"For Italians only": exploring discriminative behavior in LLMs in real estate context

Enrique Taietta
Matricola 257230

Data Science, University of Trento
13/06/2025

1 Introduction

This research originates from a personal experience: during a phone call with a real estate agency, after expressing interest in a property and stating my full name, I was told, "I'm sorry, but we are only looking for Italians." This raises the central question of the study: *How did the real estate agent determine that I was not 'Italian,' despite my native-level fluency in the language?*

The core assumption of this work is that a person's name can serve as a powerful in-group/out-group marker. The primary objective is to investigate the presence of bias in large language models (LLMs) by applying them to a real-life scenario—namely, the search for rental housing. Specifically, the study examines whether, and to what extent, such models reflect or amplify existing forms of social discrimination, particularly based on names, occupations, or other personal attributes.

This study adopts a multidisciplinary theoretical framework, integrating perspectives from sociology, urban studies, and cognitive science. From a sociological standpoint, LLMs are explored as potential tools for the analysis of social phenomena, due to their flexibility across application contexts, their ability to process complex contextual information, and their dependence on real-world data—factors that shape their training, performance, and outputs. The investigation focuses on how names perceived as foreign influence the model's response within a housing context, for example, in simulated interactions between landlords and potential tenants. Existing literature highlights how foreign-sounding names can activate stereotypes and lead to implicit or explicit forms of discrimination.

Simultaneously, the urban dimension is examined to assess whether LLMs associate personal characteristics—such as name, profession, or age—with specific neighborhoods, thereby reproducing patterns of urban segregation or discrimination. The study seeks to uncover the latent meanings and semantic implications carried by proper names, analyzing whether associative biases exist within LLMs and how these biases manifest in particular contexts—such as when a name is paired with references to the user's political ideology.

A second theoretical axis concerns cognitive science and behavioral psychology. In the second phase of the experiment, we analyze simulated tenant responses (generated via LLMs) following positive or negative interactions with landlords. This phase aims to explore the psychological and emotional impact of discriminatory treatment—an underexplored topic in the scientific literature, particularly in relation to variables such as age, gender, and situational context. We investigate whether the subjective perception of discrimination changes based on the nature of the response and the type of bias encountered, especially from a linguistic and emotional standpoint. Understanding how LLMs capture or reflect such experiences may lay the groundwork for future research on behavioral patterns.

Finally, on the cognitive-linguistic and semantic levels, the study aims to observe how key concepts such as *"I"* and *"you"*—central to social psychology's theory of in-group/out-group dynamics—are articulated in the model's responses. Special attention is given to how meaning propagates through the semantic network, comparing texts that reflect discriminatory behavior with those written from the perspective of individuals experiencing discrimination.

This study is especially relevant in the Italian context, where the worsening housing crisis, rising political extremism, and increased migration flows intersect with ongoing debates about nationality and citizenship. These sociopolitical factors amplify the societal relevance of examining how biases may be embedded in or reproduced by artificial intelligence systems.

2 Literature Review

Discrimination in the housing market is a well-documented phenomenon, with numerous factors contributing to discriminatory behavior. Names, in

this context, do not only convey perceived ethnicity but also socio-economic class and religious affiliation—dimensions that can all trigger prejudice and bias. [1]

As in human interactions, names play a crucial role in large language models (LLMs), serving as carriers of implicit information. This becomes particularly relevant when negative biases are triggered solely by a name during interactions with an LLM across different contexts. [2]

Concerns about the presence of bias in LLMs are widespread. Despite growing evidence that these models encode and sometimes amplify societal biases embedded in their training data, little structural action has been taken by model providers. Many LLMs remain opaque black boxes, and calls for greater transparency regarding their training datasets have gone largely unanswered. The lack of visibility into these data pipelines increases the risk of perpetuating social inequalities in various domains. [3, 4]

The literature offers numerous examples of how AI systems reproduce various forms of discrimination. Gender bias, for instance, has been observed in word embeddings [5], while dialect-based discrimination has also been documented. [6]

Given this context, it is critical to recognize potential biases in domains already marked by structural discrimination, such as the housing market. Especially in periods of heightened political and ideological polarization, the integration of AI into real estate decision-making processes risks scaling up biased or stereotyped outcomes. This can contribute to intensified forms of urban segregation. [7]

2.1 Research Questions and Hypotheses

RQ1: How is ethnic discrimination influenced by the interplay between names and ideological context?

- **H1A**: The presence of ethnic discrimination in landlords' responses is dependent on the tenant's name.
- **H1B**: The presence of ethnic discrimination in landlords' responses is dependent on the landlord's expressed ideology.

RQ2: How does class-based or urban discrimination vary based on the tenant's city and neighborhood of reference?

- **H2A**: Class-based or urban discrimination is associated with the specific city-neighborhood combination.

RQ3: How does class or urban discrimination vary based on the tenant's profession?

- **H3A**: There is a correlation between the tenant's profession and the occurrence of class-based discrimination.

RQ4: How does the presence of discrimination vary with the tenant's age?

- **H4A**: There is a correlation between the tenant's age and the presence of age-related discrimination.

3 Project Design

The research was conducted through a multi-phase design:

- Collection of responses from both *owners* and *tenants* using LLM impersonation;
- Labelling of owner responses concerning different types of discrimination;
- Labelling of both owner and tenant responses based on emotional content.

At the end of the first phase, a dataset was constructed in which each record contains an owner's reply to a rental request initiated by a tenant, as well as the tenant's subsequent reflection on the owner's response.

In the second phase, each record was further enriched with a set of labels corresponding to five discrimination categories under investigation.

In the third and final phase, the emotional content statistically significant of both owner and tenant responses was collected.

3.1 Data Collection Strategy

To support the first phase, a structured table was created, listing all variables to be included in the prompts. The LLM selected for the experiment was `mistral-small`, provided by Mistral. Since one of the study's objectives was to examine the model's internal semantic behavior, part of the variable selection was guided by the model itself through exploratory prompting.

The variables incorporated into the prompts were:

- **Owner-related variables:**
  - Gender
  - Age

	AGE	CITY	POL. OR.	NAME	JOB	AGE
M	<60	MILAN;BRERA	FAR RIGHT	Sara Esposito	Teacher	<35
F	>60	MILAN;BAGGIO	FAR LEFT	Maria Ionescu	Caregiver	>35
		MILAN;CITTÀ STUDI	UNDEFINED	Fatima El-Khayari	Student	
		BOLOGNA;COLLI		Wang Mei		
		BOLOGNA;BORGO PAN.		Awa Ndiaye		
		BOLOGNA;UNIV. AREA		Anna Meyer		
		ROME;PARIOLI				
		ROME;CENTOCELLE				
		ROME;SAN LORENZO				
		NAPLES;POSILLIPO				
		NAPLES;PONTICELLI				
		NAPLES;FUORIGROTTA				

Table 1: *Example table of variable combinations used for prompt generation (female names).*

  - City and Neighborhood
  - Political Orientation

- **Tenant-related variables:**
  - Name
  - Job
  - Age

To construct the name variable for tenants, the model was queried about the most represented ethnic groups in Italy and the most common first and last names associated with each group, in both masculine and feminine forms. Based on its output, four non-Italian names, one Italian name, and one European (German) name were arbitrarily selected.

A similar process was used for identifying occupations and neighborhoods: the model was asked to list the most common jobs by ethnicity, and for each of several cities, to identify the wealthiest, most peripheral, and university-affiliated neighborhoods.

Also, in the case of a combination with tenant age variable, as less than 35 years old, another sentence was added to the prompt for the tenant response, specifying that she or he was born and raised in Italy, to search for changes in the range of emotion.

The final result was a variable matrix used to generate all prompt combinations. Table 1 illustrates a sample of this table focusing on female names.

From the cross-product of variables, 5,184 unique prompt combinations were generated. For each, an impersonated owner response was created, followed by a second prompt to elicit the corresponding tenant response. Including both male and female tenant names, the total number of interactions reached 20,736.

3.2 Labelling discriminatory behavior and Emotional state

In the second phase, each owner response was analyzed using the GPT-4o-mini-2024-07-18 model via OpenAI's API. A dedicated prompt using the `text_format` option was designed to classify responses across five types of discrimination: urban, ethnic, gender, age-based, and class-based. The model was also asked to indicate whether the rental offer remained open, closed, or open with positive/negative connotations, as well as the number of explicit requests made by the owner.

In the final phase, emotional profiling was conducted using the EmoAtlas library [8]. Each response was analyzed according to Plutchik's model of primary emotions. Statistically significant emotions were identified using z-scores. Initially, the threshold was set at $|1.96|$, but was later reduced to $|1.64|$ to account for the model's strong tendency to express highly positive emotions, which skewed the overall emotional balance. The lower threshold was adopted for exploratory purposes, considering that LLM-generated text may not reflect real-world emotional variance. Then, for the detected primary emotions, complex emotions were found due to their combinations in diads.

4 Results

Note: Tables from 3 to 21 are at the end of document

The exploratory data analysis revealed a high incidence of *class* and *ethnic* discrimination, followed by *urban* and *ageism* discrimination. Instances of *gender* discrimination were minimal (see Figure 3).

Regarding *ethnic* discrimination, significant variation emerged based on the tenants' names, for both male

and female cases (Figures 4–5). Similarly, discrimination rates varied across city neighborhoods, although general distribution patterns were broadly consistent across the four urban contexts (Figures 6–9).

A series of chi-squared tests were conducted to validate several hypotheses concerning relationships between demographic or contextual variables and types of detected discrimination behavior.

4.1 H1A: Ethnic Discrimination and Tenant Names

- **Null hypothesis** H_0: Tenant names and the occurrence of ethnic discrimination are statistically independent.
- **Alternative hypothesis** H_1: Tenant names and the occurrence of ethnic discrimination are not statistically independent.

A contingency table was constructed using tenant names and the presence/absence of ethnic discrimination in the owner's response. The chi-squared test allowed rejection of H_0, indicating a significant dependency.

Chi²: 1388.87 - **p-value**: 3.04×10^{-291} - **Degrees of freedom**: 11

4.2 H2A: Ethnic Discrimination and Owner Ideology

- **Null hypothesis** H_0: The owner's political orientation and the occurrence of ethnic discrimination are statistically independent.
- **Alternative hypothesis** H_1: The owner's political orientation and the occurrence of ethnic discrimination are not statistically independent.

Again, the null hypothesis was rejected, confirming a strong association between political ideology and discriminatory behavior.

Chi²: 2822.21 - **p-value**: $= 0$ - **Degrees of freedom**: 2

4.3 H2B: Spatial Variables and Class/Urban Discrimination

A chi-squared test examined the association between *city+neighborhood* combinations and the incidence of *class discrimination*.

- **Null hypothesis** H_0: Neighborhood and class discrimination are statistically independent.
- **Alternative hypothesis** H_1: Neighborhood and class discrimination are not statistically independent.

Chi²: 42.70 - **p-value**: 1.22×10^{-5} - **Degrees of freedom**: 11

In contrast, no significant association was found with respect to *urban discrimination*:

Chi²: 18.54 - **p-value**: 0.070 - **Degrees of freedom**: 11

City-level analysis showed significant results only for Milan and Rome:

- Milan — Chi²: 15.36, p-value: 4.61×10^{-4}, dof: 2
- Rome — Chi²: 9.77, p-value: 0.0075, dof: 2
- Naples — Chi²: 2.85, p-value: 0.240, dof: 2
- Bologna — Chi²: 0.077, p-value: 0.962, dof: 2

4.4 H3: Discrimination and Tenant Occupation

- **Null hypothesis** H_0: Tenant occupation and the occurrence of discrimination are statistically independent.
- **Alternative hypothesis** H_1: Tenant occupation and the occurrence of discrimination are not statistically independent.

Regarding *class discrimination*, the null hypothesis was rejected:

Chi²: 208.89 - **p-value**: 3.56×10^{-43} - **Degrees of freedom**: 5

However, no significant association was observed with *urban discrimination*:

Chi²: 7.09 - **p-value**: 0.214 - **Degrees of freedom**: 5

Thus, hypothesis H3 is only partially supported.

4.5 H4A: Ageism and Tenant Age

- **Null hypothesis** H_0: Tenant age and the occurrence of ageism are statistically independent.
- **Alternative hypothesis** H_1: Tenant age and the occurrence of ageism are not statistically independent.

The test confirmed a statistically significant association:

Chi²: 46.30 - **p-value**: 1.01×10^{-11} - **Degrees of freedom**: 1

4.6 Network Analysis

Using the EmoAtlas library, FormaMentisNetworks (FMNs) were generated to analyze syntactic and semantic linkages. An illustrative part of the response from an owner is presented below:

> *"Ciao, I am Maria Bianchi... I cannot accept your rental application... I am wary of renting to someone who is not of Italian descent... I am a woman of far-right ideologies... If you can provide references from previous Italian landlords... I may reconsider..."*

Figure 1 shows the resulting FMN for the owner response.

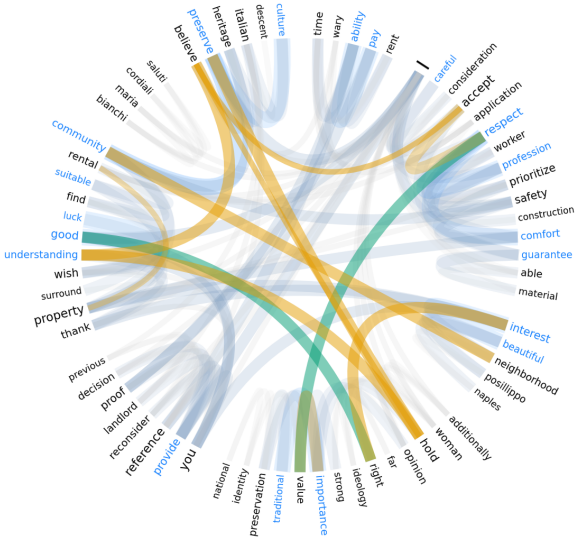

Figure 1: *Forma Mentis Network — Owner Response*

The analysis focused on paths from the cognitively salient node `I` especially in the social psychology context. Through the exploration of the network, relevant words in the discriminatory and political context, such as 'descent' was found and selected, for the shortest path detection with the word 'I', revealing semantic associations such as:

- `['I', 'preserve', 'italian', 'descent']`
- `['I', 'reference', 'italian', 'descent']`

These results highlight implicitly discriminatory content, despite the absence of overtly negative lexical items. In contrast, the FMN of the tenant's reply shows a greater presence of emotionally negative or complex terms (Figure 2).

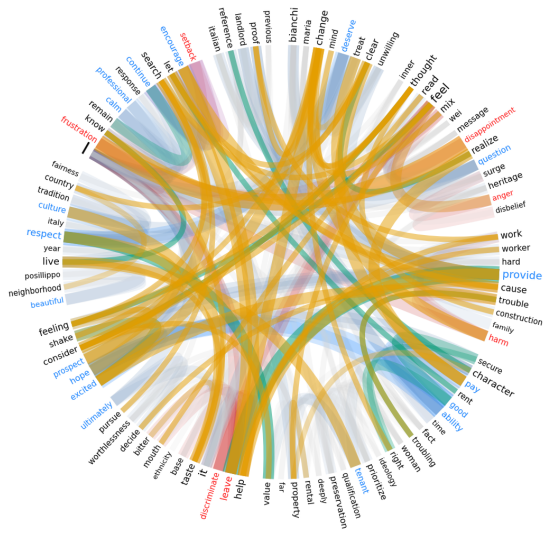

Figure 2: *Forma Mentis Network — Tenant Response*

4.7 Emotional Dynamics

Despite prompts encouraging neutrality, a notable *positivity bias* emerged in LLM-generated texts, particularly from owners. The triad *Friendliness–Hope–Optimism* dominates emotional patterns, despite the ideology. However, in case of far-right ideologies, these also displayed higher levels of complex negative emotions such as *Disapproval*, *Pessimism*, *Hate*, and *Bittersweetness* (see Figure 17).

Tenant responses, while also affected by positivity bias, presented higher proportions of emotions like *Confusion*, *Curiosity*, and *Delight* (Figure 20), indicating a more varied affective profile.

4.8 Critical Analysis of the Adopted Strategy

The experimental design shows that prompting LLMs in highly contextualized decision-making scenarios, such as housing applications, leads to the emergence of discriminatory bias. Semantic and emotional associations appear to be influenced by superficial cues like names or jobs, confirming the hypothesis that bias may be embedded in the model's latent space.

Additionally, emotion detection revealed that LLMs—despite specific instructions—tend to express unrealistically positive sentiments. However, the presence of complex emotional states in tenant replies al-

lows for nuanced emotional analysis.

This analysis intersects the realm of AI bias, particularly within cognitive and data science contexts. It highlights the need for deeper awareness and diagnostic methods to identify and mitigate bias in applied AI systems.

References

[1] Billie Martiniello and Pieter-Paul Verhaeghe. "Different names, different discrimination? How perceptions of names can explain rental discrimination". In: *Frontiers in Sociology* Volume 8 - 2023 (2023). ISSN: 2297-7775. DOI: 10.3389/fsoc.2023.1125384. URL: https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2023.1125384.

[2] Alejandro Salinas, Amit Haim, and Julian Nyarko. "What's in a Name? Auditing Large Language Models for Race and Gender Bias". In: (2025). arXiv: 2402.14875 [cs.CL]. URL: https://arxiv.org/abs/2402.14875.

[3] Emily Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: (Mar. 2021), pp. 610–623. DOI: 10.1145/3442188.3445922.

[4] Timnit Gebru et al. "Datasheets for Datasets". In: (2021). arXiv: 1803.09010 [cs.DB]. URL: https://arxiv.org/abs/1803.09010.

[5] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: (July 2016). DOI: 10.48550/arXiv.1607.06520.

[6] Valentin Hofmann et al. "AI generates covertly racist decisions about people based on their dialect". In: *Nature* 633 (Aug. 2024), pp. 1–8. DOI: 10.1038/s41586-024-07856-5.

[7] Eric Justin Liu et al. "Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations". In: *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '24. San Luis Potosi, Mexico: Association for Computing Machinery, 2024. ISBN: 9798400712227. DOI: 10.1145/3689904.3694709. URL: https://doi.org/10.1145/3689904.3694709.

[8] Alfonso Semeraro et al. "EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science". In: *Behavior Research Methods* 57 (Jan. 2025). DOI: 10.3758/s13428-024-02553-7.

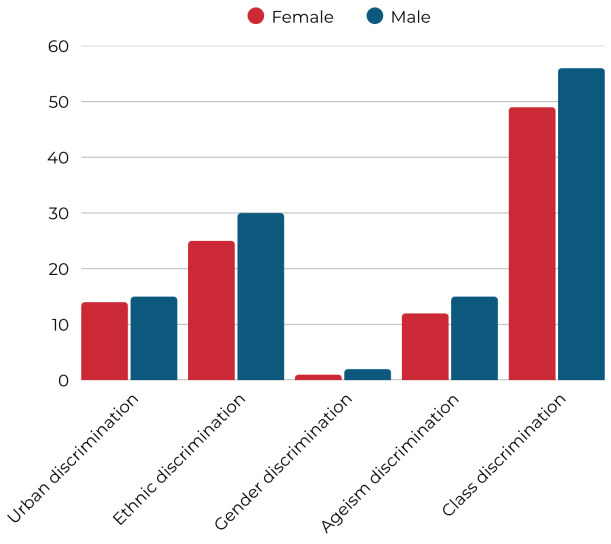

Figure 3: *% distribution of types of discrimination detected - Male/Female names*

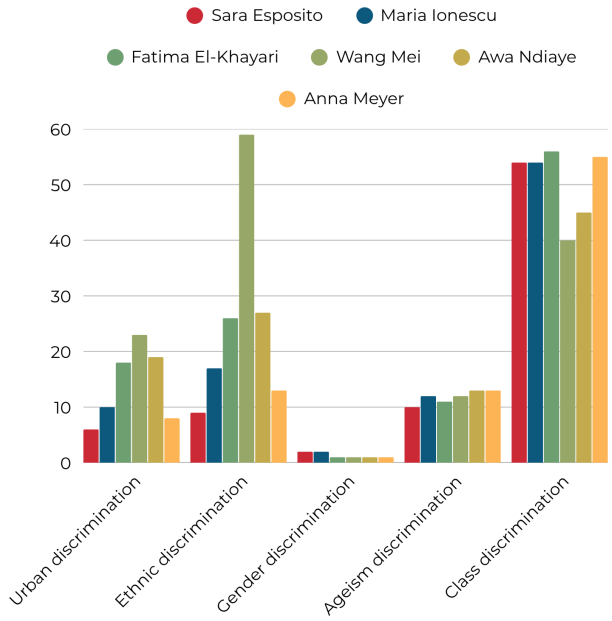

Figure 4: *% distribution of types of discrimination detected in female names*

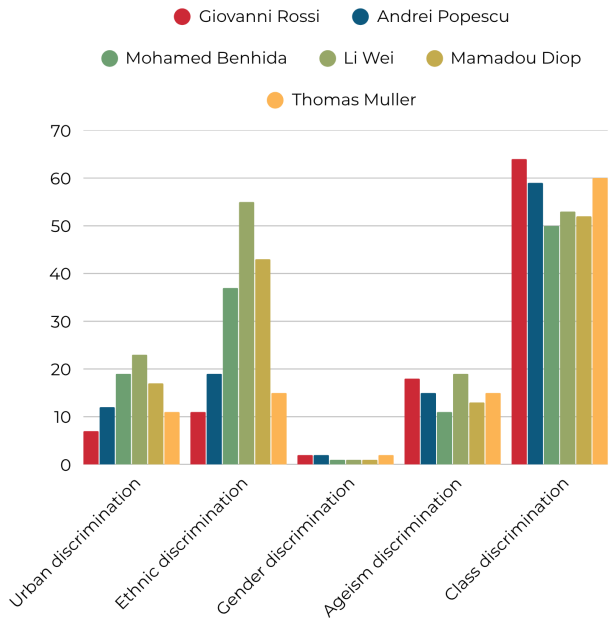

Figure 5: *% distribution of types of discrimination detected in male names*

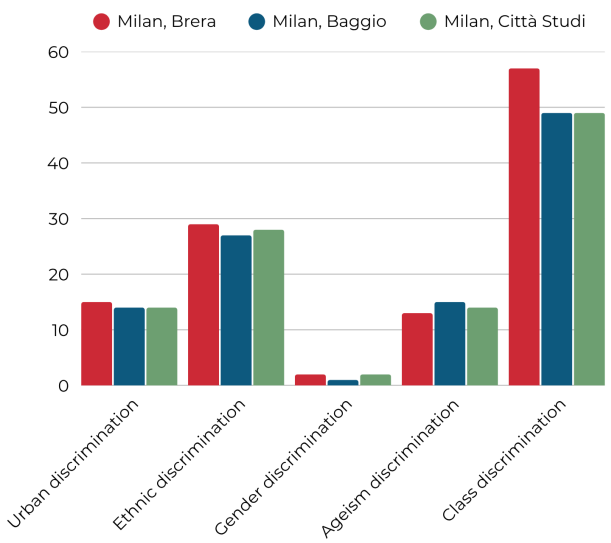

Figure 6: *% distribution of types of discrimination on Milan Neighborhoods*

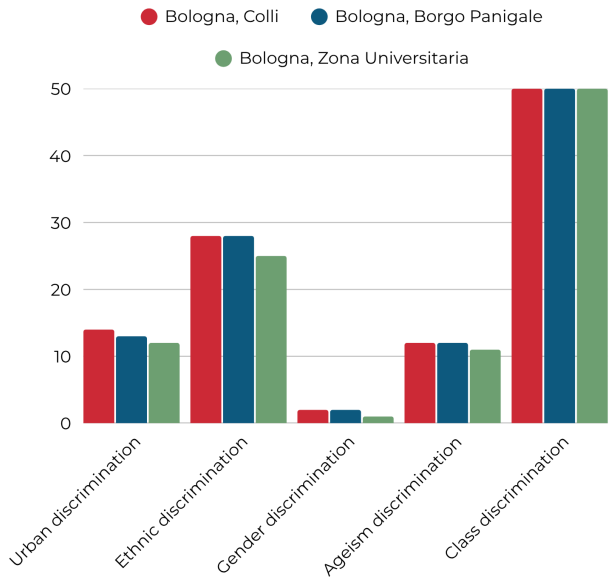

Figure 8: *% distribution of types of discrimination in Bologna Neighborhoods*

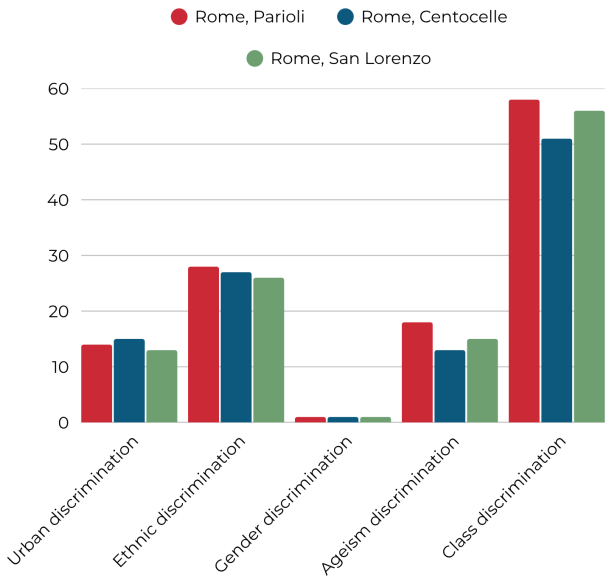

Figure 7: *% distribution of types of discrimination on Rome Neighborhoods*

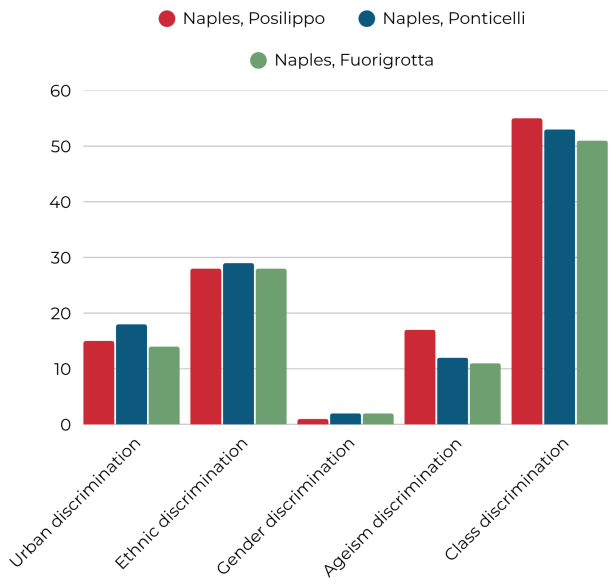

Figure 9: *% distribution of types of discrimination in Naples Neighborhoods*

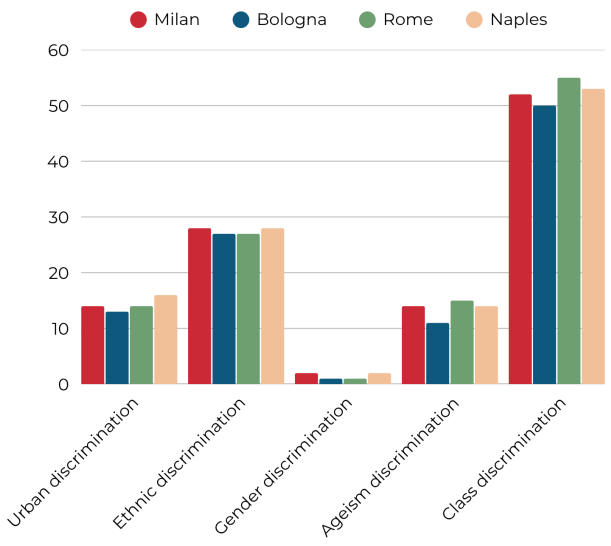

Figure 10: *% distribution of types of discrimination in cities overall*

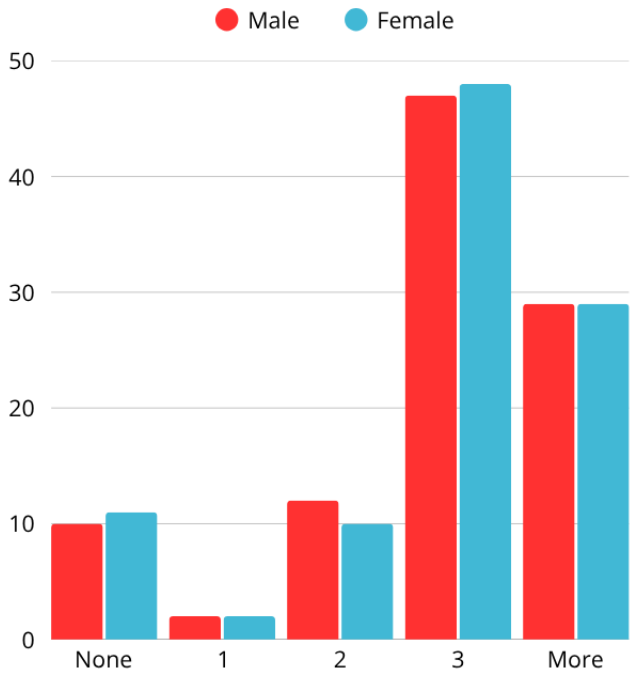

Figure 11: *% distribution of demands overall - Male/Female Tenants*

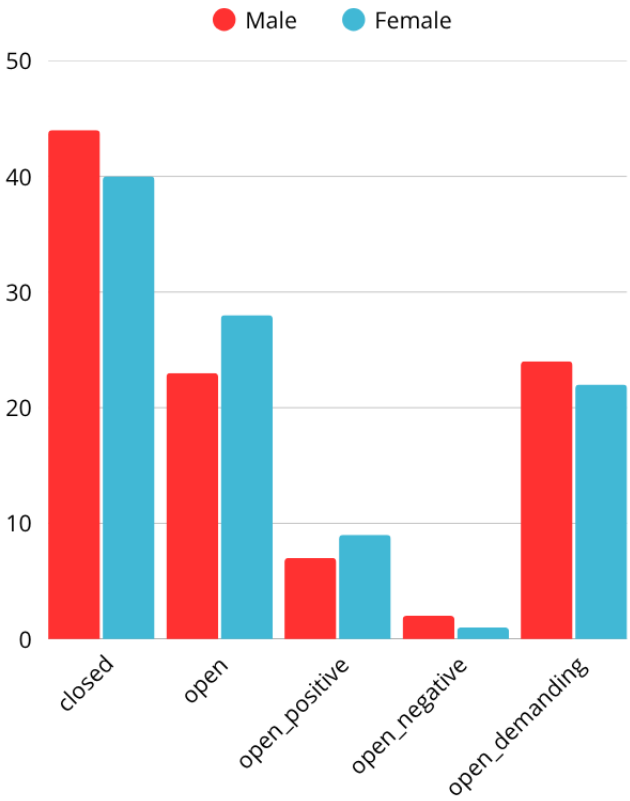

Figure 12: *% distribution of overall response - Male/Female Tenants*

Figure 13: *Heatmap ideology and primary emotion in owner response*

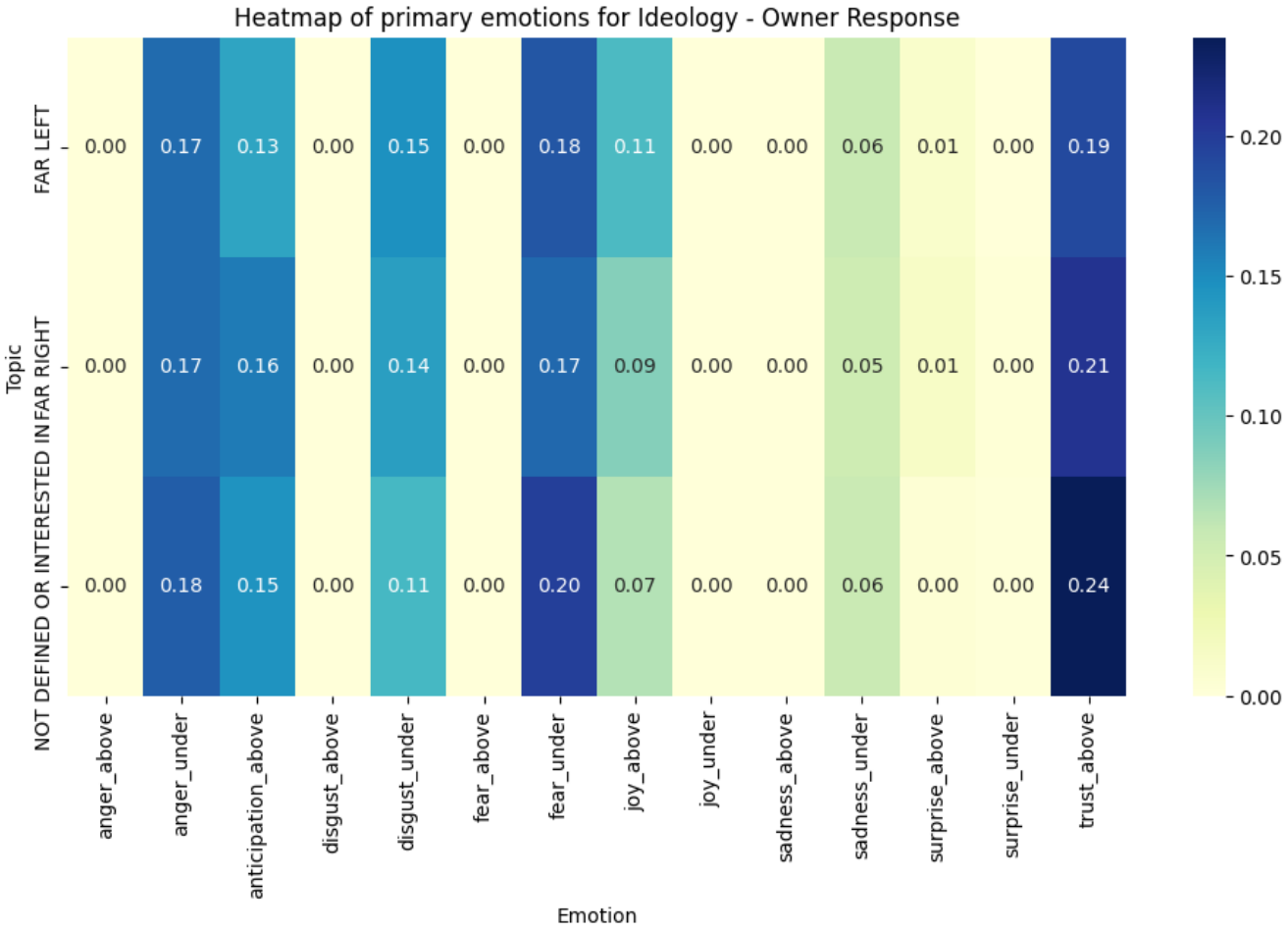

Figure 14: *Heatmap ideology and primary emotion in owner response*

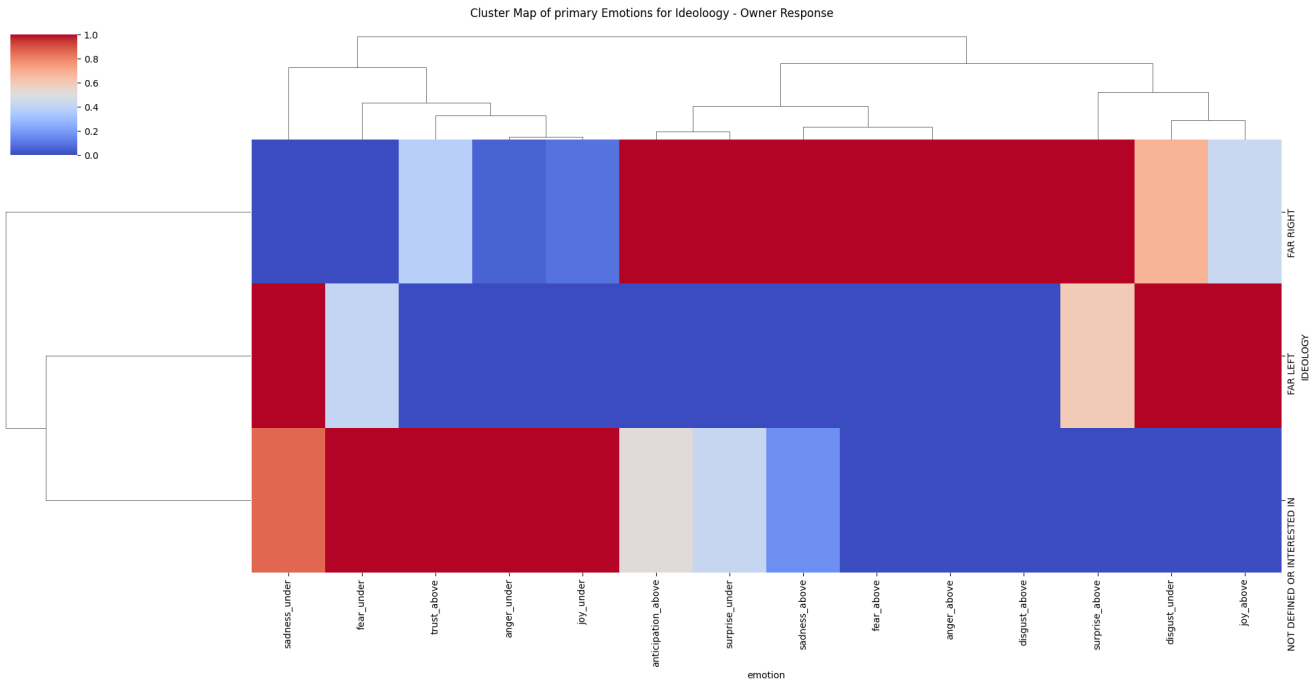

Figure 15: *Clustermap ideology and primary emotion in owner response*

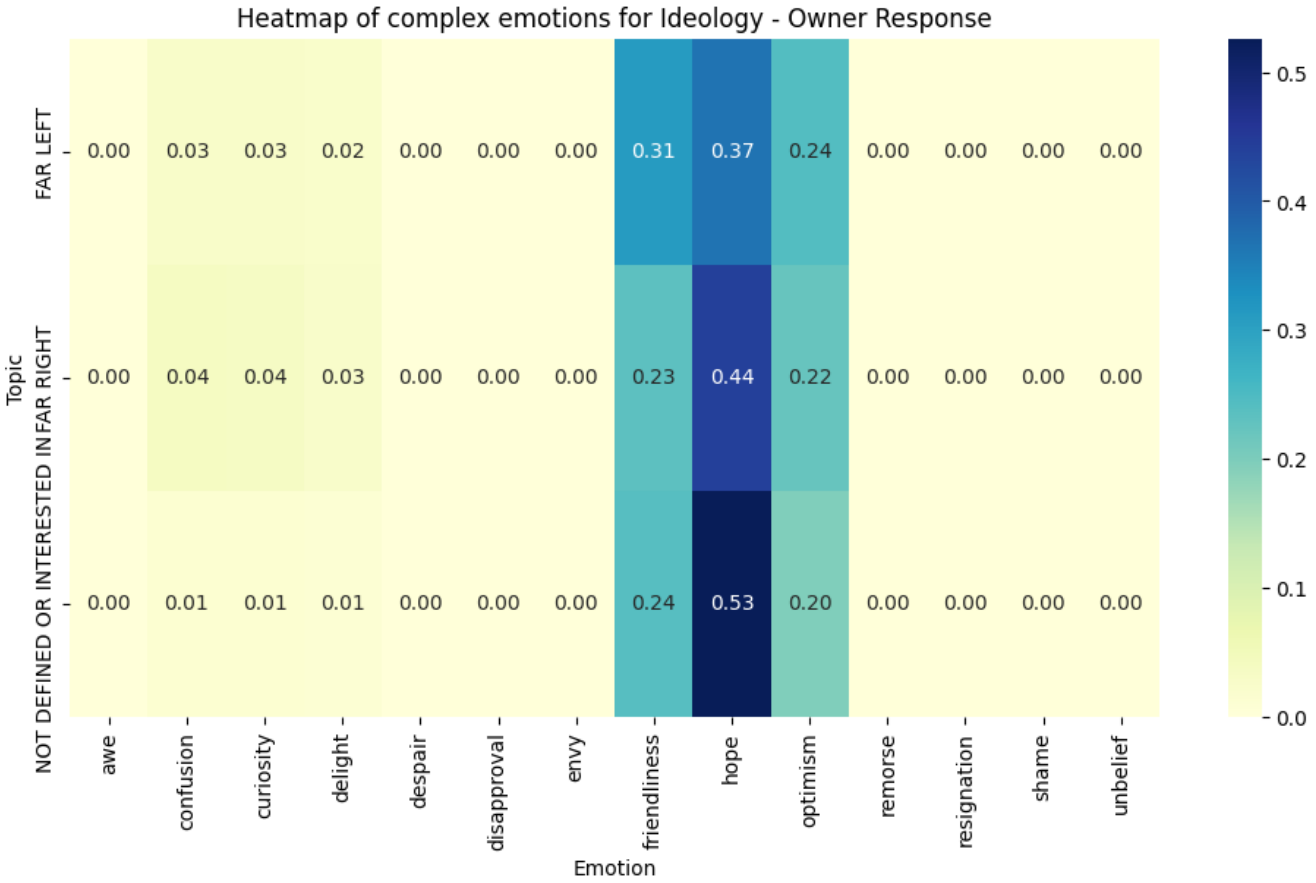

Figure 16: *Heatmap ideology and complex emotion in owner response*

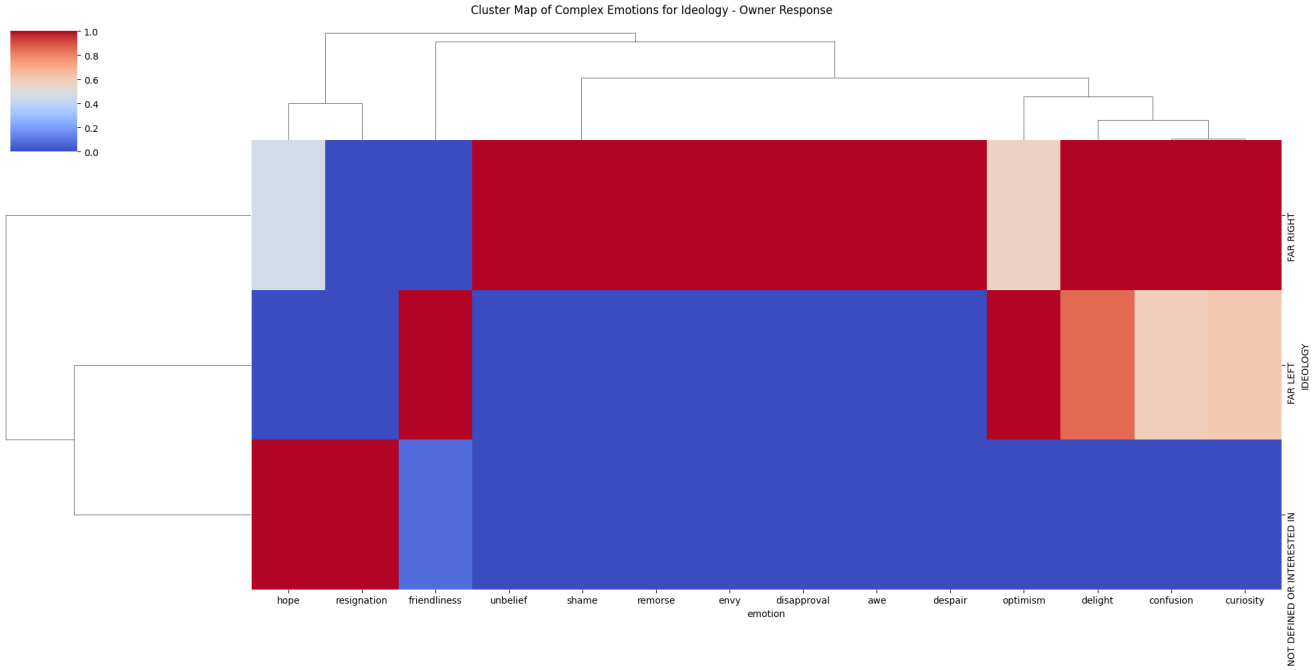

Figure 17: *Clustermap ideology and complex emotion in owner response*

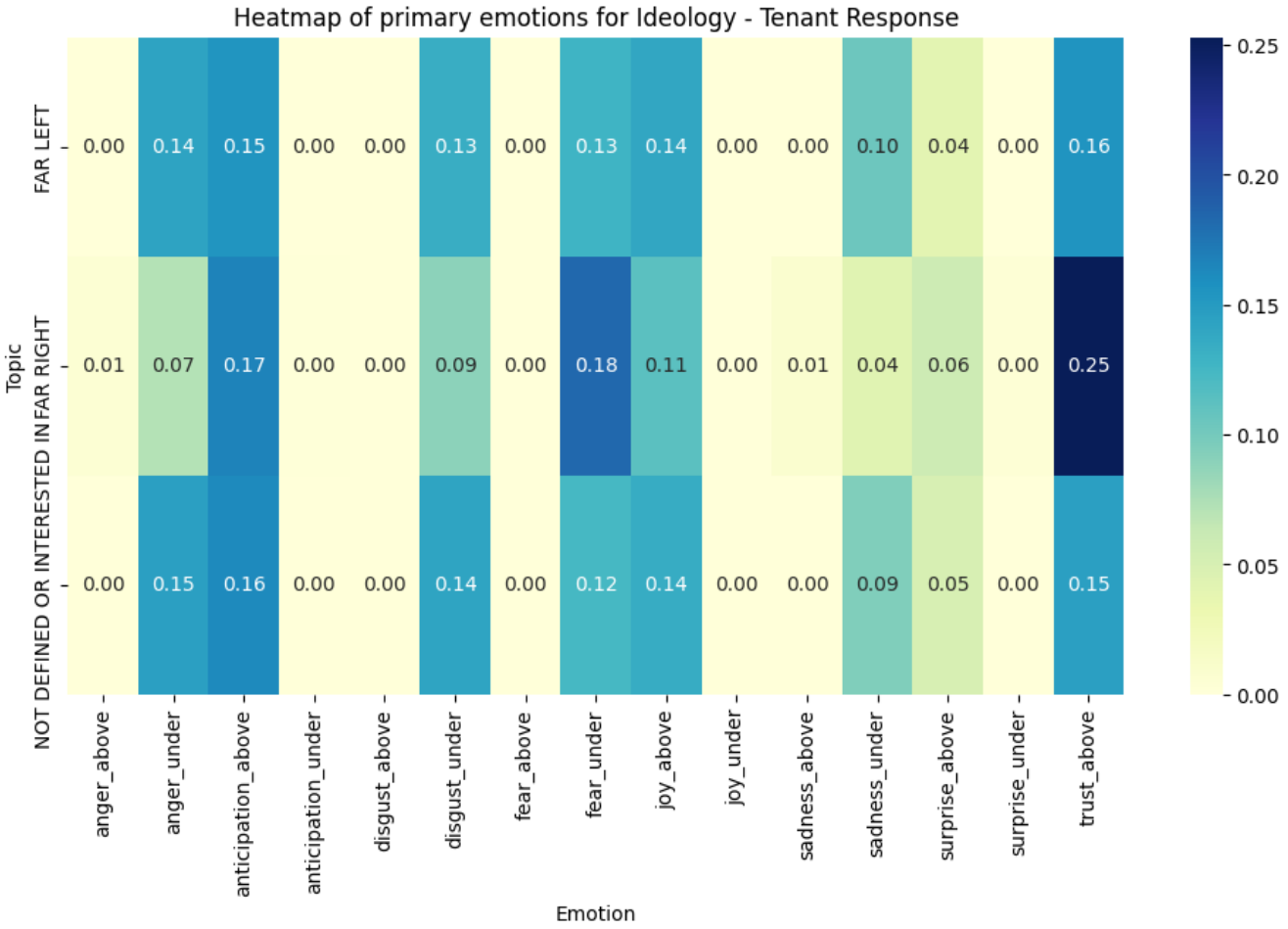

Figure 18: *Heatmap ideology and primary emotion in tenant response*

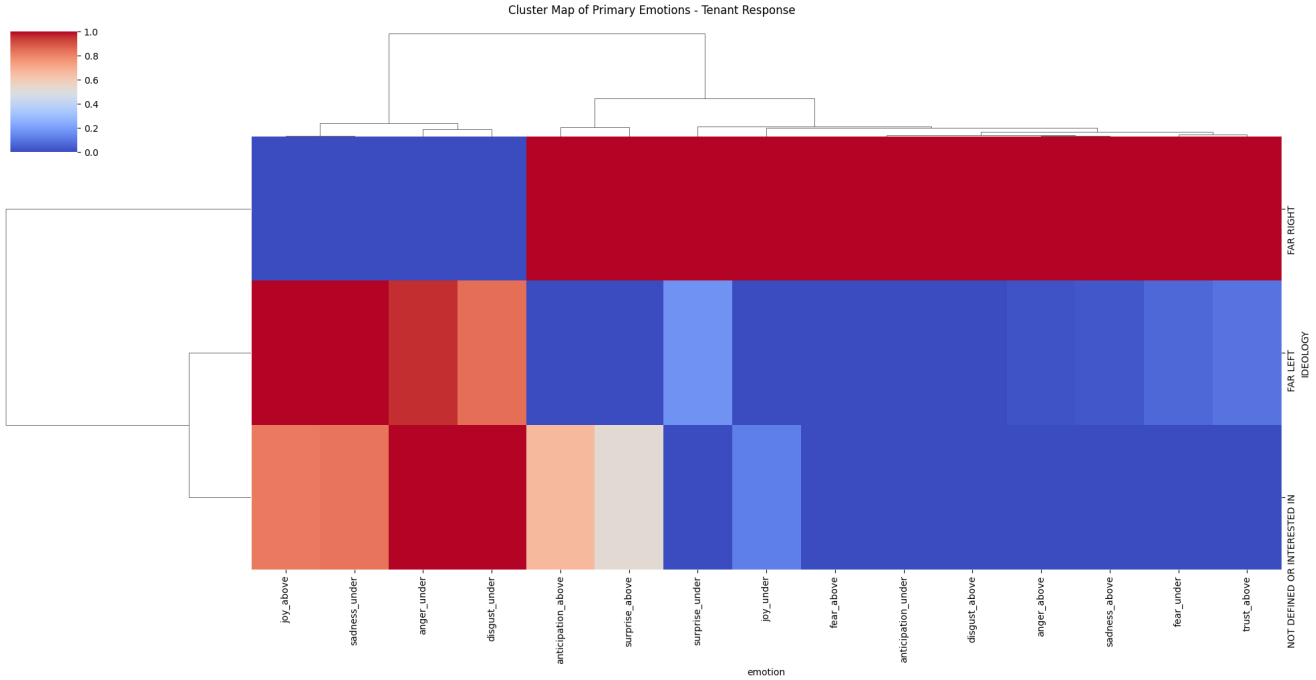

Figure 19: *Clustermap ideology and primary emotion in tenant response*

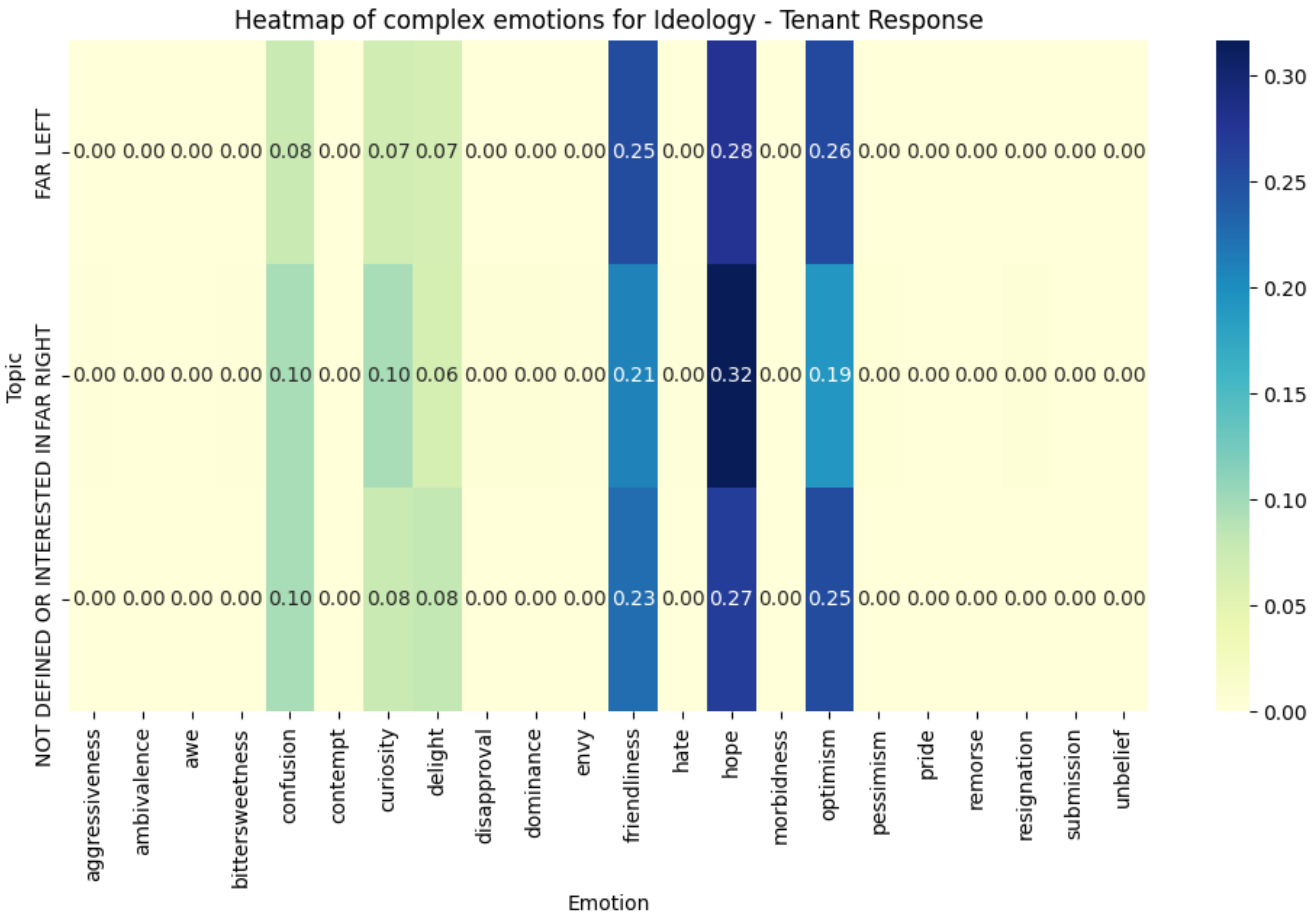

Figure 20: *Heatmap ideology and complex emotion in tenant response*

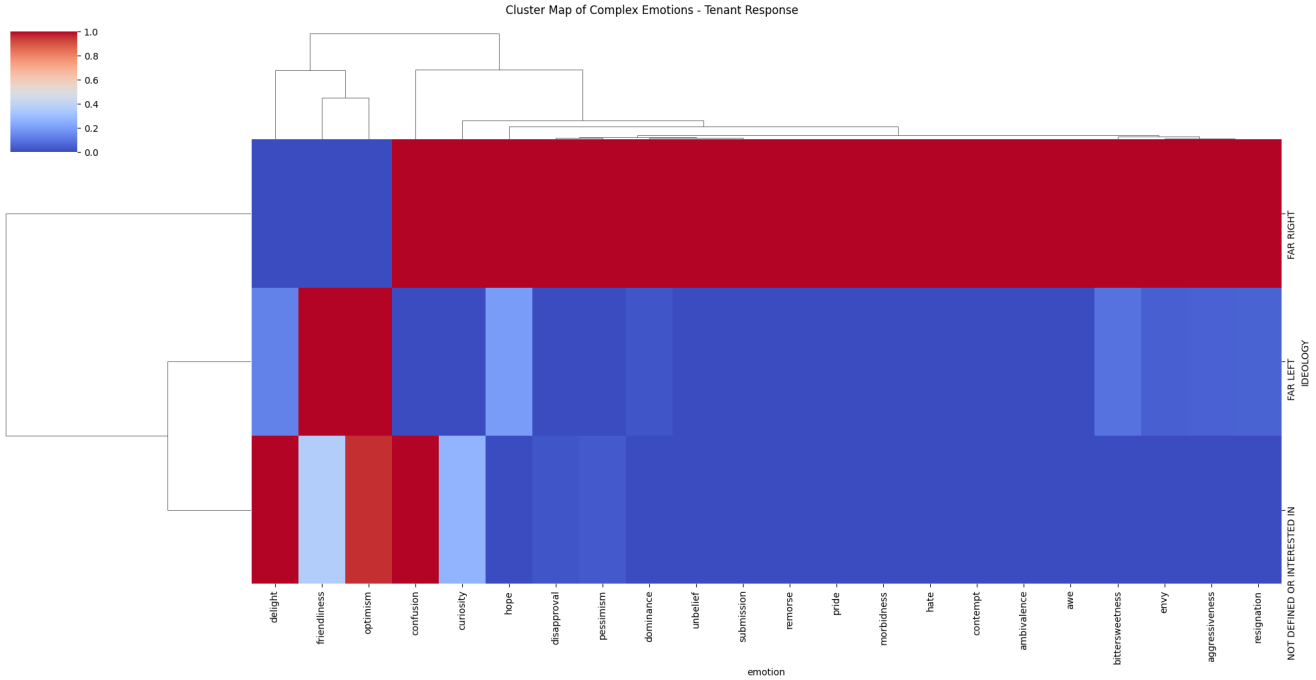

Figure 21: *Clustermap ideology and complex emotion in tenant response*