

Master in Intelligent Interactive Systems
Universitat Pompeu Fabra

Invisible Signals: Detecting Potential Selection Bias in AI-Based Resume Screening

Pau Buyreu Real

Supervisor: Carlos Castillo

July 2025



Master in Intelligent Interactive Systems
Universitat Pompeu Fabra

Invisible Signals: Detecting Potential Selection Bias in AI-Based Resume Screening

Pau Buyreu Real

Supervisor: Carlos Castillo

July 2025



Contents

1	Introduction	1
1.1	Motivation	2
1.2	Objectives	3
1.3	Structure of the Report	4
2	State of the Art	6
2.1	Algorithmic Discrimination in Recruitment	6
2.2	Proxies for Protected Characteristics in Resumes	8
2.3	Fairness Interventions and Debiasing Methods in Recruitment AI	9
3	Methods	11
3.1	Resumes Dataset Description	12
3.2	Word Classification and Text Cleaning	14
3.2.1	Text Cleaning	14
3.2.2	Word Classification	15
3.3	Resume Classification	17
3.3.1	Classification Labels	17
3.3.2	Model Architecture	18
3.4	Lexical Shift Analysis with Shifterator	20
3.5	Ablation Study Design	21
4	Results	24
4.1	Perceived Origin	25

4.1.1	Classification Metrics	25
4.1.2	General Lexical Contributions	26
4.1.3	Lexical Contributions by Word Group	27
4.2	Gender	29
4.2.1	Classification Metrics	29
4.2.2	General Lexical Contributions	30
4.2.3	Lexical Contributions by Word Group	31
4.3	Religion	32
4.3.1	Classification Metrics	32
4.3.2	General Lexical Contributions	33
4.3.3	Lexical Contributions by Word Group	34
4.4	Sexual Orientation	35
4.4.1	Classification Metrics	35
4.4.2	General Lexical Contributions	36
4.4.3	Lexical Contributions by Word Group	37
5	Discussion	39
5.1	Discussion	39
5.1.1	General Classification Capability	39
5.1.2	Keyword-Level Signals	45
5.1.3	Implications for the Protection of Demographic Attributes	47
5.2	Conclusions	49
5.2.1	General Conclusions	49
5.2.2	Limitations	50
5.2.3	Future Work	52
	List of Figures	54
	List of Tables	55

Bibliography	58
A Additional Lexical Top Words	63
A.1 Perceived Origin	64
A.2 Gender	65
A.3 Religion	66
A.4 Sexual Orientation	67
B Additional Lexical Word Group Tables	68
B.1 Perceived Origin	69
B.2 Gender	70
B.3 Religion	72
B.4 Sexual Orientation	74

Acknowledgement

Acknowledgements

I would like to sincerely thank my supervisor, Carlos, for proposing the topic and for all the guidance and support he provided throughout the development of this thesis. I am also grateful to his colleagues Jorge and Anna for their help and feedback.

My thanks also go to the contributors of the FINDHR project, including Carlos and Jorge, for granting me access to the dataset, which made it possible to work with real-world data.

Finally, I want to thank Chris for her patience and support during the writing of this thesis.

Abstract

Automated resume screening tools are now widely used in hiring processes, offering the promise of efficiency and fairness by reducing human bias. Yet recent studies have shown that these systems can still behave unfairly by picking up on subtle linguistic clues that reveal sensitive personal information. This thesis explores whether transformer-based models can infer protected attributes (gender, perceived origin, religion, or sexual orientation) from resume text, even when this information isn't stated directly.

To investigate this, the study analyzes a real-world dataset of over 900 resumes. Each document was cleaned and its words categorized into semantic groups, such as occupation-related words, location-related, skill-related, and proper nouns. The main method used is a series of lexical ablation experiments: for each demographic attribute, twelve experiments were run by including or excluding different word categories. These were combined with a lexical shift analysis using Shifterator to identify which specific words most influenced the model's predictions.

The results show that models can reliably infer gender and perceived origin. Occupation-related terms were mainly predictive of gender, while geographic references were almost direct cues in identifying perceived origin. However, attempts to predict religion and sexual orientation failed, likely due to limited language cues or imbalanced data. Interestingly, even individual words like gendered job titles (e.g., "waitress") or places names were enough to act as unintended signals.

These findings raise important concerns about fairness in algorithmic hiring. The fact that AI models can detect protected attributes even in anonymized resumes suggests that bias may persist through indirect linguistic patterns. This highlights the need for stronger audits, more transparent systems, and proactive strategies to reduce bias, such as masking certain word types or using debiasing techniques during training. It also calls for caution when relying on AI-driven tools in hiring.

Overall, this thesis adds to the field of algorithmic fairness by presenting a practical framework to identify and understand hidden bias in resume screening. It shows that

removing obvious identifiers is not enough; fairness also depends on understanding how language itself can reveal sensitive information.

Keywords: Algorithmic fairness; Resume screening; Proxy discrimination; NLP; Transformers; Selection bias

Chapter 1

Introduction

Despite decades of equal opportunity legislation and diversity initiatives, discrimination in hiring remains a persistent challenge across labor markets. People in protected groups such as women, immigrants and ethnic minorities often face disproportionate barriers when applying for jobs, even when their qualifications are equivalent to majority-group candidates. Meta-analyses of correspondence studies have shown that such applicants frequently need to submit significantly more applications to receive the same number of interview invitations [1, 2].

In recent years, automation and AI tools have been introduced to streamline recruitment processes and reduce subjective bias. Today, the vast majority of large companies, including the 99% of Fortune 500 firms, use some form of algorithmic assistance in hiring decisions [3]. These tools, from rule-based applicant tracking systems to large language models (LLMs), promise efficiency, objectivity, and scalability. However, as recent studies have demonstrated, AI does not inherently eliminate bias. On the contrary, it often reflects and amplifies the historical and structural inequalities present in the data it learns from [3, 4, 5].

Resume screening, the initial and often decisive stage of recruitment, is particularly vulnerable to bias. During this stage, recruiters form judgments based on limited information, relying on superficial cues such as names, addresses, educational background, or even writing style [6]. These cues can activate implicit stereotypes and,

consciously or unconsciously, lead to discriminatory decisions [1, 7]. Even subtle linguistic features, such as vocabulary and grammar can correlate with an applicant's social background, contributing to differential outcomes for equivalent qualifications [8].

Bias in AI Resume screening can arise from multiple sources. Training data may be imbalanced, historical hiring patterns may reflect past discrimination, and linguistic or cultural cues in resumes may be unintentionally penalized [9]. This can result in the systematic exclusion of applicants from marginalized groups, even when identity markers such as names are removed [10]. Moreover, efforts to anonymize resumes, intended to prevent bias, have at times produced counterproductive effects [6, 11].

This thesis explores how linguistic cues in resumes may enable AI-based screening systems to infer protected attributes such as gender, origin, religion, or sexual orientation, even when such information is not explicitly stated. To investigate this phenomenon, a lexical-level ablation study is conducted using a real-world resume dataset. By isolating semantic word groups and evaluating their impact on classification performance, this work aims to uncover the textual signals most responsible for proxy discrimination, and to inform future mitigation strategies.

1.1 Motivation

A growing body of evidence suggests that AI systems can infer sensitive demographic attributes such as ethnicity, gender, religion, or sexual orientation through indirect linguistic cues present in resumes [9, 4]. This is particularly concerning because even when explicit identifiers such as names or photos are anonymized, attributes like grammar, vocabulary, phrasing, or section structure may still correlate with a candidate's background [8]. These subtle signals can enable models to "re-identify" individuals from minority or marginalized groups, undermining anonymization efforts and exacerbating bias.

If such hidden cues remain unaddressed, automated resume screening tools, designed to enhance fairness and efficiency, may instead reinforce discriminatory practices.

While some mitigation strategies exist, such as debiasing algorithms or attribute obfuscation, their effectiveness depends on a deeper understanding of how and where proxy signals arise in textual data [12]. Despite the growing use of AI in hiring pipelines, relatively few studies have systematically investigated the linguistic pathways through which discrimination can emerge across multiple protected group.

This thesis is motivated by the ethical imperative to ensure fair access to employment in the context of automated decision-making. It addresses a critical gap at the intersection of algorithmic fairness and computational linguistics by identifying linguistic patterns that act as proxies for protected attributes and testing their impact through controlled ablation experiments. The findings aim to inform both technical design and policy regulation of AI-based hiring systems.

1.2 Objectives

Main Objective

To investigate the extent to which AI-based resume screening systems can infer protected attributes based only on linguistic patterns, and to identify the textual features that contribute most to such inferences.

Specific Objectives

1. To evaluate whether a machine learning classifier (transformer-based) can predict demographic group membership from resume text alone, even in the absence of explicit personal information.
2. To systematically isolate and assess the impact of different semantic word groups on classification outcomes using ablation techniques.
3. To quantify the most influential lexical features driving model predictions through divergence-based analysis (e.g., Shifterator).
4. To explore the implications of proxy linguistic cues in resume screening systems for algorithmic fairness and discrimination risk.

Research Questions

1. To what extent can AI models infer protected characteristics from textual resumes?
2. Which semantic categories or lexical patterns most strongly influence the model's predictions for different demographic attributes?
3. How does the presence or absence of specific word groups affect the identifiability of sensitive information in resume classification tasks?

1.3 Structure of the Report

This thesis is organized into six chapters, each covering a key component of the research process:

- **Chapter 1: Introduction** — Introduces the motivation behind the study, with a focus on the ethical risks of bias in automated resume screening. It defines the objectives of the work and outlines the protected attributes studied.
- **Chapter 2: State of the Art** — Reviews the literature on algorithmic bias, the role of language in demographic inference, and how modern NLP models can both perpetuate and detect these biases. It also discusses common mitigation strategies and the importance of interpretability.
- **Chapter 3: Methods** — Details the dataset used in the study, the preprocessing steps, and the experimental design. It explains how transformer-based classifiers were trained and evaluated, how lexical ablation experiments were structured, and how the Shifterator framework was used to analyze word-level contributions.
- **Chapter 4: Results** — Presents the empirical results for each of the four classification tasks: *Perceived Origin*, *Gender*, *Religion*, and *Sexual Orientation*. It includes both classification performance and lexical analyses, supported by tables and visualizations.

-
- **Chapter 5: Discussion, Conclusions and Future Work** — Interprets the results in depth, comparing the success of different classification tasks, analyzing linguistic patterns, and discussing key concepts such as proxy indicators and stereotype reinforcement. It also reflects on the ethical implications of the findings. The conclusions summarize the main insights from the study, acknowledge its limitations, and propose directions for future research, including the exploration of bias mitigation techniques and the use of larger datasets or stronger models.

Chapter 2

State of the Art

2.1 Algorithmic Discrimination in Recruitment

Algorithmic hiring tools have been shown to reflect and even exacerbate existing societal biases. Systems trained on historical hiring data, for instance, often replicate past patterns of discrimination against minority groups [3, 4].

A significant example is Amazon's now-discontinued recruitment AI, which penalized resumes containing terms like "women's" or graduates from women-only colleges for software engineering vacancies. This bias emerged not from explicit instructions but from the system learning preferences embedded in historical hiring data, reflecting a male-dominated field [8]. Similarly, large language models like ChatGPT have demonstrated systemic disparities in resume evaluations, with Arab and Asian applicants often rated lower than majority-group applicants with identical profiles [4].

Dovidio and Gaertner presented the phenomenon of **aversive racism**, a form of unconscious bias often masked by overt egalitarian attitudes. This phenomenon helps explain how discrimination can persist in seemingly neutral systems [13]. When AI models internalize social norms and patterns from data, they can unintentionally amplify these biases in ways that are difficult to detect, audit, or correct. This not only risks marginalizing applicants from protected groups but also undermines trust

in AI systems used for critical employment decisions [14].

A recurring finding is that discrimination is most likely to occur in ambiguous cases, situations where applicant qualifications are neither clearly strong nor clearly weak. In such contexts, implicit biases tend to guide decision-making more strongly [13]. While anonymization of resumes has been proposed as a mitigation strategy, its effectiveness remains inconsistent. Some studies found that anonymized resumes improved outcomes for minority applicants, while others revealed negative or negligible effects, suggesting that anonymization may sometimes reduce positive action or reinforce suspicion [10].

Recent work has extended this conversation to *intersectional discrimination*, where biases do not arise solely from individual characteristics such as gender or race, but from their interaction. As highlighted by Gohar and Cheng, systems that appear fair across individual dimensions (using mitigation techniques) may still discriminate against subgroups that lie at their intersections, such as Black women or trans individuals, due to compounding disadvantages [15]. Morina et al. provide practical metrics and post-processing techniques to audit and mitigate such intersectional bias in machine learning models, even when intersectional subgroups are underrepresented in the data [16].

Furthermore, models may learn discriminatory patterns indirectly via proxy variables. Information such as university names, residential postal codes, or extracurricular activities can encode sensitive attributes without explicitly stating them, thereby bypassing fairness constraints [8]. Chen emphasizes that algorithmic bias in recruitment often stems from both data limitations and a lack of diverse perspectives in algorithm design, advocating for technical measures like bias-aware datasets and managerial oversight frameworks [17].

The lack of transparency in many AI-powered hiring tools also contributes to this issue. Many systems operate as black boxes, limiting the ability of candidates or auditors to understand or challenge decisions. Beretta et al. argue that the incorporation of explainable AI (XAI) mechanisms is essential to improve both fairness

and trust in recruitment algorithms. Their study outlines requirements for XAI systems tailored to hiring, emphasizing the need to align explanations with stakeholder expectations [18]. Complementing this, Zhou et al. show that different types of AI explanations influence human perceptions of fairness, and that responsible design of explanations should incorporate insights from social science to be effective [19].

2.2 Proxies for Protected Characteristics in Resumes

Even in the absence of overt markers like names or photos, resumes often contain proxies: neutral-seeming features that indirectly reveal protected characteristics. These include choice of words, stylistic preferences, academic or professional affiliations, educational institutions, geographic locations, or extracurricular activities [8]. For instance, attending an all-women’s college, listing involvement in a religious student association, or using non-native syntactic structures can serve as indicators of gender, religion, or ethnicity [8].

Deshpande et al. (2020) demonstrated that even when names were removed, models trained on resume content could still infer demographic categories with high accuracy based on socio-linguistic patterns. Their “fair-tf-idf” approach aimed to reduce such correlations by adjusting term weights in the matching algorithm, but residual identification remained a challenge [5].

The qualitative study by Vinod Bhatia et al. (2024), that analyzed the same CVs dataset as this thesis, a dataset that contains over 1,000 real-world CVs, reinforced these concerns by showing how intersectional identities manifest through resume content. Markers related to ethnic background, religion, gender, or socio-economic class can be embedded in CVs through implicit cultural references, phrasing, or educational trajectories [8]. These signals are often interpreted as indicators of fit, competence, or desirability, even when such inferences are unjustified.

As said, anonymization may not be sufficient to eliminate the impact of these proxies. Linguistic style itself can trigger biased evaluations. Chu et al. (2024) found that applicants from Asian and Hispanic backgrounds were more frequently perceived as

relying on AI-generated content, due in part to writing style heuristics, which in turn lowered their competence and authenticity evaluations [9].

In a large-scale audit of resume screening algorithms, Binns et al. (2018) demonstrated that location, university names, and even certain extracurriculars were associated with lower scores for minority candidates, even after controlling for qualifications. They argue that the presence of “covert indicators” can lead to unintentional discrimination by creating false associations between identity and job suitability [20].

Moreover, word embeddings themselves can act as conduits of bias. Bolukbasi et al. (2016) showed that embeddings capture deep cultural associations, such as analogies like “man is to computer programmer as woman is to homemaker.” These statistical relationships, when used in recruitment scoring, risk reintroducing stereotype-based evaluations even in otherwise anonymized systems [21].

Together, these findings highlight a central paradox in AI hiring: even if explicit identity markers are removed, the algorithm can still “see through” anonymization by learning from patterns in the data. This will be further investigated in this thesis.

2.3 Fairness Interventions and Debiasing Methods in Recruitment AI

While this thesis does not directly implement debiasing techniques, the findings suggest an urgent need for such interventions. This section provides an overview of the current literature on fairness-aware methods in AI recruitment systems, offering context for the risks uncovered and directions for future research.

Recent literature categorizes interventions into three levels: pre-processing, in-processing, and post-processing techniques.

Pre-processing approaches focus on data modification before model training. Counterfactual data augmentation, creating synthetic samples that alter protected attributes (e.g., replacing gender pronouns) has proven effective in mitigating bias in

language models [22]. Applied to hiring, this technique can balance representation in resume data, reducing proxy-based discrimination.

In-processing methods incorporate fairness objectives directly into model training. Adversarial debiasing, for instance, trains a primary classifier to predict job fit while simultaneously training an adversary to predict protected attributes; the classifier is penalized when the adversary succeeds [23]. Such adversarial frameworks have been successfully used to remove gender signals from word embeddings in recruitment scenarios, showing real-world reductions in wage and job recommendation disparities [23].

Post-processing methods adjust model outputs to satisfy fairness constraints. Techniques such as equalized odds ensure equal false positive and negative rates across protected groups by modifying decision thresholds [24]. Despite their appeal, these methods remain rare in commercial hiring tools, in part due to regulatory complexity and business integration challenges.

Beyond technical strategies, socio-technical audits are critical. Fabris et al. emphasize the role of multidisciplinary audits that combine fairness metrics, stakeholder input, and legal review throughout the development lifecycle of hiring AI systems [25]. These audits help identify bias sources and document mitigation steps, increasing accountability.

Finally, Meade et al. conducted an empirical survey of debiasing methods in pre-trained language models, reporting mixed results: while techniques like data augmentation reduce bias, they sometimes degrade model accuracy and vary in effectiveness depending on the attribute (e.g., gender vs ethnicity) [22]. These trade-offs highlight the need for evaluation frameworks specific to hiring contexts, where fairness and performance must be carefully balanced.

Chapter 3

Methods

This chapter outlines the methodology used in the thesis. It covers the dataset, how words were classified, the steps taken to clean and prepare the text, the classification setup, and the tools used to interpret potential bias. An overview of the full process is shown in Figure 1, and each part is explained in more detail in the following sections.

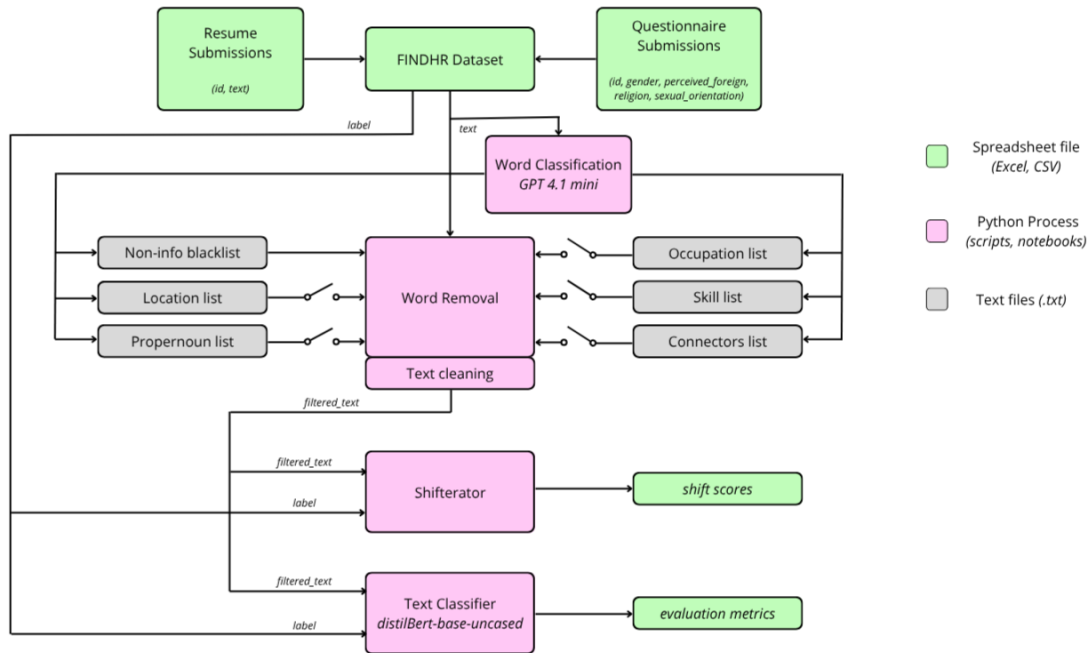


Figure 1: Overview of the methodological pipeline including word classification, ablation, lexical shift analysis, and resume classification.

3.1 Resumes Dataset Description

The dataset used in this study was collected through a data donation campaign run as part of the FINDHR project (<https://findhr.eu/>), as described by Bhatia et al. [8]. The campaign invited people living in the European Economic Area or Switzerland who were working or looking for work to anonymously donate their resumes through an online survey. Before participating, donors had to confirm they were over 18, give informed consent, and fill out a demographic questionnaire with sensitive information such as gender, age, sexual orientation, ethnicity, religion, and disability status.

After collecting the data, the original authors [8] processed the resumes by automatically translating them into English and applying structural parsing to organize the content clearly. The parsing step was key for later computational analysis and to ensure consistency across all documents, especially for quantitative research like

this thesis.

Out of the 1211 total resumes, only those written in English and processed through parsing (either manually or using Eden-AI) were used in this work. This resulted in a final dataset of 921 resumes. This filtering was necessary given the nature of the study, which relies on transformer-based models. To get reliable and comparable results, it was important that all resumes followed a similar structure and language style.

Each entry in the dataset includes both the resume text and the associated metadata collected through the survey. The structure is as follows:

- **ID:** A unique identifier assigned to each anonymous submission. It serves as a linking key between the resume text and the corresponding responses in the data questionnaire.
- **Resume:** A single string containing all parsed text from the submitted CV. This includes section titles and labels, concatenated into one document for each entry.
- **Gender:** One of the following options: *Man*, *Woman*, *Non-binary*, or *Self-Identify*.
- **Sexual Orientation:** One of the following values: *Asexual*, *Bisexual*, *Heterosexual*, *Homosexual*, or *Self-Identify*.
- **Perceived Foreignness:** Indicates whether the individual perceives themselves as foreign in the country where they live. Options are: *No, never*; *Yes, always*; or *Yes, sometimes*.
- **Religion:** One of the following: *Buddhism*, *Christianity*, *Hinduism*, *Islamic*, *Judaism*, *None*, *Other*, or *Secular*.

3.2 Word Classification and Text Cleaning

The main goal of this thesis is to find out whether an AI model used for resume screening could differentiate protected groups from a full resume dataset, which could lead to discriminatory outcomes. To explore this, the study looks at the language used in resumes. Unlike essays or cover letters, resumes usually don't follow a long, continuous narrative. Instead, they are made up of short sections, bullet points, and isolated phrases where individual words carry most of the meaning. Often, there's little or no connection between nearby words, which makes full-sentence or paragraph-level analysis harder and less useful.

Because of this structure, two options were considered for how to represent the text. One was to focus only on professional summaries, where writing style might reveal more about the candidate. However, not all resumes had this section, and using only those would have reduced the size of the dataset. The other option, which was ultimately chosen, was to analyze resumes at the word level, treating each word on its own and ignoring how words relate within sentences or paragraphs.

To make this word-level analysis work, it was necessary to create clear and meaningful word groups that could be used in different classification experiments. The idea was to sort the words in the resumes into distinct semantic categories. These groups were then used to include or exclude specific types of words during experiments, helping to understand how different kinds of words influence the model's ability to detect protected attributes.

3.2.1 Text Cleaning

Before starting the full word classification process, it was important to clean the resume texts by removing elements that had been added during parsing and were not written by the participants themselves. These included things like structural labels, metadata tags, and automatic section headers such as *end_date*, *first_name*, or *self_summary*, as well as web-related terms like *https* or *.com*. To do this, a manually created blacklist was used to filter out all these unwanted tokens, making

sure that only the original text written by the users remained in the dataset.

Once the blacklist had been applied, a second cleaning step was done to remove any leftover noise, such as punctuation, formatting characters, or broken structures, that no longer served any purpose. A custom function using regular expressions was used to clean up these elements, including repeated punctuation, stray delimiters (like slashes or pipes), empty fields, or random numbers. This helped ensure that the final text was clean and consistent, ready for word-level analysis.

3.2.2 Word Classification

All the words taken from the cleaned resumes were saved into a single `.txt` file, with one word per line. These words were then sorted by their *shift_score* (which is explained later in Section 3.4), and any repeated words were removed so that each word would only be classified once.

The classification was done using a transformer-based model, specifically GPT-4.1 mini. Transformer models have shown strong performance in understanding unstructured text [26]. In this case, each word was classified individually using a custom prompt designed for resume content, asking the model to assign the word to one of six predefined semantic groups:

- **Location:** Geographic or administrative areas (e.g., cities, countries, regions, demonyms). *Examples: barcelona, usa, europe, san francisco, city.*
- **Proper noun:** Proper names of people, organizations, universities, or non-English terms not referring to locations. *Examples: Microsoft, Paul, Carrefour, Securitas, Universitat.*
- **Occupation:** Job titles, professional domains, fields of study, or work-related terms. *Examples: hospital, nurse, cancer, degree, education.*
- **Skill:** Abilities, competencies, working styles, or personal traits. *Examples: teamwork, enthusiasm, problem, development, collaboration.*

- **Connectors:** Prepositions, conjunctions, and pronouns. This category was explicitly restricted to exclude nouns, verbs, adjectives, and adverbs. *Examples: im, non, like, dont, furthermore.*
- **Non-Informational:** Non-alphabetic tokens, such as numbers or symbols, that do not convey lexical meaning in isolation. *Examples: first_name, https, location, country_code, com.*

Each word was automatically routed to its respective list file (e.g., `location_list.txt`, `skill_list.txt`, etc.) based on the GPT model’s classification. In addition, all single-letter and two-letter words were included in the `connectors` group, regardless of their classification, as their semantic contribution was considered ambiguous or negligible.

Although most of the word classification was done automatically using the GPT model, all assignments were manually reviewed afterward. If a word was misclassified or unclear in its context, it was moved to a more suitable category. This helped ensure that the word groups were accurate and consistent for the experiments that followed.

The final classification produced six separate word groups, each containing a different type of term. The number of words in each group reflects the variety of vocabulary found in the resumes and forms the basis for the ablation experiments explained in the next sections.

Word Group	Number of Words
Non-Informational	1,450
Location	2,049
Propernoun	6,463
Occupation	5,074
Skill	3,031
Connectors	715
Total	18,782

Table 1: Final word count per semantic group after classification and review.

3.3 Resume Classification

3.3.1 Classification Labels

This thesis explores whether a resume screening model can pick up on sensitive personal attributes just by analyzing the text, which could lead to biased decisions. To study this, four binary classification tasks were created, each based on a different protected attribute gathered from the survey. Each task tries to predict one of these categories using only the text of the resume. The classification targets are as follows:

- **Perceived Origin:** This task labels resumes according to whether the candidate perceives themselves as foreign in the country where they reside. Label 0 corresponds to *Local* (response: “No, never”), and label 1 to *Foreign* (responses: “Yes, always” and “Yes, sometimes”). The dataset contains 517 local resumes (62.0%) and 317 foreign resumes (38.0%).
- **Gender:** This task distinguishes between gender categories. Label 0 corresponds to *Man*, and label 1 to *Woman*. Resumes from individuals “Non-binary” or “Self-Identify” were excluded due to insufficient sample size. The final dataset contains 396 Man resumes (48.5%) and 421 Woman resumes (51.5%).

- **Religion:** In this task, label 0 corresponds to candidates who identified as *Secular or Non-Religious* (including the responses “Secular” and “None”), while label 1 represents *Religious* individuals (including “Buddhism”, “Christianity”, “Hinduism”, “Islamic”, “Judaism”, and “Other”). This classification includes 147 secular resumes (27.8%) and 382 religious resumes (72.2%).
- **Sexual Orientation:** This task compares heterosexual individuals with those part of the LGBT+ community. Label 0 corresponds to *Heterosexual*, and label 1 includes *Asexual, Bisexual, Homosexual, and Self-Identify*. The dataset includes 643 heterosexual resumes (84.0%) and 122 LGBT resumes (16.0%).

These four classification settings are used to evaluate whether a model trained on resume text is capable of predicting sensitive personal attributes. Consistent class balancing and preprocessing ensure fair experimental conditions across all tasks.

3.3.2 Model Architecture

For tokenization and all classification tasks, the model used was `distilbert-base-uncased`, a lightweight and efficient transformer model available through the `Transformers` library by Hugging Face. DistilBERT is a smaller version of the original BERT model, created through knowledge distillation. It keeps most of BERT’s ability to understand language while being faster and more compact [27].

Even though it is smaller, DistilBERT has been shown to perform well on many natural language processing tasks, including binary classification [28]. To adapt it to the task, a classification layer was added on top of the transformer, allowing the model to learn from the training data. It uses contextual embeddings and self-attention to capture important information from the text.

Since most of the classification tasks in this thesis involve imbalanced classes, the training process included a weighted loss to reduce bias toward the majority class. These class weights were calculated based on how frequent each class was, so that the model would penalize mistakes in the minority class more heavily. The weights

were added directly into the loss function, helping the model learn more fairly across both classes.

Training was done using the `Trainer` API from the Hugging Face `Transformers` library. The main training settings included a learning rate of 2×10^{-5} , batch size of 8 for both training and evaluation, a weight decay of 0.01, and a maximum of 20 epochs. The model that performed best on the validation set (based on lowest evaluation loss) was automatically selected. Accuracy was used as the main metric during training.

This setup made sure that the model was aware of class imbalances and was trained under consistent and controlled conditions. This was important to allow fair comparisons across the different classification tasks based on protected attributes.

Throughout all four classification tasks, class weights were systematically adjusted across multiple training runs to explore the classifier’s full potential. Instead of relying on a single predefined weighting scheme, each sub-experiment was repeated with different class weight configurations. This approach allowed the model to be tested under a range of balance conditions, helping to identify the setup that produced the most reliable and fair results for each task, regardless of the original class distribution.

To evaluate the performance of the binary classifiers, five metrics were used: **Precision** and **Recall** for each class, plus **Accuracy** and **F1 Macro Average**. Precision tells us how many of the model’s predictions for a class were actually correct, while recall shows how well the model found all the actual examples of that class. Accuracy shows the percentage of total correct predictions, but doesn’t account for class imbalance, so it can be misleading in uneven datasets (high accuracy although low recall and precision for the minority class). F1 Macro Average averages the F1 scores of both classes equally, which helps balance the view, but it can still be misleading if one class strongly dominates the other.

To make sure the model was truly learning to distinguish between classes, the classification was considered successful only if all six metrics (precision and recall for both

classes, accuracy, and F1 Macro Average) were above 0.5. This threshold helps filter out models that perform no better than random, or that just favor the majority class. Although this rule might exclude some borderline classifiers that could still be valid, it adds a level of reliability. Among the models that meet this threshold, those with higher *Accuracy* and *F1 Macro Average* are considered stronger overall, since they reflect more balanced and consistent performance.

3.4 Lexical Shift Analysis with Shifterator

As shown in the Results chapter, some of the classification experiments confirm that a resume screening model can pick up on protected attributes, even when they aren't directly mentioned in the text. This highlights a clear risk of unintended discrimination, since certain word choices or patterns may indirectly reveal sensitive information. To better understand which words are responsible for these inferences, interpretability tools are essential. In this thesis, the *Shifterator* library is used to carry out a lexical shift analysis, helping to identify the specific terms that most influence the model's ability to tell groups apart.

Shifterator works by comparing how often each word appears in two different groups of texts and calculates how much each word contributes to the difference between them. In this case, the `JSDivergenceShift` method was used, splitting the dataset into two groups based on the real labels for each protected attribute. A `CountVectorizer` was applied to turn the resume texts into word frequency dictionaries for each group. Then, using the Jensen-Shannon divergence, the tool measured how distinct the distributions were. The final result is a set of shift scores, where the highest positive or negative values show which words most strongly signal the difference between the two groups.



Figure 2: Lexical shift graph for the *Gender* classification task.

Figure 2 shows an example of a lexical shift graph generated by Shifterator for the *Gender* classification task. The graph displays how much each word contributes to the difference between resumes labeled as *Man* and those labeled as *Woman*, with the horizontal axis showing the strength and direction of each word’s contribution. Words on the right are more common in resumes from women, while those on the left appear more often in resumes from men. In this example, as in many others throughout the thesis, the contribution is clearly unbalanced. This means that most of the strong lexical signals come from one class (in this case, *Woman*), which is why the graph mainly highlights terms on one side.

3.5 Ablation Study Design

To analyze the linguistic features most responsible for the inference of protected attributes, this thesis employs an ablation-based methodology [29]. For each of the four classification tasks (*Perceived Origin*, *Gender*, *Religion*, and *Sexual Orientation*) a series of 12 controlled experiments were conducted based on the semantic

word groups previously defined.

Each experiment involves two components: (1) resume classification using the fine-tuned DistilBERT model and (2) lexical shift analysis using Shifterator to identify the top contributing words in each class. The goal is to observe how the inclusion or exclusion of specific word groups affects the model’s ability to predict protected attributes, and to interpret which linguistic features dominate in each context.

The ablation experiments are categorized as follows:

- **Only In:** All word groups are removed from the resumes except one. This allows testing whether a single group contains sufficient information to enable classification above chance.
- **Only Out:** A single word group is removed, while the remaining ones are kept intact. This reveals whether that specific group was contributing significantly to the classification task.
- **All In:** No word groups are removed; this represents the full resume text (excluding the Non-Informational blacklist).
- **All Out:** All semantic word groups are removed, resulting in an empty resume. This acts as a control to confirm that classification fails in the absence of meaningful input.

This results in a total of 12 experiments per classification task: five *Only In*, five *Only Out*, one *All In*, and one *All Out*. The word groups considered are: *Location*, *Propernoun*, *Occupation*, *Skill*, and *Connectors*. It is important to note that the *Non-Informational blacklist* (comprising parsing tags and parsing-generated tokens) is excluded from all experiments, as these elements are not part of the original resume content and are always removed prior to classification.

Each of the 12 ablation experiments is repeated multiple times (at least 10) using different class weight configurations in the loss function. This repetition ensures

that the model’s classification potential is thoroughly explored, particularly in imbalanced settings. In the results chapter, only the best performing version of each experiment, according to the defined evaluation metrics, is reported and analyzed.

This ablation strategy helps to better understand how each group of words contributes to the detection of sensitive attributes. It offers useful insight into which types of language might act as indirect signals for demographic information in automated resume screening systems.

Chapter 4

Results

This chapter presents the results of the ablation experiments carried out to explore whether transformer-based classifiers can infer protected attributes from the text of resumes. The analysis is divided into four sections, one for each classification task studied: *Perceived Origin*, *Gender*, *Religion*, and *Sexual Orientation*.

Each section is structured in two parts. The first part shows the results of the 12 ablation experiments, reporting metrics such as precision, recall, accuracy, and F1 Macro Average. These results are organized into two tables: one for the *Only In* setups (including *All In*) and one for the *Only Out* setups (including *All Out*). Experiments where all metrics are above 0.5 are considered high recovery risk classifications.

The second part focuses on lexical analysis using shift scores generated by Shifterator. For each task, the 10 most influential words are shown for both classes. These are followed by an analysis of key word groups, where the most impactful terms are examined. Extended tables with additional word data, especially for groups with less influence, are available in the Appendix B.

4.1 Perceived Origin

4.1.1 Classification Metrics

Table 2 and Table 3 present the classification results for the *Perceived Origin* task, obtained through lexical ablation experiments. In the *All In* setup (Table 2), the classifier shows solid performance across all metrics, reaching an F1 Macro Average of 0.63 and an accuracy of 0.64. These results suggest that the model is able to infer the perceived origin of candidates fairly well from the text in their resumes.

Group Retained	Precision (Local / Foreign)	Recall (Local / Foreign)	Accuracy	F1 Macro Avg.	High Recovery Risk
Location	0.72 / 0.57	0.76 / 0.53	0.67	0.64	✓
Proper Nouns	0.66 / 0.45	0.72 / 0.38	0.59	0.55	✗
Occupation	0.64 / 0.38	0.29 / 0.73	0.46	0.45	✗
Skill	0.60 / 0.37	0.26 / 0.71	0.43	0.42	✗
Connectors	0.64 / 0.40	0.47 / 0.57	0.51	0.51	✗
All In	0.73 / 0.52	0.65 / 0.62	0.64	0.63	✓

Table 2: Performance metrics for *Only In* ablation experiments — **Perceived Origin** classification task.

Group Removed	Precision (Local / Foreign)	Recall (Local / Foreign)	Accuracy	F1 Macro Avg.	High Recovery Risk
Location	0.66 / 0.42	0.59 / 0.49	0.55	0.54	✗
Proper Nouns	0.71 / 0.49	0.63 / 0.57	0.61	0.60	✗
Occupation	0.72 / 0.51	0.65 / 0.59	0.63	0.62	✓
Skill	0.70 / 0.51	0.70 / 0.51	0.63	0.61	✓
Connectors	0.68 / 0.48	0.69 / 0.46	0.60	0.58	✗
All Out	0.67 / 0.45	0.67 / 0.46	0.59	0.56	✗

Table 3: Performance metrics for *Only Out* ablation experiments — **Perceived Origin** classification task.

Among the individual word groups, the *Location* group stands out. It is the only category that reaches a high recovery risk classifier when retained in isolation, with

an F1 Macro Average of 0.64. Additionally, its removal (*Only Out*, Table 3) causes a noticeable drop in performance (F1: 0.54), showing that location-related terms are highly informative for this classification task.

None of the other groups (*Proper Nouns*, *Occupation*, *Skill*, or *Connectors*) achieve strong classification performance on their own, as their metrics are either similar to or worse than the *All Out* (or *Connectors In*) baselines. This suggests that their contribution is either limited or redundant in this context.

Finally, it is important to highlight that in the *Only Out* experiments (Table 3), most group removals still lead to reasonably good classification results. This is probably because the *Location* group remains present in these setups, helping the model to effectively differentiate between local and foreign candidates.

4.1.2 General Lexical Contributions

Table 4 lists the top 10 words with the highest shift scores for each class in the *Perceived Origin* task. These results point to the specific lexical cues that contribute most to the classification. Most of the top words linked to the *Foreign* class are names of countries or cities, which shows that geographic terms play a strong role. On the other hand, many of the top words for the *Local* class belong to different semantic categories.

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
course	-2.58×10^{-4}	colombia	9.79×10^{-3}
training	-2.44×10^{-4}	country	5.67×10^{-3}
spain	-1.50×10^{-4}	buenos	5.47×10^{-3}
degree	-1.48×10^{-4}	aires	5.24×10^{-3}
es	-1.40×10^{-4}	lima	4.32×10^{-3}
madrid	-9.02×10^{-5}	development	3.85×10^{-3}
manager	-8.97×10^{-5}	romania	3.60×10^{-3}
education	-7.40×10^{-5}	peru	3.47×10^{-3}
higher	-6.66×10^{-5}	van	3.45×10^{-3}
quality	-5.47×10^{-5}	city	3.38×10^{-3}

Table 4: Top 10 most influential words by shift score for each class in the **Perceived Origin** classification task.

Table 4 presents the 10 words with the highest shift scores for each class in the *Perceived Origin* task. These words help to understand which lexical elements had the strongest influence on the classification. Most of the words linked to the *Foreign* class are names of countries or cities, showing the clear importance of geographic references. In contrast, many of the top words for the *Local* class come from other areas, not necessarily tied to location.

Additional Relevant Terms

When examining the top 30 words more closely (listed in Appendix A), some interesting contrasts appear. A good example is the pair *bachelor* (Foreign, position 17) and *degree* (Local, position 4). Although they are not exact synonyms, they often describe the same education level. Their placement in opposite classes suggests regional differences in how this concept is expressed.

Another relevant comparison is between *administrator* (Foreign, position 21) and *manager* (Local, position 7). These job titles may not always be interchangeable, but they are commonly used for similar roles. The fact that they are linked to opposite classes again shows how subtle language choices may reflect perceived origin.

4.1.3 Lexical Contributions by Word Group

Table 5 breaks down the top 100 influential words by word group. While the *Occupation* group includes the highest number of words (41), it is the *Location* group that stands out the most in terms of overall impact, with the highest combined shift score ($|\Delta| = 0.0574$). This supports the earlier results from the ablation experiments (see Table 2), where *Location* was the only group that enabled high recovery risk classification on its own.

Group	# in Top 100	Sum $ \Delta $
Location	30	0.0574
Propernoun	7	0.0116
Occupation	41	0.0202
Skill	15	0.0122
Connectors	3	0.0015

Table 5: Lexical group representation within the top 100 most influential words (both classes combined) — **Perceived Origin**.

To further explore these contributions, Table 6 presents the top 10 most influential location-related terms for each class. Words associated with the *Local* class are mostly European places (e.g., *spain*, *madrid*, *barcelona*, *italy*), while those associated with the *Foreign* class are primarily from Latin America (e.g., *colombia*, *lima*, *venezuela*, *cuba*).

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
spain	-2.05×10^{-3}	colombia	4.45×10^{-2}
madrid	-1.42×10^{-3}	buenos	2.48×10^{-2}
barcelona	-4.13×10^{-4}	country	2.45×10^{-2}
valencia	-2.94×10^{-4}	aires	2.38×10^{-2}
italy	-2.55×10^{-4}	lima	1.97×10^{-2}
netherlands	-2.52×10^{-4}	romania	1.64×10^{-2}
center	-1.98×10^{-4}	peru	1.58×10^{-2}
salamanca	-1.82×10^{-4}	venezuela	1.41×10^{-2}
san	-1.46×10^{-4}	city	1.37×10^{-2}
catalonia	-1.35×10^{-4}	cuba	1.32×10^{-2}

Table 6: Top 10 most influential **Location** words by shift score for each class in the **Perceived Origin** classification task.

The top words for the remaining word groups (*Proper Nouns*, *Occupation*, *Skill*, *Connectors*) can be found in Appendix B.

4.2 Gender

4.2.1 Classification Metrics

Table 7 and Table 8 show the classification performance of the model on the *Gender* task using lexical ablation experiments.

In the *All In* condition (Table 7), the classifier shows strong overall performance. It reaches high values for both precision and recall, with an F1 Macro Average and accuracy of 0.70. These results suggest that the model can effectively predict a candidate’s gender from their resume, performing clearly above the control baseline (*All Out*, Macro Average: 0.50).

Group Retained	Precision	Recall	Accuracy	F1	High Recovery
	(Man / Woman)	(Man / Woman)		Macro Avg.	Risk
Location	0.52 / 0.56	0.54 / 0.54	0.54	0.54	✓
Proper Nouns	0.59 / 0.66	0.68 / 0.56	0.62	0.62	✓
Occupation	0.70 / 0.77	0.78 / 0.68	0.73	0.73	✓
Skill	0.58 / 0.63	0.65 / 0.56	0.60	0.60	✓
Connectors	0.50 / 0.52	0.13 / 0.88	0.52	0.43	✗
All In	0.72 / 0.69	0.62 / 0.78	0.70	0.70	✓

Table 7: Performance metrics for *Only In* ablation experiments — **Gender** classification task.

In the *Only In* experiments, all word groups except *Connectors* produce high recovery risk classifiers on their own. The most important group is clearly *Occupation*, which achieves the best performance overall (F1: 0.73, Accuracy: 0.73), even slightly better than the full model in the *All In* condition. *Proper Nouns* and *Skill* also show good results (F1: 0.62 and 0.60, respectively). The *Location* group leads to a high recovery risk classifier as well, but its performance is much weaker (F1: 0.54), suggesting that location-related words are not especially useful for predicting gender.

Group Removed	Precision	Recall	Accuracy	F1	High Recovery
	(Man / Woman)	(Man / Woman)		Macro Avg.	Risk
Location	0.72 / 0.69	0.63 / 0.78	0.71	0.70	✓
Proper Nouns	0.73 / 0.77	0.76 / 0.73	0.74	0.74	✓
Occupation	0.57 / 0.60	0.59 / 0.58	0.59	0.59	✓
Skill	0.74 / 0.69	0.61 / 0.80	0.71	0.70	✓
Connectors	0.75 / 0.65	0.52 / 0.84	0.68	0.67	✓
All Out	0.62 / 0.55	0.23 / 0.87	0.67	0.50	✗

Table 8: Performance metrics for *Only Out* ablation experiments — **Gender** classification task.

The *Only Out* results (Table 8) also highlight how important the *Occupation* group is. When this group is removed, the model’s performance drops significantly (F1: 0.59). This suggests that job-related words are key for identifying gender in resumes. Still, the classifier stays above the threshold thanks to the contribution of the remaining word groups.

4.2.2 General Lexical Contributions

Table 9 lists the 10 words with the highest shift scores for each gender class in the *Gender* classification task. Most of the words are clearly connected to a person’s job or professional role, such as *assistant*, *administrative*, *engineering*, or *manager*. These results align with the importance of the *Occupation* group shown earlier.

In contrast to the *Perceived Origin* task, here there are not any pairs of near-synonyms appearing on opposite sides. However, one word stands out: *waitress*, ranked seventh for the *Woman* class. This is one of the few explicitly gendered terms and shows that the model can pick up on direct lexical markers of gender.

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
manager	-2.24×10^{-4}	administrative	9.70×10^{-3}
business	-1.96×10^{-4}	assistant	8.81×10^{-3}
language	-1.35×10^{-4}	social	5.65×10^{-3}
certificate	-1.24×10^{-4}	management	5.47×10^{-3}
course	-8.38×10^{-5}	customer	5.01×10^{-3}
development	-6.90×10^{-5}	hr	4.13×10^{-3}
program	-6.08×10^{-5}	waitress	3.73×10^{-3}
operator	-5.74×10^{-5}	city	3.58×10^{-3}
engineering	-5.39×10^{-5}	university	3.33×10^{-3}
law	-5.06×10^{-5}	lima	2.91×10^{-3}

Table 9: Top 10 most influential words by shift score for each class in the **Gender** classification task.

4.2.3 Lexical Contributions by Word Group

Table 10 summarizes the number and cumulative impact of words from each lexical group within the top 100 most influential terms. As in the previous analyses, the *Occupation* group clearly dominates, both in terms of frequency and total shift score. This confirms its critical role in the classification of gender from resumes.

Group	# in Top 100	Sum $ \Delta $
Location	21	0.0346
Propernoun	2	0.0059
Occupation	53	0.0610
Skill	20	0.0230
Connectors	1	0.0001

Table 10: Lexical group representation within the top 100 most influential words (both classes combined) — **Gender**.

To better understand this pattern, Table 11 shows the top 10 occupation-related words that influence each class the most. The results reflect common gender associations in the job market: words like *assistant*, *administrative*, and *hr* are closely tied to the *Woman* class, while terms such as *business*, *manager*, and *engineering* are more strongly linked to the *Man* class.

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
course	-4.82×10^{-4}	administrative	2.34×10^{-2}
manager	-4.44×10^{-4}	assistant	2.07×10^{-2}
business	-3.60×10^{-4}	customer	1.15×10^{-2}
certificate	-1.91×10^{-4}	hr	1.05×10^{-2}
language	-1.62×10^{-4}	waitress	9.60×10^{-3}
project	-1.49×10^{-4}	degree	6.13×10^{-3}
diploma	-1.13×10^{-4}	service	6.13×10^{-3}
services	-9.69×10^{-5}	accounting	5.81×10^{-3}
law	-8.98×10^{-5}	cashier	5.65×10^{-3}
secondary	-7.64×10^{-5}	driver	5.51×10^{-3}

Table 11: Top 10 most influential **Occupation** words by shift score for each class in the **Gender** classification task.

4.3 Religion

4.3.1 Classification Metrics

Table 12 and Table 13 present the results of the lexical ablation experiments conducted for the *Religion* classification task. Unlike previous sections, none of the evaluated configurations result in a high recovery risk classifier, as always there is a metric that falls below the established 0.50 threshold.

Group Retained	Precision	Recall	Accuracy	F1	High Recovery
	(Secular / Religious)	(Secular / Religious)		Macro Avg.	Risk
Location	0.34 / 0.77	0.52 / 0.62	0.59	0.55	X
Proper Nouns	0.32 / 0.77	0.58 / 0.53	0.55	0.52	X
Occupation	0.30 / 0.75	0.52 / 0.56	0.55	0.52	X
Skill	0.33 / 0.78	0.59 / 0.55	0.56	0.53	X
Connectors	0.32 / 0.78	0.62 / 0.51	0.54	0.52	X
All In	0.43 / 0.78	0.42 / 0.79	0.69	0.60	X

Table 12: Performance metrics for *Only In* ablation experiments — **Religion** classification task.

In the *Only In* condition (Table 12), all the individual lexical groups show low

precision for the *Secular* class, even though recall values stay relatively high. Even when all word groups are included (*All In*), the classifier still doesn't meet the threshold because both precision and recall for secular candidates are too low.

Group Removed	Precision	Recall	Accuracy	F1	High Recovery
	(Secular / Religious)	(Secular / Religious)		Macro Avg.	Risk
Location	0.36 / 0.79	0.55 / 0.62	0.60	0.56	✗
Proper Nouns	0.45 / 0.80	0.48 / 0.78	0.70	0.63	✗
Occupation	0.43 / 0.85	0.69 / 0.65	0.66	0.63	✗
Skill	0.55 / 0.79	0.38 / 0.88	0.75	0.65	✗
Connectors	0.34 / 0.77	0.52 / 0.62	0.59	0.55	✗
All Out	0.28 / 0.75	0.72 / 0.31	0.42	0.42	✗

Table 13: Performance metrics for *Only Out* ablation experiments — **Religion** classification task.

The *Only Out* experiments (Table 13) support the same conclusion. While some experiments, like removing the *Skill* group, result in slightly better accuracy, none of the models meet the threshold to be considered high recovery risk. In most cases, like the *Skill Out*, the accuracy is high only because the model fails to correctly identify secular candidates. This suggests that the model is not able to learn useful patterns for predicting religion from the text, no matter which word groups are included or left out.

One possible reason for these poor results could be the class imbalance in the dataset (72% Religious vs. 28% Secular). However, a similar imbalance was present in the *Perceived Origin* task, which still produced high recovery risk classifications. This suggests a deeper issue: the words in the resumes do not provide enough information to distinguish between secular and religious candidates effectively.

4.3.2 General Lexical Contributions

Even though the model couldn't produce high recovery risk results for the *Religion* task, a lexical analysis was still carried out using the true class labels. This analysis can still reveal patterns in word usage between *Secular* and *Religious* resumes, even

if the model was not able to use those differences successfully.

Table 14 lists the 10 most influential words for each class. Most of the words associated with the *Religious* class (like *management*, *course*, *accounting*) are common professional terms that don't seem directly linked to religion. The same is true for the *Secular* class, where words like *development*, *software*, and *language* also don't suggest any religious connection.

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
degree	-1.28×10^{-4}	course	7.55×10^{-3}
development	-9.31×10^{-5}	spain	4.92×10^{-3}
data	-8.63×10^{-5}	management	4.21×10^{-3}
design	-7.56×10^{-5}	sa	3.62×10^{-3}
computer	-6.62×10^{-5}	certificate	3.42×10^{-3}
public	-5.71×10^{-5}	venezuela	3.34×10^{-3}
valencia	-5.45×10^{-5}	berlin	3.28×10^{-3}
university	-5.34×10^{-5}	lima	3.12×10^{-3}
technician	-5.19×10^{-5}	marketing	2.85×10^{-3}
social	-5.15×10^{-5}	alicante	2.81×10^{-3}

Table 14: Top 10 most influential words by shift score for each class in the **Religion** classification task.

4.3.3 Lexical Contributions by Word Group

Looking at the broader set of the top 100 most influential words (see Table 15) confirms this finding. While the *Occupation* group appears most often and has the highest total shift score, the *Location* group has a higher average shift per word. However, this doesn't seem to carry meaningful information for the classification task. When reviewing the detailed list of location-related words (Appendix B), there are no clear patterns or terms that suggest any link to religion. This will be revisited in the Discussion chapter.

Group	# in Top 100	Sum $ \Delta $
Location	24	0.0448
Propernoun	2	0.0022
Occupation	45	0.0421
Skill	21	0.0109
Connectors	2	0.0052

Table 15: Lexical group representation within the top 100 most influential words (both classes combined) — **Religion**.

4.4 Sexual Orientation

4.4.1 Classification Metrics

Table 16 and Table 17 present the results of the lexical ablation experiments for the *Sexual Orientation* classification task. Similar to the *Religion* task, none of the configurations in these experiments resulted in a high recovery risk classifier.

Group Retained	Precision (Hetero / LGBT)	Recall (Hetero / LGBT)	Accuracy	F1 Macro Avg.	High Recovery Risk
Location	0.87 / 0.33	0.89 / 0.29	0.80	0.60	✗
Proper Nouns	0.81 / 0.13	0.52 / 0.38	0.50	0.41	✗
Occupation	0.91 / 0.28	0.71 / 0.63	0.69	0.59	✗
Skill	0.85 / 0.17	0.57 / 0.46	0.56	0.47	✗
Connectors	0.86 / 0.27	0.91 / 0.17	0.80	0.55	✗
All In	0.91 / 0.27	0.67 / 0.67	0.67	0.58	✗

Table 16: Performance metrics for *Only In* ablation experiments — **Sexual Orientation** classification task.

In the *Only In* condition (Table 16), most word groups result in high precision for the *Heterosexual* class, but very low precision and recall for the *LGBT* class. Even when all lexical groups are included (*All In*), the classifier still doesn't meet the threshold. This is mainly because the model struggles to correctly identify LGBT examples, leading to an unbalanced prediction scores.

Group Removed	Precision	Recall	Accuracy	F1	High Recovery
	(Hetero / LGBT)	(Hetero / LGBT)		Macro Avg.	Risk
Location	0.87 / 0.22	0.69 / 0.46	0.65	0.53	✗
Proper Nouns	0.90 / 0.25	0.66 / 0.63	0.65	0.56	✗
Occupation	0.87 / 0.22	0.69 / 0.46	0.65	0.53	✗
Skill	0.90 / 0.30	0.75 / 0.59	0.73	0.62	✗
Connectors	0.89 / 0.27	0.76 / 0.50	0.71	0.58	✗
All Out	0.85 / 0.18	0.71 / 0.33	0.65	0.50	✗

Table 17: Performance metrics for *Only Out* ablation experiments — **Sexual Orientation** classification task.

The *Only Out* results (Table 17) support the same conclusion. Even though some setups lead to relatively high accuracy, the precision for the LGBT class never reaches the required threshold. This shows that the model struggles to keep balanced predictions, and in every case the LGBT class ends up with either poor recall or low precision.

A likely reason for this is the strong imbalance in the dataset: 84% of the examples belong to the *Heterosexual* class, and only 16% to the *LGBT* class. This makes the model biased towards the majority class, leading to good accuracy overall but weak performance for the minority group. Also, unlike other tasks such as *Perceived Origin* (heavily influenced by *Location* terms) or *Gender* (driven by *Occupation* words), no word group stands out as a clear signal for sexual orientation. This lack of distinctive patterns makes the task especially difficult.

4.4.2 General Lexical Contributions

Although the model didn’t perform well in classification, we still carried out a lexical contribution analysis using the Shifterator tool. This analysis is based on the true labels and may still uncover meaningful language differences between resumes labeled as *Heterosexual* and *LGBT*.

Table 18 shows the 10 words that most contributed to the classification for each group. The words come from various semantic categories, including *Location* (e.g.,

spain, madrid, sweden, córdoba), *Occupation* (e.g., *manager, administrative, school*), and *Skill* (e.g., *english*). However, there’s no clear pattern or strong divide between the two classes, which reinforces the idea that language alone may not carry strong signals for this particular task.

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
spain	-4.25×10^{-4}	city	5.70×10^{-3}
training	-2.12×10^{-4}	research	4.18×10^{-3}
manager	-2.04×10^{-4}	sweden	4.03×10^{-3}
es	-1.88×10^{-4}	school	3.75×10^{-3}
sales	-1.68×10^{-4}	córdoba	3.48×10^{-3}
business	-1.67×10^{-4}	madrid	3.14×10^{-3}
administrative	-1.21×10^{-4}	english	3.06×10^{-3}
sa	-9.46×10^{-5}	country	2.72×10^{-3}
commercial	-9.39×10^{-5}	germany	2.64×10^{-3}
international	-9.21×10^{-5}	science	2.61×10^{-3}

Table 18: Top 10 most influential words by shift score for each class in the **Sexual Orientation** classification task.

4.4.3 Lexical Contributions by Word Group

To provide a broader view, Table 19 summarizes the distribution of the top 100 most influential words by word group. While the *Occupation* group contributes the highest number of terms (44), the *Location* group exhibits the largest cumulative shift score. However, this does not necessarily indicate a strong lexical signal: the classification metrics did not reflect any clear benefit from location-related terms, and the dominance of this group in the shift scores might be influenced by the strong class imbalance (84% Heterosexual vs. 16% LGBT), which can distort score magnitudes.

Group	# in Top 100	Sum $ \Delta $
Location	29	0.0486
Propernoun	3	0.0001
Occupation	44	0.0366
Skill	17	0.0144
Connectors	5	0.0038

Table 19: Lexical group representation within the top 100 most influential words (both classes combined) — **Sexual Orientation**.

Further inspection of the group-specific tables (see Appendix B) supports this interpretation. Even within the *Location* group, there is no consistent or interpretable pattern distinguishing LGBT from Heterosexual resumes. This reinforces the idea that the observed lexical differences may be coincidental or dataset-specific, rather than indicative of generalizable linguistic cues tied to sexual orientation.

Chapter 5

Discussion

5.1 Discussion

5.1.1 General Classification Capability

One of the main goals of this study was to evaluate whether textual content in resumes can be used to infer protected attributes. Based on the results from the ablation experiments, three different types of classification outcomes were observed.

The first case is represented by the *Perceived Origin* task, where the model was clearly able to identify strong lexical cues, especially related to geographic locations, that made classification possible. In this scenario, a specific group of words (e.g., location names) carries a large part of the discriminative power, making the classification task relatively straightforward.

The second case is found in the *Gender* classification task. Here, classification was successful across several word groups, with *Occupation*-related terms contributing the most. However, even when this group was removed, the model still achieved high recovery risk performance by relying on other lexical signals. This indicates a more distributed use of information across multiple semantic categories.

Finally, the third case involves *Religion* and *Sexual Orientation*. In both tasks,

the classifier failed to achieve high recovery risk results regardless of which lexical features were used. Even when all word groups were included, the model was unable to capture meaningful patterns to differentiate between classes. This suggests that either the labels are not strongly reflected in the text, or that the available lexical cues are too subtle for the model to learn from.

Perceived Origin

Among the four classification tasks studied, *Perceived Origin* is clearly the one with the strongest and most direct link between the label and the lexical information available in the resume. The model relies primarily on location-related terms, such as country or city names, to infer whether a candidate is perceived as *Local* or *Foreign*. This is not only evident in the lexical contribution tables but also in the classification metrics, which show a clear dependency on the presence of the *Location* word group.

In the ablation experiments, when the classifier is given access only to location-related words (*Only In*), it already reaches a high recovery risk classification performance. Conversely, removing this group (*Only Out*) causes a noticeable drop in the model's ability to make accurate predictions. This shows that the *Location* category is not just important but central to the classification of perceived origin.

This pattern is highly intuitive: when the goal is to infer where someone is perceived to be from, the places they mention in their resume are naturally the most revealing clues. While other word groups such as *Occupation* or *Skill* also contribute to the classification with some key words, their effect is secondary. The presence of specific non-European place names, particularly from Latin America (e.g., *Colombia*, *Peru*, *Venezuela*), plays a decisive role in the model's decision-making, and is a reflection of the demographic composition of the dataset used.

Despite this strong association, the overall classification metrics are not exceptionally high, and this is to be expected. First, it is important to consider that the task is based on perceived origin rather than actual nationality or objective geographic background. This label is subjective: someone born abroad but who has spent most

of their life in Europe might be perceived as local, while another individual born in Europe but who has studied or worked in another continent could be seen as foreign.

Second, the dataset does not include the country in which the person currently resides or identifies as local. This adds another layer of ambiguity. While it is known that all respondents work in Europe, there is no way to align their current location with their perceived origin. For example, a Spanish candidate working in Germany might perceive themselves as foreign in that context, but in the dataset, where Spanish resumes likely dominate the local category, that person could be misclassified.

These ambiguities mean that the model often picks up on more clear-cut signals, such as whether someone has lived, studied, or worked outside Europe. Thus, the classification task effectively becomes a detection of whether the candidate has international experience in non-European countries, particularly Latin American ones. This bias is likely driven by the dataset’s composition and the labels provided, but extrapolating this to a real-case, a situation like this, could result in a bias against people from Latin America.

To improve future studies on perceived origin, it would be beneficial to include information about both the candidate’s country of origin and the country where they currently live or work. This would allow for a more targeted and nuanced analysis, particularly in intra-European cases where perceived foreignness is less clear-cut. Still, the current results show that even under these limitations, transformer models are capable of detecting meaningful patterns, especially in cases where the perceived origin differs significantly from the European norm. This highlights the importance of considering this attribute in fairness evaluations and reinforces the potential for unintended bias in automated systems based on resume text.

Gender

The *Gender* classification task also produced successful results, although the link between the label and specific word groups is less direct than in the case of *Perceived Origin*. Here, the predictive capacity of the model appears to be more evenly dis-

tributed across different lexical categories, rather than relying on a single dominant group. All word groups except *Connectors* enabled high recovery risk classifiers when isolated in the *Only In* experiments, and the combined condition (*All In*) achieved strong overall metrics.

An important factor that likely contributed to these results is the balanced nature of the dataset. Unlike other tasks in this study, the distribution of the gender labels is nearly even, with 48% of samples labeled as *Man* and 52% as *Woman*. This gives the model equal exposure to both classes during training, which supports the learning of representative linguistic patterns for each.

Among all word groups, *Occupation* stands out as the most influential. It not only achieves the highest classification scores when used alone but also causes the most significant drop in performance when removed. This confirms its central role in enabling gender classification. The presence of occupation-specific terms such as *assistant*, *hr*, and *cashier* for the *Woman* class, and *manager*, *business*, and *engineering* for the *Man* class, reflects well-established sociolinguistic patterns. These associations, while based on stereotypes, are backed by the data in this dataset and observable in the lexical shift scores.

Although these job terms are not strictly tied to one gender, they tend to appear disproportionately due to real-world inequalities in the labor market. As a result, models trained on such data learn to associate these roles with specific gender labels, not because of explicit discrimination, but because of statistically grounded patterns. This is confirmed through the lexical analysis with Shifterator, which independently measures word distributions across classes without being influenced by model behavior.

From an ethical standpoint, this phenomenon raises concerns. On one hand, it could be argued that certain word associations may unintentionally protect candidates. For example, a female engineer might be misclassified as male based on technical vocabulary, potentially avoiding gender-based discrimination in screening systems. However, this effect is unreliable and potentially harmful. If the model does detect

her gender, the same engineering-related terms might then become detrimental if the classifier has learned that engineering is more typical of male candidates. In scenarios where gender is interpreted negatively, the presence of non-matching vocabulary could worsen the outcome. Moreover, applicants from female-dominated fields (such as *administrative*) might be unfairly penalized for simply belonging to an occupational sector correlated with women, especially when the target job is unrelated to the background.

Unlike in the case of *Perceived Origin*, where one group (*Location*) alone carried most of the predictive signal, gender classification appears more distributed. Groups such as *Skill*, *Proper Nouns*, and to a lesser extent *Location*, also contribute meaningfully. This suggests that gender-related cues are embedded in a wider range of linguistic features and not limited to job titles.

These cues are not always explicit but, like the *Occupational* cues, follow social or statistical patterns. For instance, *Skill* terms linked to women tend to be more socially oriented (e.g., *care*, *communications*, *relations*, *empathetic*), whereas those linked to men relate more to knowledge and analysis (e.g., *development*, *analysis*, *knowledge*). These trends behave similarly to occupation-based associations. In the case of *Proper Nouns*, although the differences are more subtle, tools or platforms more common in tertiary-sector work (statistically more common to women [30]), such as *Microsoft*, *Adobe*, or *Google*, appear more frequently in resumes classified as female, a pattern also discussed in the literature. Finally, while the *Location* group also produces a high recovery risk classifier, its role is likely related to biases specific to this dataset rather than generalizable gender patterns.

Religion and Sexual Orientation: Low Recovery Risk Classifications

In contrast to the *Perceived Origin* and *Gender* classification tasks, the models trained to predict *Religion* and *Sexual Orientation* did not achieve high recovery risk results in any of the ablation experiments. Despite using the same methodology and evaluation thresholds, none of the configurations produced a classifier with balanced performance across both classes. These failures point to deeper issues, both

methodological and conceptual, that affect the detectability of these two protected attributes through text classification.

There are several important distinctions that separate these two tasks from the others. First, unlike *Perceived Origin* or *Gender*, which are both binary classifications with clearly defined opposing labels (e.g., *Local vs. Foreign* or *Man vs. Woman*), the categories used in *Religion* and *Sexual Orientation* are far more complex. In both cases, the binary label groups actually contain multiple subcategories: the *Religious* class includes Christians, Muslims, Jews, Hindus, and others; the *LGBT* class includes homosexual, bisexual, asexual, and other identities. These internal variations within each class introduce heterogeneity in the associated language, potentially making intra-class differences stronger than inter-class ones. As a result, the model may struggle to find consistent linguistic patterns that separate the two broad groups.

A second and very relevant factor is class imbalance, especially in the case of *Sexual Orientation*. In this task, only 16% of the resumes (122 examples) correspond to LGBT individuals. This makes it very difficult for the model to identify patterns tied to this minority class, particularly when no single word group (like *Location* or *Occupation*) appears to provide strong discriminative features. In the absence of a clearly dominant lexical signal, the model tends to default to predicting the majority class, leading to poor recall or precision for the minority.

Third, it is possible that the language used in resumes does not, in fact, contain sufficient lexical cues to infer religious beliefs or sexual orientation in a generalizable way. Resumes are typically written in a professional and neutral tone, and candidates may intentionally avoid revealing information about these personal aspects. While the linguistic differences might exist in some cases, they may be too subtle or inconsistent to be captured by the model, especially given the small sample size available in this study.

It is also worth noting that, even if some indirect signals exist, their detectability is likely reduced by the way labels are grouped. For example, *Location* was hypothe-

sized to be a potential proxy for religion, as religious affiliation often correlates with geographic or cultural background. However, since all religious groups were merged into a single *Religious* class, the classifier may have been unable to learn any signal. For instance, both *Christian* and *Secular* candidates are commonly found in Europe, making it hard for location terms to act as reliable discriminators. Creating finer-grained or more targeted binary splits could have helped, but the dataset was too imbalanced to support that approach.

Ultimately, the main limitation in this part of the study is the size and structure of the dataset. A larger and more balanced dataset, both in terms of sample count and class definition, would be needed to determine whether religion or sexual orientation can be inferred from resume text with acceptable accuracy. With the current data, there is no strong evidence that such classification is feasible, at least not using lexical cues alone.

5.1.2 Keyword-Level Signals

Although the main focus of this thesis has been on analyzing broad lexical categories (e.g., *Occupation*, *Location*) and their impact on the classification of protected attributes, the word-level analysis performed with Shifterator reveals that some individual terms can have a disproportionately large effect on classification outcomes. In the tasks where classification was successful, *Perceived Origin* and *Gender*, two types of influential words stand out: synonym pairs used differently across classes, and terms that directly encode protected attributes in their meaning.

Lexical Synonyms Across Classes

In the *Perceived Origin* task, several of the most influential terms did not belong to the dominant category of *Location*, but still contributed significantly to classification. Some of these were near-synonyms, words with similar meanings but distributed asymmetrically across classes. This phenomenon can be understood through sociolinguistic patterns: individuals with different cultural or linguistic backgrounds may refer to the same concept using different words. These subtle variations create

strong cues for classifiers.

A notable example is the pair *bachelor* (Foreign, rank 17) and *degree* (Local, rank 4). While not perfect synonyms, since *degree* is more general, both terms often refer to the same level of education. The term *bachelor* is more common in English-speaking contexts like the United States and the United Kingdom, which are influential in many Latin American countries. Meanwhile, some European candidates often use the broader term *degree*, especially in Spain. This difference becomes a linguistic signal that the model can pick up on, even though the intended meaning is the same.

Another similar case involves *administrator* (Foreign, rank 21) and *manager* (Local, rank 7). Although they do not always denote the exact same role, they are often used interchangeably to describe positions of responsibility. Their divergent usage across classes provides another clear example of how lexical choice can reflect underlying demographic differences and be leveraged by the model for classification.

Attribute-Embedded Terms

A second type of influential term consists of words that directly encode information about the protected attribute within their semantic meaning. These are particularly relevant in the context of the *Gender* classification task. Although English is relatively gender-neutral compared to languages like Spanish or French, where gender often appears in adjectives or noun endings, some English words, especially occupational terms, still carry explicit gender markers.

One clear example from the results is the word *waitress*, which appears as the seventh most influential word for the *Woman* class. These terms inherently include gender information, and any classifier with basic linguistic understanding would be able to associate them with the corresponding gender. This shows how single words can unintentionally expose protected attributes, even if candidates do not explicitly mention them.

This finding underscores the importance of individual lexical choices when considering attribute privacy. While high-level features may contribute to overall patterns,

a single word with embedded gender, origin, or cultural meaning can be enough to reveal sensitive information.

An interesting counterexample can be found in the *Sexual Orientation* task. Despite the presence of the acronym *LGBT* in some resumes, the word did not appear among the most influential terms. This is likely due to its extremely low frequency in the dataset, which limited its impact both in training the model and in Shifterator’s shift score computation. Unlike the model classifier, Shifterator does not possess embedded linguistic knowledge; it simply measures distributional differences. Therefore, even a semantically loaded word will go unnoticed if it does not appear frequently enough to affect token distributions significantly.

In summary, individual keywords can sometimes carry more predictive weight than entire categories. These terms can either reflect subtle sociolinguistic patterns (as with synonyms) or encode protected information directly (as with gendered occupational nouns). When present in sufficient quantities, these words can substantially affect classifier behavior, even when broader category-level patterns are weak.

5.1.3 Implications for the Protection of Demographic Attributes

The findings of this thesis demonstrate that transformer-based classifiers are capable of detecting protected demographic attributes in resumes, even when these attributes are not explicitly mentioned. As shown throughout the experiments, information embedded in word choices, semantic categories, or even subtle lexical patterns can lead models to infer characteristics such as gender and perceived origin. This raises important concerns: if these models are trained on biased data, whether due to historical patterns, societal inequalities, or imbalanced hiring practices, they may replicate and reinforce discriminatory behavior in automatic resume screening processes.

For this reason, both job applicants and recruiters must be aware of the risks associated with algorithmic decision-making. From the applicant’s perspective, certain precautions can help limit unwanted exposure of protected attributes. Specifically,

candidates should avoid using “attribute-embedded” words: terms that explicitly include demographic cues in their meaning. For instance, in the case of gender, using the masculine job titles (e.g., *waiter* instead of *waitress*, or avoiding gendered language altogether) can reduce the risk of revealing gender information to the model. In cases where synonymous terms exist, choosing the version more common across groups (or at least less correlated with one specific minority class) can help reduce the chance of unintended profiling.

However, for broader lexical groups such as *Occupation*, *Skill*, or *Location*, it is neither realistic nor beneficial for candidates to avoid these categories. These terms are essential for communicating qualifications and professional backgrounds. Removing or altering them would make the resume less informative and could negatively impact the candidate’s chances in any selection process, algorithmic or otherwise.

Therefore, the responsibility for protecting against biased model behavior falls on employers and developers of Resume Screening Systems. Organizations that aim to promote fairness and reduce discrimination in hiring should implement strategies that mitigate the impact of sensitive attributes during the automatic filtering phase. This is especially relevant for the first stages of screening, where decisions are often made without human oversight.

One promising approach is the partial anonymization or masking of specific word groups during the initial selection phase. For example, geographic information, shown to be a strong proxy for perceived origin, could be removed or hidden from the first pass of screening. Human evaluators, in later stages, would still have access to this information when contextually appropriate, but it would not influence early automatic decisions that are more prone to undetected bias.

In addition, the development and deployment of screening algorithms must include rigorous bias mitigation techniques. Existing strategies such as reweighing, data augmentation, or adversarial debiasing can help reduce the influence of biased patterns in the training data. It is equally important to conduct ethical evaluations of the models used, especially when applying techniques like bootstrapping or self-

supervision, which may inadvertently reinforce latent biases.

Overall, ensuring fairness in automatic screening systems requires a combined effort: candidates must be aware of how linguistic choices can reveal protected information, while recruiters and model developers must proactively adopt technical and procedural safeguards to minimize discrimination. Only through such coordinated practices can we move toward more equitable and accountable use of AI in hiring.

5.2 Conclusions

5.2.1 General Conclusions

This thesis has studied whether transformer-based models can detect protected demographic attributes from resume text, even when this information is not stated directly. To explore this, a set of lexical ablation experiments were carried out using a dataset of resumes, focusing on four attributes: *Perceived Origin*, *Gender*, *Religion*, and *Sexual Orientation*.

The results show that some of these attributes can indeed be inferred through language. In particular, the model was able to classify *Perceived Origin* and *Gender* with satisfactory performance. In the case of origin, location-related words (like countries or cities) played a central role. For gender, terms related to occupation were the most relevant, although other categories such as skills or personal traits also helped the model make distinctions.

On the other hand, the classification tasks for *Religion* and *Sexual Orientation* did not show good results. This may be because the differences in language between the groups were too small or unclear, or because there wasn't enough data for the model to learn meaningful patterns.

Apart from the classification performance, a second part of the study focused on analysing the importance of specific words using the Shifterator tool. This analysis helped identify both broad trends like the importance of occupations and individual words that had a strong influence on classification, including some near-synonyms

or words that indirectly reveal protected information.

In short, the findings confirm that resume texts often include subtle linguistic clues that allow AI models to guess sensitive attributes. Even though candidates may not write this information directly, the words they use can give away demographic details such as gender or perceived origin.

This is especially relevant for automated screening systems, where decisions are made without human review. If the model has learned biases from the training data, then detecting these attributes becomes a risk. Candidates could be unfairly excluded based on characteristics that should not affect their chances, like gender, ethnicity, or religion.

Understanding which attributes are more visible through language helps highlight where protection efforts should be focused. If some words act as strong proxies for sensitive traits, we can think about ways to reduce their influence, either by modifying the data, adapting the model, or adding post-processing steps to correct unfair outcomes.

Ultimately, these results support the idea that automatic systems should be carefully designed and evaluated to avoid reproducing social biases. This includes being aware of how much demographic information can be inferred from language, and taking active steps to reduce the potential for discrimination.

5.2.2 Limitations

This study presents a number of limitations that should be taken into account when interpreting the results. The most important one is the size of the dataset used. With only 921 resumes available, the amount of data is relatively small for training and evaluating language models. This limitation is partly due to the fact that the original goal of the data collection was to conduct a qualitative study [8], where a smaller number of detailed responses was sufficient. As a result, the dataset was not designed with large-scale machine learning in mind. In contrast, most commercial resume screening systems are trained on much larger datasets, often containing tens

or hundreds of thousands of examples. Therefore, the conclusions drawn here may not fully reflect how large-scale systems behave in real-world scenarios.

Another limitation is the class imbalance found in some of the protected attributes. In particular, the *Sexual Orientation* attribute has a very uneven distribution, with only 16% of the samples belonging to the LGBT group. This makes it difficult for the model to learn meaningful patterns about that group and also limits the ability of Shifterator to detect representative word-level differences. Similar issues affect the *Religion* attribute, where one class combines many different identities (e.g., Christianity, Judaism, Islam), which could lead to internal variation that weakens the model’s ability to classify accurately.

The transformer model used in this thesis, *DistilBERT*, is a simplified version of BERT, designed to be faster and lighter, while retaining performance [31]. This makes DistilBERT an efficient and reliable choice for studies like this one, where computational resources are limited. However, it is important to acknowledge that larger and more powerful models have recently demonstrated superior understanding and accuracy in complex language tasks [32].

Therefore, although DistilBERT provides a reasonable and well-established baseline, it is likely that newer and more advanced models, especially those trained specifically for resume screening or built with cutting-edge architectures, could detect subtler linguistic patterns or hidden biases. In this sense, the findings in this thesis are likely conservative, and future research using more powerful models may uncover further risks or nuances in the automatic detection of protected attributes.

Finally, it is important to mention that the classification part (using the transformer) and the word-level analysis (using Shifterator) were done independently. That is, the words highlighted by Shifterator are based only on the real labels and word frequencies, not on what the model actually learned. Similarly, the model was trained without any direct influence from the shift score analysis. While it is reasonable to expect some overlap between both views, we cannot guarantee that the patterns found by Shifterator were the same ones used by the classifier. Some methods, such

as LIME or SHAP, offer the possibility to inspect model-specific explanations, but they typically focus on individual predictions and would require far more computing power to apply to the full dataset.

5.2.3 Future Work

This thesis has demonstrated that transformer-based models can infer certain protected attributes from resume text, highlighting both the potential and the risks of using automated screening systems. However, several aspects of the current study could be extended or refined in future research.

First, the most immediate improvement would be the use of a larger and more diverse dataset. With only 921 resumes, this study faced significant limitations in data availability, especially for minority classes such as *LGBT* or *Secular*. A larger dataset would allow for more balanced training, more robust classifier evaluation, and a more detailed analysis of linguistic patterns across subgroups.

Second, integrating model interpretability methods could enhance the analysis of how classifiers make decisions. Techniques such as SHAP or LIME could be used to complement the Shifterator-based word contribution analysis by directly inspecting the model's internal reasoning. These methods, if applied at scale, could offer a more faithful mapping between learned weights and observed word-level statistics.

Third, future studies could experiment with different label formulations. In this thesis, all classification tasks were binary, but some classes grouped together heterogeneous identities (e.g., multiple religions or orientations). Exploring alternative label groupings or multi-label classification strategies might give richer insights into how models perceive and reproduce social categories in text.

Finally, a natural extension of this work would be to apply and evaluate bias mitigation techniques directly on the dataset. Approaches such as data reweighing, augmentation, or adversarial debiasing could be used to reduce the presence of proxy signals related to protected attributes. Running the same classification experiments on debiased data would allow for a direct comparison of model behavior before and

after mitigation, offering practical insights into the effectiveness of different bias-reduction strategies.

Overall, expanding this line of work with better data, stronger models, and more advanced analytical tools would provide a clearer picture of how automatic systems process sensitive information, and how they might be improved to reduce unintended discrimination.

List of Figures

1	Overview of the methodological pipeline including word classification, ablation, lexical shift analysis, and resume classification.	12
2	Lexical shift graph for the <i>Gender</i> classification task.	21

List of Tables

1	Final word count per semantic group after classification and review.	17
2	Performance metrics for <i>Only In</i> ablation experiments — Perceived Origin classification task.	25
3	Performance metrics for <i>Only Out</i> ablation experiments — Perceived Origin classification task.	25
4	Top 10 most influential words by shift score for each class in the Perceived Origin classification task.	26
5	Lexical group representation within the top 100 most influential words (both classes combined) — Perceived Origin	28
6	Top 10 most influential Location words by shift score for each class in the Perceived Origin classification task.	28
7	Performance metrics for <i>Only In</i> ablation experiments — Gender classification task.	29
8	Performance metrics for <i>Only Out</i> ablation experiments — Gender classification task.	30
9	Top 10 most influential words by shift score for each class in the Gender classification task.	31
10	Lexical group representation within the top 100 most influential words (both classes combined) — Gender	31
11	Top 10 most influential Occupation words by shift score for each class in the Gender classification task.	32
12	Performance metrics for <i>Only In</i> ablation experiments — Religion classification task.	32

13	Performance metrics for <i>Only Out</i> ablation experiments — Religion classification task.	33
14	Top 10 most influential words by shift score for each class in the Religion classification task.	34
15	Lexical group representation within the top 100 most influential words (both classes combined) — Religion	35
16	Performance metrics for <i>Only In</i> ablation experiments — Sexual Orientation classification task.	35
17	Performance metrics for <i>Only Out</i> ablation experiments — Sexual Orientation classification task.	36
18	Top 10 most influential words by shift score for each class in the Sexual Orientation classification task.	37
19	Lexical group representation within the top 100 most influential words (both classes combined) — Sexual Orientation	38
20	Top 30 most influential words by shift score for each class in the Perceived Origin classification task.	64
21	Top 30 most influential words by shift score for each class in the Gender classification task.	65
22	Top 30 most influential words by shift score for each class in the Religion classification task.	66
23	Top 30 most influential words by shift score for each class in the Sexual Orientation classification task.	67
24	Top 10 most influential Propernoun words by shift score — Perceived Origin	69
25	Top 10 most influential Occupation words by shift score — Perceived Origin	69
26	Top 10 most influential Skill words by shift score — Perceived Origin	70
27	Top 10 location-related words by shift score in the Gender classification task.	70

28	Top 10 proper noun words by shift score in the Gender classification task.	71
29	Top 10 skill-related words by shift score in the Gender classification task.	71
30	Top 10 location-related words by shift score for each class in the Religion classification task.	72
31	Top 10 proper noun words by shift score for each class in the Religion classification task.	72
32	Top 10 occupation-related words by shift score for each class in the Religion classification task.	73
33	Top 10 skill-related words by shift score for each class in the Religion classification task.	73
34	Top 10 location-related words by shift score for each class in the Sexual Orientation classification task.	74
35	Top 10 propernoun-related words by shift score for each class in the Sexual Orientation classification task.	74
36	Top 10 occupation-related words by shift score for each class in the Sexual Orientation classification task.	75
37	Top 10 skill-related words by shift score for each class in the Sexual Orientation classification task.	75

Bibliography

- [1] Derous, E. & Ryan, A. M. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal* **29**, 113–130 (2019). URL <https://onlinelibrary.wiley.com/doi/full/10.1111/1748-8583.12217><https://onlinelibrary.wiley.com/doi/abs/10.1111/1748-8583.12217><https://onlinelibrary.wiley.com/doi/10.1111/1748-8583.12217>.
- [2] Lippens, L., Dalle, A., D’hondt, F., Verhaeghe, P. P. & Baert, S. Understanding ethnic hiring discrimination: A contextual analysis of experimental evidence. *Labour Economics* **85**, 102453 (2023).
- [3] Wilson, K. & Caliskan, A. Gender, race, and intersectional bias in resume screening via language model retrieval (2024). URL <https://arxiv.org/abs/2407.20371v2>.
- [4] Lippens, L. Computer says ‘no’: Exploring systemic bias in chatgpt using an audit approach. *Computers in Human Behavior: Artificial Humans* **2**, 100054 (2023). URL <http://arxiv.org/abs/2309.07664><http://dx.doi.org/10.1016/j.chbah.2024.100054>.
- [5] Deshpande, K. V., Pan, S. & Foulds, J. R. Mitigating demographic bias in ai-based resume filtering. *UMAP 2020 Adjunct - Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* 268–275 (2020). URL https://www.researchgate.net/publication/342906317_Mitigating_Demographic_Bias_in_AI-based_Resume_Filtering.

- [6] Lacroux, A. & Martin-Lacroux, C. Anonymous résumés: An effective pre-selection method? *International Journal of Selection and Assessment* **28**, 98–111 (2020). URL <https://onlinelibrary.wiley.com/doi/full/10.1111/ijsa.12275><https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsa.12275><https://onlinelibrary.wiley.com/doi/10.1111/ijsa.12275>.
- [7] Adamovic, M. Analyzing discrimination in recruitment: A guide and best practices for resume studies. *International Journal of Selection and Assessment* **28**, 445–464 (2020). URL <https://onlinelibrary.wiley.com/doi/full/10.1111/ijsa.12298><https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsa.12298><https://onlinelibrary.wiley.com/doi/10.1111/ijsa.12298>.
- [8] Bhatia, K. V., Capasso, M., Arora, P., Castillo, C. & Saldivar, J. Proxy discrimination risks in hiring: A qualitative analysis of a set of real cvs (2024). URL <https://papers.ssrn.com/abstract=5048771>.
- [9] Chu, H., Men, L. R., Liu, S., Yuan, S. & Sun, Y. Nationality, race, and ethnicity biases in and consequences of detecting ai-generated self-presentations (2024). URL <https://arxiv.org/abs/2412.18647v1>.
- [10] Behaghel, L., Crépon, B. & Barbanchon, T. L. Unintended effects of anonymous resumes (2014).
- [11] Åslund, O. & Skans, O. N. Do anonymous job application procedures level the playing field? *Industrial and Labor Relations Review* **65**, 82–107 (2011).
- [12] Parasurama, P. & Sedoc, J. Degendering resumes for fair algorithmic resume screening (2021). URL <https://arxiv.org/pdf/2112.08910>.
- [13] Dovidio, J. F. & Gaertner, S. L. Aversive racism and selection decisions: 1989 and 1999 (2000).
- [14] Bartkoski, T., Lynch, E., Witt, C. & Rudolph, C. A meta-analysis of hiring discrimination against muslims and arabs. *Personnel Assessment and Decisions: Number 4* (2018). URL <https://doi.org/10.25035/pad.2018.02.001>.

- [15] Gohar, U. & Cheng, L. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *IJCAI International Joint Conference on Artificial Intelligence* **2023-August**, 6619–6627 (2023). URL <http://arxiv.org/abs/2305.06969><http://dx.doi.org/10.24963/ijcai.2023/742>.
- [16] Morina, G., Oliinyk, V., Waton, J., Marusic, I. & Georgatzis, K. Auditing and achieving intersectional fairness in classification problems (2019). URL <https://arxiv.org/pdf/1911.01468>.
- [17] Chen, Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications* **10**, 1–12 (2023). URL <https://www.nature.com/articles/s41599-023-02079-x>.
- [18] Beretta, A. *et al.* Requirements of explainable ai in algorithmic hiring (2024). URL <https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems>.
- [19] Zhou, J., Chen, F. & Holzinger, A. Towards explainability for ai fairness. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **13200 LNAI**, 375–386 (2022). URL https://link.springer.com/chapter/10.1007/978-3-031-04083-2_18.
- [20] Binns, R. *et al.* 'it's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions. *Conference on Human Factors in Computing Systems - Proceedings* **2018-April** (2018). URL [/doi/pdf/10.1145/3173574.3173951?download=true](https://doi.org/10.1145/3173574.3173951?download=true).
- [21] Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* 4356–4364 (2016). URL <https://arxiv.org/pdf/1607.06520>.
- [22] Meade, N., Poole-Dayana, E. & Reddy, S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *Proceedings of the*

- Annual Meeting of the Association for Computational Linguistics* **1**, 1878–1898 (2021). URL <https://arxiv.org/pdf/2110.08527>.
- [23] Rus, C., Luppés, J., Oosterhuis, H. & Schoenmacker, G. H. Closing the gender wage gap: Adversarial fairness in job recommendation. *CEUR Workshop Proceedings* **3218** (2022). URL <https://arxiv.org/pdf/2209.09592>.
- [24] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 3323–3331 (2016). URL <https://arxiv.org/pdf/1610.02413>.
- [25] Fabris, A. *et al.* Fairness and bias in algorithmic hiring: a multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology* **1** (2024). URL <http://arxiv.org/abs/2309.13933><http://dx.doi.org/10.1145/3696457>.
- [26] Yerramreddy, D. R., Marasani, J., Gowtham, P. S. V., Abhishek, S. & Anjali. An empirical analysis of topic categorization using palm, gpt and bert models. *2023 Innovations in Power and Advanced Computing Technologies, i-PACT 2023* (2023).
- [27] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019). URL <https://arxiv.org/pdf/1910.01108>.
- [28] Büyüköz, B. & Hürriyetöç, A. Analyzing the generalizability of deep contextualized language representations for text classification (2023). URL <https://arxiv.org/pdf/2303.12936>.
- [29] Kumari, N. *et al.* Ablating concepts in text-to-image diffusion models. *Proceedings of the IEEE International Conference on Computer Vision* 22634–22645 (2023). URL <https://arxiv.org/pdf/2303.13516>.
- [30] Biesialska, M., Solans, D., Luque, J. & Segura, C. On the relationship of social gender equality and grammatical gender in pre-trained large language models (2024). URL <http://ceur-ws.org>.

- [31] Barbon, R. S. & Akabane, A. T. Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: A case study. *Sensors 2022, Vol. 22, Page 8184* **22**, 8184 (2022). URL <https://www.mdpi.com/1424-8220/22/21/8184/html><https://www.mdpi.com/1424-8220/22/21/8184>.
- [32] Kheddar, H. Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. *Information Fusion* **124**, 103347 (2025). URL <http://arxiv.org/abs/2408.07583>.

Appendix A

Additional Lexical Top Words

This appendix provides complementary lexical analysis results that were not included in the main body of the thesis for readability and focus. These additional tables highlight influential terms ranked by their contribution to each classification task according to the shift score metric. While the most relevant and representative lexical items were already discussed in Chapter 4.

These tables may offer further insight into the linguistic patterns observed across the different tasks, especially in relation to protected attributes.

A.1 Perceived Origin

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
course	-2.58e-4	colombia	9.79e-3
training	-2.44e-4	country	5.67e-3
spain	-1.50e-4	buenos	5.47e-3
degree	-1.48e-4	aires	5.24e-3
es	-1.40e-4	lima	4.32e-3
madrid	-9.02e-5	development	3.85e-3
manager	-8.97e-5	romania	3.60e-3
education	-7.40e-5	peru	3.47e-3
higher	-6.66e-5	van	3.45e-3
quality	-5.47e-5	city	3.38e-3
administrative	-4.85e-5	venezuela	3.12e-3
operator	-4.77e-5	cuba	2.91e-3
level	-4.61e-5	planning	2.57e-3
word	-4.32e-5	melilla	2.44e-3
information	-4.30e-5	argentina	2.40e-3
learning	-4.19e-5	university	2.38e-3
community	-4.12e-5	bachelor	2.32e-3
law	-4.01e-5	management	2.14e-3
specialist	-3.53e-5	chile	2.11e-3
production	-3.38e-5	havana	1.98e-3
italy	-3.27e-5	administrator	1.90e-3
internship	-3.25e-5	leone	1.85e-3
netherlands	-3.22e-5	corunna	1.84e-3
computer	-3.17e-5	cancer	1.83e-3
sl	-3.14e-5	gestión	1.74e-3
valencia	-2.95e-5	surgery	1.74e-3
analytics	-2.59e-5	united	1.68e-3
occupational	-2.59e-5	process	1.66e-3
foundation	-2.54e-5	florence	1.62e-3
hr	-2.28e-5	copenhagen	1.62e-3

Table 20: Top 30 most influential words by shift score for each class in the **Perceived Origin** classification task.

A.2 Gender

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
manager	-2.24e-4	administrative	9.70e-3
business	-1.96e-4	assistant	8.81e-3
language	-1.35e-4	social	5.65e-3
certificate	-1.24e-4	management	5.47e-3
course	-8.38e-5	customer	5.01e-3
development	-6.90e-5	hr	4.13e-3
program	-6.08e-5	waitress	3.73e-3
operator	-5.74e-5	city	3.58e-3
engineering	-5.39e-5	university	3.33e-3
law	-5.06e-5	lima	2.91e-3
services	-4.79e-5	degree	2.83e-3
information	-4.61e-5	melilla	2.83e-3
sa	-4.51e-5	service	2.82e-3
diploma	-4.46e-5	country	2.75e-3
network	-4.06e-5	van	2.56e-3
seville	-4.02e-5	accounting	2.51e-3
technical	-3.86e-5	research	2.48e-3
python	-3.29e-5	zurich	2.48e-3
usa	-2.94e-5	training	2.44e-3
states	-2.82e-5	canada	2.42e-3
windows	-2.77e-5	zaragoza	2.42e-3
qualification	-2.71e-5	communication	2.39e-3
level	-2.38e-5	paraguay	2.36e-3
council	-2.38e-5	romania	2.32e-3
legal	-2.28e-5	cashier	2.26e-3
secondary	-2.25e-5	shop	2.07e-3
industry	-2.20e-5	toledo	2.02e-3
expert	-2.10e-5	córdoba	2.00e-3
head	-2.06e-5	environmental	1.98e-3
programming	-2.03e-5	communications	1.91e-3

Table 21: Top 30 most influential words by shift score for each class in the **Gender** classification task.

A.3 Religion

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
degree	-1.28×10^{-4}	course	7.55×10^{-3}
development	-9.31×10^{-5}	spain	4.92×10^{-3}
data	-8.63×10^{-5}	management	4.21×10^{-3}
design	-7.56×10^{-5}	sa	3.62×10^{-3}
computer	-6.62×10^{-5}	certificate	3.42×10^{-3}
public	-5.71×10^{-5}	venezuela	3.34×10^{-3}
valencia	-5.45×10^{-5}	berlin	3.28×10^{-3}
university	-5.34×10^{-5}	lima	3.12×10^{-3}
technician	-5.19×10^{-5}	marketing	2.85×10^{-3}
social	-5.15×10^{-5}	alicante	2.81×10^{-3}
team	-4.86×10^{-5}	business	2.73×10^{-3}
engineer	-4.56×10^{-5}	microsoft	2.68×10^{-3}
teaching	-4.52×10^{-5}	sales	2.66×10^{-3}
quality	-4.43×10^{-5}	manager	2.59×10^{-3}
years	-4.16×10^{-5}	colombia	2.54×10^{-3}
support	-4.01×10^{-5}	accounting	2.51×10^{-3}
collaboration	-3.91×10^{-5}	administrative	2.46×10^{-3}
web	-3.78×10^{-5}	peru	2.44×10^{-3}
consultant	-3.50×10^{-5}	assistant	2.38×10^{-3}
mails	-3.34×10^{-5}	vienna	2.37×10^{-3}
specialist	-3.30×10^{-5}	food	2.37×10^{-3}
foundation	-3.23×10^{-5}	operations	2.35×10^{-3}
phones	-3.07×10^{-5}	madrid	2.23×10^{-3}
online	-2.89×10^{-5}	country	2.20×10^{-3}
present	-2.82×10^{-5}	van	2.11×10^{-3}
catalan	-2.82×10^{-5}	commercial	2.03×10^{-3}
software	-2.74×10^{-5}	romania	2.03×10^{-3}
qualification	-2.61×10^{-5}	buenos	2.01×10^{-3}
fp	-2.52×10^{-5}	city	1.97×10^{-3}
economics	-2.49×10^{-5}	canada	1.95×10^{-3}

Table 22: Top 30 most influential words by shift score for each class in the **Religion** classification task.

A.4 Sexual Orientation

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
spain	-4.25e-4	city	5.70e-3
training	-2.12e-4	research	4.18e-3
manager	-2.04e-4	sweden	4.03e-3
es	-1.88e-4	school	3.75e-3
sales	-1.68e-4	córdoba	3.48e-3
business	-1.67e-4	madrid	3.14e-3
administrative	-1.21e-4	english	3.06e-3
sa	-9.46e-5	country	2.72e-3
commercial	-9.39e-5	germany	2.64e-3
international	-9.21e-5	science	2.61e-3
technical	-7.38e-5	customer	2.58e-3
marketing	-6.95e-5	prague	2.47e-3
operator	-6.69e-5	social	2.43e-3
work	-6.50e-5	poland	2.31e-3
new	-6.39e-5	havana	2.29e-3
diploma	-5.45e-5	town	2.24e-3
construction	-5.03e-5	brno	2.18e-3
information	-5.00e-5	wide	2.10e-3
financial	-4.99e-5	uk	2.08e-3
consulting	-4.54e-5	london	2.01e-3
intelligence	-4.49e-5	fashion	1.98e-3
food	-4.48e-5	community	1.89e-3
years	-4.39e-5	lausanne	1.88e-3
valencia	-4.31e-5	zurich	1.85e-3
hours	-4.25e-5	corporate	1.84e-3
health	-4.12e-5	media	1.84e-3
maintenance	-3.77e-5	law	1.83e-3
real	-3.76e-5	durham	1.81e-3
knowledge	-3.64e-5	rome	1.81e-3
estate	-2.91e-5	utrecht	1.76e-3

Table 23: Top 30 most influential words by shift score for each class in the **Sexual Orientation** classification task.

Appendix B

Additional Lexical Word Group Tables

This appendix contains supplementary tables with the top lexical contributors for each word group not highlighted in the main results. These tables follow the same format and methodology described in Chapter 4 but include groups considered less central to the core findings. The goal is to provide transparency and completeness for future reference or replication.

B.1 Perceived Origin

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
carrefour	-1.63e-4	university	6.74e-2
aux	-1.09e-4	microsoft	3.45e-2
inglés	-1.06e-4	van	2.51e-2
pablo	-1.01e-4	universidad	1.61e-2
pedro	-8.62e-5	excel	1.50e-2
merlin	-8.58e-5	gestión	1.44e-2
seguridad	-8.58e-5	new	1.22e-2
carmen	-5.75e-5	crm	1.01e-2
securitas	-5.75e-5	paul	6.43e-3
uned	-5.75e-5	coursera	6.16e-3

Table 24: Top 10 most influential **Propernoun** words by shift score — **Perceived Origin**.

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
course	-3.35×10^{-4}	bachelor	6.45×10^{-3}
manager	-2.91×10^{-4}	administrator	5.66×10^{-3}
degree	-2.36×10^{-4}	cancer	5.47×10^{-3}
administrative	-1.87×10^{-4}	surgery	5.34×10^{-3}
computer	-1.15×10^{-4}	process	4.77×10^{-3}
school	-8.58×10^{-5}	patients	4.27×10^{-3}
information	-6.05×10^{-5}	nurse	4.27×10^{-3}
specialist	-5.35×10^{-5}	programme	4.13×10^{-3}
coordinator	-3.72×10^{-5}	customer	4.00×10^{-3}
education	-3.40×10^{-5}	construction	3.91×10^{-3}

Table 25: Top 10 most influential **Occupation** words by shift score — **Perceived Origin**.

Class 0: Local		Class 1: Foreign	
Word	Shift Score	Word	Shift Score
social	-3.09e-4	development	1.56e-2
digital	-1.80e-4	planning	1.02e-2
german	-1.63e-4	intercultural	8.65e-3
teamwork	-1.13e-4	internal	7.14e-3
intelligence	-9.65e-5	agility	6.53e-3
networks	-8.54e-5	explanations	5.08e-3
senior	-7.54e-5	enthusiasm	4.68e-3
windows	-7.54e-5	wants	4.58e-3
team	-6.13e-5	passionate	4.51e-3
collaboration	-6.06e-5	problem	4.42e-3

Table 26: Top 10 most influential **Skill** words by shift score — **Perceived Origin**.

B.2 Gender

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
seville	-2.12e-4	city	2.13e-2
usa	-1.54e-4	country	1.65e-2
states	-1.49e-4	lima	1.64e-2
murcia	-7.65e-5	melilla	1.58e-2
alicante	-7.52e-5	zurich	1.40e-2
capital	-6.21e-5	canada	1.37e-2
paris	-5.68e-5	zaragoza	1.34e-2
chile	-5.46e-5	paraguay	1.33e-2
sierra	-5.46e-5	romania	1.31e-2
puerto	-4.26e-5	toledo	1.14e-2

Table 27: Top 10 location-related words by shift score in the **Gender** classification task.

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
universitat	-2.87e-4	university	8.72e-2
jorge	-1.25e-4	microsoft	4.57e-2
studio	-1.21e-4	van	2.21e-2
instituto	-1.15e-4	excel	1.63e-2
juan	-9.72e-5	word	1.23e-2
mvs	-8.19e-5	gmbh	1.17e-2
simón	-8.19e-5	adobe	9.13e-3
bolívar	-8.19e-5	google	6.73e-3
carlos	-7.66e-5	crm	6.45e-3
dept	-6.83e-5	aux	5.92e-3

Table 28: Top 10 proper noun words by shift score in the **Gender** classification task.

Class 0: Man		Class 1: Woman	
Word	Shift Score	Word	Shift Score
development	-3.06e-4	social	2.51e-2
analysis	-1.68e-4	management	1.66e-2
level	-1.43e-4	communication	1.03e-2
knowledge	-9.45e-5	care	8.43e-3
powerpoint	-6.30e-5	training	7.44e-3
advanced	-5.78e-5	communications	7.31e-3
political	-5.55e-5	quality	6.22e-3
ability	-4.06e-5	german	5.62e-3
urls	-3.90e-5	linux	5.31e-3
personal	-3.87e-5	work	4.86e-3

Table 29: Top 10 skill-related words by shift score in the **Gender** classification task.

B.3 Religion

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
valencia	-3.48×10^{-4}	spain	2.60×10^{-2}
online	-1.86×10^{-4}	venezuela	1.76×10^{-2}
london	-1.05×10^{-4}	berlin	1.69×10^{-2}
kingdom	-9.13×10^{-5}	lima	1.64×10^{-2}
las	-9.07×10^{-5}	alicante	1.48×10^{-2}
european	-8.71×10^{-5}	colombia	1.34×10^{-2}
state	-7.35×10^{-5}	peru	1.28×10^{-2}
sant	-7.13×10^{-5}	vienna	1.23×10^{-2}
nijmegen	-4.74×10^{-5}	madrid	1.18×10^{-2}
estate	-4.55×10^{-5}	country	1.14×10^{-2}

Table 30: Top 10 location-related words by shift score for each class in the **Religion** classification task.

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
radboud	-3.29×10^{-4}	microsoft	6.10×10^{-2}
studio	-1.66×10^{-4}	university	3.64×10^{-2}
province	-1.50×10^{-4}	excel	1.89×10^{-2}
ibm	-9.77×10^{-5}	word	1.55×10^{-2}
dept	-9.77×10^{-5}	new	1.55×10^{-2}
señora	-7.12×10^{-5}	crm	9.89×10^{-3}
konecta	-7.12×10^{-5}	coursera	8.88×10^{-3}
nuestra	-7.12×10^{-5}	gmbh	8.11×10^{-3}
centre	-5.32×10^{-5}	antonio	7.10×10^{-3}
trello	-5.32×10^{-5}	juan	6.88×10^{-3}

Table 31: Top 10 proper noun words by shift score for each class in the **Religion** classification task.

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
data	-1.63×10^{-4}	course	1.61×10^{-2}
degree	-1.46×10^{-4}	certificate	7.30×10^{-3}
education	-1.35×10^{-4}	philosophy	6.22×10^{-3}
master	-1.25×10^{-4}	marketing	6.09×10^{-3}
courses	-8.14×10^{-5}	accounting	5.58×10^{-3}
sciences	-7.54×10^{-5}	food	5.57×10^{-3}
consulting	-7.13×10^{-5}	pharmacy	5.28×10^{-3}
design	-4.49×10^{-5}	operations	5.17×10^{-3}
risk	-3.80×10^{-5}	sales	5.05×10^{-3}
trade	-3.20×10^{-5}	business	4.89×10^{-3}

Table 32: Top 10 occupation-related words by shift score for each class in the **Religion** classification task.

Class 0: Secular		Class 1: Religious	
Word	Shift Score	Word	Shift Score
social	-2.71×10^{-4}	management	2.04×10^{-2}
quality	-1.84×10^{-4}	handling	1.04×10^{-2}
support	-1.23×10^{-4}	primary	9.44×10^{-3}
group	-9.97×10^{-5}	technical	8.06×10^{-3}
collaboration	-8.67×10^{-5}	higher	7.56×10^{-3}
real	-7.66×10^{-5}	control	5.98×10^{-3}
relations	-6.80×10^{-5}	explanations	5.97×10^{-3}
italian	-5.87×10^{-5}	cleaning	5.77×10^{-3}
team	-5.79×10^{-5}	wide	4.85×10^{-3}
vocational	-3.88×10^{-5}	programming	4.31×10^{-3}

Table 33: Top 10 skill-related words by shift score for each class in the **Religion** classification task.

B.4 Sexual Orientation

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
spain	-3.27e-3	city	2.54e-2
international	-5.86e-4	sweden	2.03e-2
valencia	-2.81e-4	córdoba	1.75e-2
usa	-1.94e-4	madrid	1.46e-2
estate	-1.88e-4	germany	1.28e-2
online	-1.69e-4	prague	1.25e-2
global	-1.56e-4	poland	1.17e-2
york	-1.52e-4	havana	1.15e-2
barcelona	-1.27e-4	country	1.12e-2
murcia	-1.07e-4	brno	1.10e-2

Table 34: Top 10 location-related words by shift score for each class in the **Sexual Orientation** classification task.

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
gestión	-3.12e-4	university	5.98e-2
radboud	-2.93e-4	microsoft	3.88e-2
juan	-2.06e-4	adobe	2.46e-2
banco	-1.91e-4	excel	1.10e-2
francisco	-1.65e-4	word	9.55e-3
pablo	-6.81e-5	gil	8.30e-3
carlos	-6.81e-5	universiteit	8.30e-3
jorge	-3.25e-5	illustrator	7.61e-3
mar	-2.54e-5	jgu	6.53e-3
ceip	8.63e-6	upc	6.53e-3

Table 35: Top 10 propernoun-related words by shift score for each class in the **Sexual Orientation** classification task.

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
business	-4.44e-4	research	1.01e-2
administrative	-3.54e-4	school	8.98e-3
commercial	-2.10e-4	science	6.06e-3
information	-1.05e-4	town	6.01e-3
office	-9.73e-5	customer	5.82e-3
financial	-9.44e-5	fashion	5.35e-3
sales	-8.83e-5	corporate	4.76e-3
operator	-6.83e-5	media	4.49e-3
construction	-6.71e-5	law	4.48e-3
health	-6.41e-5	laboratory	4.03e-3

Table 36: Top 10 occupation-related words by shift score for each class in the **Sexual Orientation** classification task.

Class 0: Heterosexual		Class 1: LGBT	
Word	Shift Score	Word	Shift Score
management	-1.99e-3	wide	1.09e-2
training	-7.96e-4	english	9.35e-3
programming	-3.55e-4	social	8.39e-3
work	-2.32e-4	analysis	6.63e-3
digital	-2.30e-4	german	6.54e-3
writing	-9.00e-5	content	6.28e-3
ability	-8.20e-5	algorithms	5.49e-3
italian	-7.73e-5	complex	5.12e-3
cleaning	-7.66e-5	lean	5.11e-3
strategic	-6.24e-5	a2	4.77e-3

Table 37: Top 10 skill-related words by shift score for each class in the **Sexual Orientation** classification task.