background as categorical predictors of the emotional expression, and reveals that women have approximately 36% higher odds of expressing emotional language compared to men.

The second examined objective vulnerability indicators —normative employment status, job insecurity, difficulty making ends meet, unexpected financial strain, and economic independence— revealing that job insecurity is the only statistically significant predictor of emotional expression. Contrary to expectations, respondents experiencing high job insecurity have approximately 25% lower odds of being classified as emotional compared to those with stable employment. Other vulnerability indicators showed no significant effect. This neutrality paradox indicates that individuals facing higher structural instability are more likely to adopt linguistically neutral definitions of precariousness.

Topic analysis The Structural Topic Model identified 20 stable latent themes that capture the core narratives of precariousness. Incorporating material precarity indicators (such as economic strain and job insecurity) maintained the thematic boundaries of the corpus. The topics were defined in (Di Liberatore, 2024) and are reported below:

ID	Label (Di Liberatore, 2024)
T1	Resource-driven economic insecurity
T2	Unaffordable independent living
T3	Poor working conditions
T4	Worker exploitation and vulnerability
T5	Overtime without proper compensation
T6	Illegal and unsafe working conditions
T7	Not being valued in the workplace
T8	No access to housing or food
T9	Inability to meet living standards
T10	Impossibility to save money
T11	Living an unstable life
T12	Impossibility to plan a future
T13	Struggle to make ends meet
T14	Lack of resources
T15	Exploitation and Abusive Treatment
T16	Not being able to afford basic needs
T17	Impossibility of reaching economic independence
T18	Low wages
T19	Youth unemployment and insecurity
T20	Job insecurity and insufficient stability

Table 1: The 20 latent topics of precariousness in the VulnYouth corpus from Di Liberatore (2024).

Emotional predictors of topic prevalence
Beyond identifying the 20 latent themes, we examine through a logistic regression, whether specific predicted emotions are systematically associated with the probability of discussing particular topics.

The results reveal differentiated thematic–emotional patterns within the corpus, although their strength varies across emotions. *Sadness* is significantly predicted by a cluster composed of Topic 9 and Topic 13. For *Disappointment*, the most relevant predictors are Topic 14, Topic 13, and Topic 6. *Disapproval* is primarily characterized by Topic 3 and *Neutral* category is significantly associated with Topic 5 and Topic 18. Across these last three categories, the effect sizes and Area Under the Curve (AUC) values remain in a moderate range, indicating that while the associations are statistically reliable, these emotions are thematically diffuse and not tightly concentrated around a single structural condition.

In contrast, *Fear* shows the strongest effect sizes among the emotions considered. It is predicted by a combination of Topic 2, Topic 1, and Topic 3. However, this result must be interpreted with caution, as fear represents a very small proportion of the corpus.

5. Conclusion

This thesis has investigated the subjective experience of precariousness among Spanish youth by applying advanced Natural Language Processing techniques to open-ended survey responses from the *VulnYouth* study. While traditional sociological research often relies on predefined indicators of job insecurity, this study aims to uncover the latent themes and emotional undercurrents that young people use to define their own condition by integrating Structural Topic Modeling (STM) with Transformer-based emotion detection (BERT and RoBERTa).

A primary finding of this research is the specific emotional signature of precariousness in Spain. Contrary to the expectation that high job insecurity might fuel *Anger* (0.5%) or *Fear* (1.6%), our categorical emotion analysis (using RoBERTa) reveals a landscape dominated by *Disappointment* (51.2%) and *Sadness* (48.8%).

By aggregating the 28 emotional labels into positive, negative, mixed, and neutral categories, this study identifies a strange sociological phenomenon: the Neutrality Paradox. Logistic regression analysis reveals that respondents facing the highest levels of objective job insecurity and economic strain are significantly more likely to adopt neutral, detached language rather than employing emotional descriptor.

Ultimately, the Structural Topic Model identified 20 latent themes that corroborate the multidimensional nature of precariousness, and to move beyond descriptive mapping, this thesis examined whether specific emotions are statistically associated with the probability of discussing particular themes.

Methodological Impact
Transformer architectures (RoBERTa and BERT) proved highly effective at detecting nuanced emotional states in short, sparse sociological survey texts where traditional lexicon-based approaches fail. Furthermore, integrating Structural Topic Modeling (STM) with socio-economic covariates enabled the identification of latent themes within the open-ended responses, allowing for a more structured

and effective categorization of the unstructured text.

References

[1] Mireia Bolibar. Macro-structural determinants of precarious employment: A combined analysis of european labor markets. *International Sociology*, 35(1):65–89, 2019.

[2] Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, 2017. Association for Computational Linguistics.

[3] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, 2020. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[5] IGOP, Institut de Govern i Polítiques Públiques. Vulnyouth project, 2023.

[6] Elena Di Liberatore. Precariousness and mental health among spanish youth. Master's thesis, Politecnico di Milano, 2024. Advisor: Alessandra Guglielmi, Co-advisor: Lara Maestripieri.

[7] L. Maestripieri, A. Lanau, R. Soler-i Martí, and M. Acebillo-Baqué. Intertwined precariousness and precarity: Disentangling a phenomenon that characterizes spanish youth. *International Journal of Social Welfare*, 2024.

[8] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[10] Leah F. Vosko, Martha MacDonald, and Iain Campbell. *Gender and the Contours of Precarious Employment*. Routledge, London; New York, 2009.