Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 159-178 recibido 05-12-2024 revisado 21-01-2025 aceptado 08-02-2025

On the Impact of Syntactic Infusion for Gender Categorization Across Contextual Dimensions

Sobre el impacto de la integración sintáctica en la categorización de género a través de dimensiones contextuales

Inés Veiga Menéndez, Alberto Muñoz-Ortiz, David Vilares
Universidade da Coruña, CITIC
Departamento de Ciencias de la Computación y Tecnologías de la Información
Campus de Elviña s/n, 15071, A Coruña, Spain
{i.veiga1, alberto.munoz.ortiz, david.vilares}@udc.es

Abstract: This paper investigates how incorporating syntactic information can enhance the categorization of text into multiple gender dimensions, defined by our own identity (*as* category), the person we are addressing (*to* category), or the individual we are discussing (*about* category). Specifically, we explore the use of dependency grammars to integrate explicit syntactic embeddings while leveraging the strengths of pre-trained masked language models (MLMs). Our goal is to determine if dependency grammars add value beyond the implicit syntactic understanding already captured by MLMs. We begin by establishing a baseline using standard MLMs. Next, we propose a neural architecture that explicitly integrates dependency-based structures into this baseline, enabling a comparative analysis of performance and variations. Finally, in addition to evaluating the results, we analyzed the training dynamics of the two proposed variants to provide additional insights into their behavior during the fine-tuning stage. Explicit syntactic information boosts performance in single-task setups, though its gains fade in multitask scenarios.
Keywords: Gender Classification, Dependency Grammars, Training Dynamics.

Resumen: Este artículo investiga cómo incorporar información sintáctica puede mejorar la clasificación de textos en múltiples dimensiones de género, definidas por nuestra propia identidad (categoría *as*), la persona a la que nos dirigimos (categoría *to*) o el individuo del que se habla (categoría *about*). En concreto, exploramos el uso de gramáticas de dependencias para integrar representaciones sintácticas explícitas, complementando las representaciones de modelos de lenguaje enmascarados preentrenados (MLMs). Nuestro objetivo es determinar si las gramáticas de dependencias aportan algo más allá de la comprensión sintáctica implícita ya capturada por los MLMs. Para ello, primero establecemos un modelo base usando un MLM estándar. A continuación, proponemos una arquitectura neuronal que integra en este modelo estructuras basadas en dependencias de forma explícita, permitiendo comparar del rendimiento y las variaciones. Finalmente, evaluamos los resultados y analizamos las dinámicas las dinámicas de entrenamiento de las dos variantes propuestas para ofrecer información adicional sobre su comportamiento durante la etapa de ajuste fino. La información sintáctica explícita mejora el rendimiento en configuraciones de tarea única, aunque sus beneficios disminuyen en escenarios multitarea.
Palabras clave: Clasificación de Género, Gramáticas de Dependencias, Dinámicas de Entrenamiento.

1 Introduction

Languages inherently manifest biases across various dimensions, often shaped by societal norms and cultural contexts (Bolukbasi et al., 2016; Sap et al., 2020). These biases might be reflected in our choice of words, expressions, or even sentence structures, revealing underlying attitudes and assumptions. This often becomes evident in how we refer to individuals, as factors such as age, professional roles, hierarchical status, or other attributes that can shape our choice of language. Gender is

ISSN 1135-5948 DOI 10.26342/2025-74-11

no stranger to these biases (Sun et al., 2019; Vashishtha, Ahuja, and Sitaram, 2023) and has, in fact, been one of the most widely studied dimensions across text (Gonen and Goldberg, 2019; Kaneko et al., 2022; Garrido-Muñoz, Montejo-Ráez, and Santiago, 2022), language and vision (Ross, Katz, and Barbu, 2021; Harrison, Gualdoni, and Boleda, 2023; Fraser and Kiritchenko, 2024), and computer vision domains (Wang et al., 2019; Schwemmer et al., 2020; Wang et al., 2024).

In the case of text-based models, Dinan et al. (2020) proposed breaking down gender classification—from a contextual point of view—into three dimensions: the *as* category, which represents the identity of the person expressing the message; the *to* category, referring to the person being addressed; and the *about* category, which concerns the individual being discussed. Addressing these dimensions in natural language processing (NLP) typically involves a classification approach; and classification models based on fine-tuning masked language models (MLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have demonstrated strong performance across a range of classification tasks. Among these, multidimensional gender classification—predicting the *as*, *to*, and *about* categories—is the focus of this work. However, Dinan et al. (2020) also noted convergence issues in some cases, suggesting opportunities to explore alternative strategies.

The *as*, *to*, and *about* categories convey syntactic meaning. It is known that MLMs capture lexical, syntactic, and semantic structures of language within their latent representations during pre-training, even when the input lacks explicit structure (Hewitt and Manning, 2019; Muñoz-Ortiz, Vilares, and Gómez-Rodríguez, 2023; Waldis et al., 2024). On the other hand, syntactic parsers achieve excellent results in extracting syntactic trees, especially for English texts (Berzak et al., 2016). This context raises the following question: Can MLM-based classification models effectively handle the syntactic nuances of tasks involving these *as*, *to*, and *about* categories, or could their performance be improved by explicitly adding syntactic information to the input data? To explore this hypothesis, we propose leveraging dependency parsing (Nivre, 2010)—a syntactic framework that represents sentence structure as a graph of binary relationships between words, with each relationship defining the syntactic role of a word within the sentence. By integrating dependency-based representations, we aim to evaluate if infusing syntax can provide additional insights or improve the model's ability to disentangle the *as*, *to*, and *about* categories more effectively than relying solely on contextualized word representations.

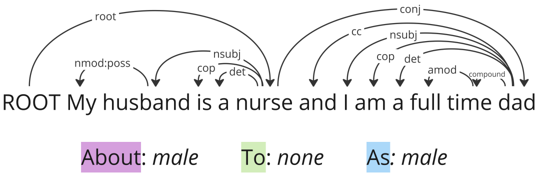


Figure 1: Dependency tree illustrating the syntactic relationships and their connection to the *as*, *to*, and *about* gender categories.

Figure 1 illustrates a dependency tree, highlighting the roles of the *about* and *as* dimensions. The word 'nurse' is unambiguously identified as referring to a man due to its syntactic link to the word 'husband', whose referent in the *about* dimension is male. Hypothetically, while predictive models might incorrectly associate 'nurse' with a woman due to statistical bias, the syntactic dependency could provide a more precise determination of gender. Similarly, 'dad', when directly to the pronoun 'I', could hypothetically allow the speaker's gender to be inferred within the *as* dimension.

Contribution We investigate the integration of syntactic information into text classification tasks across multiple dimensions: the *as*, *to*, and *about* categories. By leveraging dependency grammars, which capture word relationships (e.g., identifying direct objects or subjects), we enhance the model's ability to interpret linguistic structure. Our approach emphasizes cost-efficient, task-specific models, leveraging pre-trained models like BERT for effective and practical classification within resource constraints. First, we train a multi-task learning (MTL) sequence classification model using a language encoder and a hard-sharing decoding architecture to predict three tasks, one for each category. Next, we design an adapter to integrate syntactic information, added on top of the MLM representations. Finally, we apply training dynamic strategies to investigate whether the inclusion of syntax not only impacts the final results but also influences model convergence and reveals patterns in the training data.

2 Related Work

Next, we provide an overview of related work, focusing on gender classification in NLP and how syntactic features have been used to enhance representations.

2.1 Gender Classification in NLP

Challenges arise at every stage of an NLP system, from data collection to downstream applications, as systemic patterns in the data can propagate and amplify throughout the pipeline. The specific task of gender classification has been similarly affected by this issue, leading to studies summarizing efforts to address disparities and imbalances in NLP systems (Sun et al., 2019; Stanczak and Augenstein, 2021; Bartl et al., 2024).

First, the use of datasets scraped from public sources makes it challenging to control if models are trained on biased data. For instance, gender differences have been observed depending on the gender of the author of a text (Garimella et al., 2019; Newman et al., 2008) or the person being discussed (Marjanovic, Stańczak, and Augenstein, 2022; Asr et al., 2021). This initial lack of control over dataset cretion contributes to the propagation of these differences in the next steps.

Word embeddings and pre-trained language models (PLMs), both masked and autoregressive, are particularly susceptible to this issue due to the vast amount of data they are trained on. Word embeddings reflect and amplify gender biases present in the training data (Sun et al., 2019; Bolukbasi et al., 2016), while PLMs have been shown to reproduce various societal biases, including gender bias (Kurita et al., 2019; May et al., 2019; Nangia et al., 2020; Nadeem, Bethke, and Reddy, 2021; Thakur, 2023), which leak into downstream tasks (Stanovsky, Smith, and Zettlemoyer, 2019; Tal, Magar, and Schwartz, 2022; Sheng et al., 2019). Increasing model size, while often improving performance, can exacerbate these biases (Tal, Magar, and Schwartz, 2022; Bender et al., 2021).

Accurately classifying gender is crucial for tasks like author profiling, which involves identifying an author's profile—including gender—based on writing style, but this can be undermined by stereotypical biases (Chen, Roth, and Falenska, 2024). Gender classification is particularly challenging in non-normative contexts. While LLMs demonstrate high accuracy in predicting male and female names, their performance significantly drops for gender-neutral names (You et al., 2024). Similarly, NLP models face difficulties in handling same-gender relationships (Sobhani and Delany, 2024).

Due to the complex nature of biases, mitigating them remains a challenging task (Gonen and Goldberg, 2019). Various mitigation approaches have been developed (Bartl, Nissim, and Gatt, 2020; Garimella et al., 2021), from specialized training data (Webster et al., 2018) and debiasing techniques (Hall Maudslay et al., 2019) to targeted interventions like balanced demographic representation (Sheng et al., 2020; Ghanbarzadeh et al., 2023). However, these methods face particular challenges when applied to languages with grammatical gender (Bartl, Nissim, and Gatt, 2020). Adding context, such as previous sentence or speaker information, can reduce bias in machine translation (Basta, Costa-jussà, and Fonollosa, 2020).

2.2 Enriching Representations with Syntactic Features

Syntactic parsing has traditionally been seen as an important aspect of NLP, contributing to more advanced language understanding in tasks such as sentiment analysis (Barnes et al., 2021; Tian, Chen, and Song, 2021), machine translation (Han et al., 2013; Bugliarello and Okazaki, 2020), and question answering (Perera and Nand, 2016; Reddy et al., 2016), *inter alia.* However, the success of pretrained models with substantial expressive power has started to challenge this view (Glavaš and Vulić, 2021).

Even so, incorporating explicit syntactic information has been shown to improve MLMs. Bai et al. (2021) proposed a framework that integrates syntactic information during MLM pretraining, leading to better results across various downstream tasks. Zheng, Fan, and Li (2024) improved cross-lingual transfer by integrating both lexical and syntactic information into multilingual BERT, while Iwamoto et al. (2023) leveraged syntactic information to mitigate catastrophic forgetting in cased BERT models. Similarly, it has been shown that these benefits can also extend to fine-tuned MLMs in applications such as sentence matching (Liu et al., 2020), machine translation (Zhang et al., 2019), and sentiment classification (Cho, Jung, and Hockenmaier, 2023).

3 The Problem

This work focuses on multidimensional gender classification based on Dinan et al. (2020), which defines three dimensions of gender across contextual dimensions:

- *About*: The gender of the person being discussed. In the sentence *'My sister is a lawyer'*, the *about* dimension would be labeled as female.
- *To*: The gender of the person being addressed. In the sentence *'Nice to meet you, Belinda'*, the *to* dimension would be labeled as female.
- *As*: The gender of the speaker. In the sentence *'My husband is a nurse, and I am a full-time dad'*, the *as* dimension would be labeled as male.

To evaluate this problem, Dinan et al. (2020) introduced the **MD Gender** dataset, designed to evaluate classification models trained to predict one or more of these dimensions. It is composed of 2 345 dialogs in North American English, which are manually annotated in one of more of the mentioned categories. Annotations are included only for the relevant dimensions (*about*, *to*, or *as*), while the others are marked with a placeholder value (*nil*). The labels in the MD Gender dataset are: ABOUT:male, ABOUT:female, PARTNER:male, PARTNER:female, SELF:male, and SELF:female, where PARTNER corresponds to the *to* category and SELF corresponds to the *as* category. Annotations in each category are split roughly in half between male and female.

To train models for this task, the authors proposed or referenced several datasets with annotations that could be mapped to the *about*, *to*, and *as* dimensions. In this work, we rely on the Conv AI dataset to train our models. Its training set comprises 17 878 English multi-turn conversations, which were automatically annotated in the three dimensions of gender by the authors of MD Gender. While we attempted to use other datasets, they were either not released, no longer maintained, or we did not receive a response from the authors despite multiple contact attempts. The Conv AI dataset originally included the labels *male*, *female*, *gender-neutral*[1], and *unknown*, the latter used for cases where the label is indeterminate. The training samples are annotated for only one of the three contextual gender dimensions.

4 Models for multidimensional gender classification

Next, we present: (1) a brief overview of the baseline models used as a starting point for tackling the multidimensional text classification problem, and (2) our methodology for integrating syntactic elements to assess their influence.

4.1 Baseline models

We use two common approaches in NLP: single- and multi-task learning (MTL):

- Single-task: A separate model is trained for each classification task: one for the *about* dimension and two others independently for the *as* and *to* dimensions.
- Multi-task: A single unified model with a MTL framework. It uses a shared encoder to process the input, followed by three decoders, each responsible for predicting one of the dimensions.

Formally, for **single-task models**, the input to these models is a sequence of words, $\mathbf{X} = [w_1, w_2, \ldots, w_n]$, where w_i represents the i-th word in the text. The output is a categorical label $y \in \{1, 2, \ldots, K\}$, where K is the number of classes for the specific dimension. For sequence-to-label models, we require two main components. First, a text encoder, Enc($\mathbf{X}$), which transforms the input sequence into a contextual representation $\mathbf{H}$. It could be a bidirectional LSTM or a pre-trained transformer-based language model (e.g., BERT, RoBERTa, as we will be using in this work). Second, a classifier head, Dec($\mathbf{H}$), which maps the encoded representation $\mathbf{H}$ to the output label probabilities

[1] We fully recognize and respect that gender is a non-binary and multidimensional concept. However, the decision to exclude the *gender-neutral* label in this work was not ours but a consequence of the MD Gender dataset lacking an equivalent representation for this category. We acknowledge this as a limitation of our approach and a shortcoming of the evaluation framework.

through a fully connected layer and a softmax activation. The model is trained by minimizing the cross-entropy loss, expressed as $\mathcal{L}_{\text{single}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{y}_i^{(k)}\log\hat{\mathbf{y}}_i^{(k)}$, where $\mathbf{y}_i^{(k)}$ is the true label distribution, and $\hat{\mathbf{y}}_i^{(k)}$ is the predicted probability for class k.

For **multi-task models**, the input $X = [w_1, w_2, \ldots, w_n]$ is the same, but shared across tasks, and the output is a set of predictions $\{y_{\text{about}}, y_{\text{as}}, y_{\text{to}}\}$, one for each dimension. The MTL model consists of a shared encoder, $\text{Enc}(\mathbf{X})$, which generates a common representation $\mathbf{H}$ for the input text. This shared representation is then passed to task-specific decoder heads, $\text{Dec}_{\text{about}}, \text{Dec}_{\text{as}}, \text{Dec}_{\text{to}}$, each of which produces a prediction for its respective dimension. We train the models by jointly minimizing a combined loss, expressed as $\mathcal{L}_{\text{multi}} = \mathcal{L}_{\text{about}} + \mathcal{L}_{\text{as}} + \mathcal{L}_{\text{to}}$, where each component loss $\mathcal{L}_{\text{dimension}}$ represents the cross-entropy loss for the corresponding task.

4.2 Syntax-infused models

Next, we describe our approach for integrating syntactic embeddings and contextualized word vectors and review key concepts and notation related to dependency parsing.

4.2.1 Basics and Notation

Let $G = (V, E)$ represent a dependency graph, where V is the set of nodes and $E \subseteq V \times D \times V$ is the set of edges. Here, D denotes the set of possible dependency types. A dependency graph is formally defined as a collection of triplets of the form (v, d, w), where $v, w \in V$ and $d \in D$. Each triplet (v, d, w) represents a directed edge from node v to node w, with the edge labeled by the dependency type d. Here, we consider v and w as integers referring to the input words indexed at the v-th and w-th positions of the input sentence[2]. In this work, we focus on dependency trees, which are a specialized form of dependency graphs with additional restrictions. Specifically, a dependency tree has a single-head constraint, meaning each node (except the root, which serves as the origin of all dependencies) has exactly one incoming edge. Additionally, the structure must be acyclic. For example, in Figure 1, the word 'a' is dependent on the word 'nurse' and is connected by a dependency labeled as *det*.

One of the main challenges in combining syntactic structures and word sequences is that generic encoders, like MLMs, process inputs as flat sequences of words, while syntactic information is represented as a tree. This mismatch makes it difficult to effectively combine sequence-level and tree-structured information. As a result, models that try to incorporate both often end up with rigid and overly specialized architectures.

Traditionally, this has been addressed using classic parsing approaches, such as transition- or graph-based methods (Nivre, 2010). More recently, an alternative approach has been developed that represents dependency trees as sequences of labels, assigning each word a label that encodes part of the tree (Li et al., 2018; Strzyz, Vilares, and Gómez-Rodríguez, 2019; Amini, Liu, and Cotterell, 2023). This strategy was developed to make tree computation more straightforward with standard tagging models, but it also provides a way to represent trees as flat sequences, making them easier to integrate with word embeddings. Next, we describe the key concepts of parsing as sequence labeling and the encodings proposed in (Strzyz, Vilares, and Gómez-Rodríguez, 2019; Strzyz, Vilares, and Gómez-Rodríguez, 2020) that we will be using in this work.

4.2.2 Parsing as Sequence Labeling

Sequence labeling involves learning a one-to-one mapping between elements of an input sequence (e.g., words, characters, or subword units) and a set of labels. Each input element is assigned one label, enabling structured predictions over the sequence. This approach has proven useful in tasks like parsing, offering a faster and more efficient way to compute syntactic or semantic structures compared to traditional methods. However, we will use these linearizations with a different objective: incorporating syntactic information into multidimensional contextual classification.[3]

[2] The first actual word of the sentence is considered to be indexed at 1. Additionally, it is common to add a dummy word, indexed at position 0, which points to the actual root of the sentence.

[3] Parsing linearizations have previously been explored to enhance other tasks. For example, constituent parsing linearizations have been applied to improve semantic role labeling (Johansson and Nugues, 2008), dependency parsing has been utilized as part of a pre-training phase for low-resource languages (Rotman and Reichart, 2019), and even as components in pipeline systems for tasks such as sentiment analysis (Imran, Kellert, and Gómez-Rodríguez, 2024).

Building a linearized tree involves assigning each word w_i a label (x_i, l_i), where l_i represents the dependency type, and x_i encodes a subset of the tree's arcs. We next describe how x_i is encoded, as l_i consistently denotes the dependency type. Figure 2 shows examples for the encodings used, explained below.

Absolute Positional Encoding In dependency trees, each dependent node has one head. Each edge is represented as $(w_i, d_i, w_j) \in E$, where $w_j \in V$ is the head, $w_i \in V$ is the dependent, and $d_i \in D$ is the dependency type between w_i and w_j. Due to the single-head restriction, a naive encoding is to assign each w_i a label of the from j representing the absolute position of its head word in the sentence. As an example, consider Figure 2, where the label $(3, \mathrm{obj})$ assigned to the word '*you*' signifies that its syntactic head, '*meet*' , is positioned at index 3 in the sentence with the dependency relation *obj*[4].

Relative Positional Encoding A variation of absolute positional encoding, where, instead of storing the absolute position of the head, each word is assigned an offset of the form $j - i$ to encode the component x_i of the syntactic label. The main motivation is to achieve a more compact and standardized representation. For instance, there are frequent left dependencies where the head of a word is the immediately preceding term (i.e., $j - i = -1$), or more generally, where the head is k positions to the left or right. These offsets reflect recurring syntactic patterns. In absolute positional encoding, the associated labels depend on the position of w_i in the sequence, leading to variability. In contrast, relative encoding assigns the same label to equivalent dependency arcs, regardless of their position in the sequence. This standardization simplifies the representation, and may emphasize underlying syntactic patterns. Theoretically, this standardization enhances the encoding's learnability, which was the main reason for its introduction.

Part-of-speech-based encoding It integrates both grammatical categories and relative positioning within the dependency tree to define offsets between dependents and heads. For a pair of words (w_i, w_j) it assigns the dependent word w_i a label x_j, represented as a tuple (p_i, o_j). Here, p_i denotes the PoS tag of the head word w_i, and o_j specifies the number of words to the right or left that share the same PoS tag as the head. Positive values of o_j represent words to the right, while negative values indicate words to the left. For example, in Figure 2, the dependency arc links *'you'* (w_i) and *'meet'* (w_j). Here, *'meet'* is the first verb to the left of *'you'*, so the label for *'you'* is $(\mathrm{VERB}, -1, \mathrm{obj})$, i.e. *'you'* is related to the nearest preceding verb through an object dependency. This representation explicitly encodes simple syntactic patterns within the label itself. Still, it requires precomputed or dynamically generated labels for real data, which we obtain by running the Stanza PoS tagger (Qi et al., 2020).

Bracketing encoding Each word w_j is assigned a label in the form (x_j, l_j), where x_j is a text string encoding the incoming and outgoing arcs of w_j and its neighbors, and l_j represents the dependency relation label connecting w_j to its parent node. The string x_j adheres to the following regular expression: `(<)?((\)*|(/)*)(>)?`, where `<` indicates that w_{j-1} has an incoming arc from the right; a repetition of \ k times means that w_j has k outgoing arcs to the left; a repetition of `/` k times signifies that w_{j-1} has k outgoing arcs to the right; and `>` indicates that w_j has an incoming arc from the left. This representation captures the structural relationships between a word, its neighbors, and the dependency arcs within the tree.

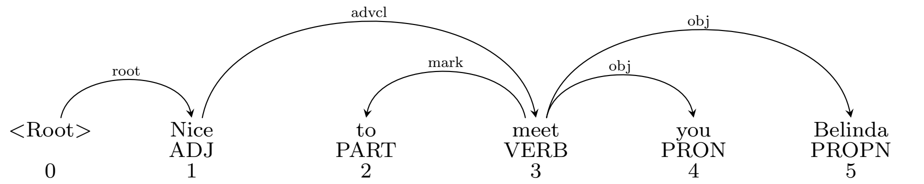
We use Stanza's dependency parser to compute trees that are later linearized using these encodings.

5 *Training dynamics*

When working with classification models, larger datasets are generally preferred due to their potential for improving overall performance. However, their size can complicate the quality analysis and obscure insights into training dynamics. In response to this, Swayamdipta et al. (2020) introduced a method to analyze the behavior of a classification model through individual sample assessments. They applied this method to a natural language inference task, categorizing sentence pairs as linked, contradictory, or neutral.

Formally, given a dataset $\{(x_i, y_i)\}_{i=1}^{N}$ of size N, where x_i represents the i-th observation and y_i the true label, we define the train-

[4] This labeling approach corresponds to the CoNLL-U format, a standard for encoding dependency trees in plain text.



	1	2	3	4	5
Naive positional encoding	(0,root)	(3,mark)	(1,advcl)	(3,obj)	(3,obj)
Relative positional encoding	(-1,root)	(+1,mark)	(-2,advcl)	(-1,obj)	(-2,obj)
Relative PoS-based encoding	(ROOT,-1,root)	(VERB,+1,mark)	(ADJ,-1,advcl)	(VERB,-1,obj)	(VERB,-1,obj)
Bracketing-based encoding	(/,root)	(<,mark)	(\>//,advcl)	(>,obj)	(>,obj)

Figure 2: Examples of the different encodings used in this work. Positional encodings mark where the head of a token is, while bracketing encodings represent some of the incoming and outgoing arcs of a token.

ing dynamics of the dataset through statistical characterization across E training epochs. Following Swayamdipta et al. (2020), we compute the following statistical parameters:

- Correctness: It measures how often the model correctly predicts the true label y_i throughout the training epochs:

$$\text{Correctness}_i = \frac{1}{E}\sum_{e=1}^{E}\mathbf{1}(\hat{y}_{i,e} = y_i)$$

- Confidence: It measures how certain the model is of having assigned the correct label. The *confidence* value for a sample (x_i, y_i) is defined as the average probability of correctly predicting the label across all training epochs:

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(\hat{y}_i|x_i)$$

- Variability: It measures how consistent the model is in predicting labels across different training cycles. A low variability indicates that the model consistently assigns the same label (whether correct or incorrect), while high variability suggests that the model is more uncertain in its label assignments:

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta^{(e)}}(\hat{y}_i|x_i) - \hat{\mu}_i)^2}{E}}$$

To evaluate these parameters, we capture the probabilities of correct classification for each sample within a tensor and review this data after training to assess confidence and variability. Concurrently, a separate tensor logs the accuracy of each prediction, assisting in the determination of correctness. Upon analyzing these metrics, we create a *data map* that visually organizes training samples according to their variability and confidence. This map groups samples into categories like *easy-to-learn* (with consistent and confident predictions), *hard-to-learn* (marked by low confidence and possibly mislabeled or uninformative data), and *ambiguous* (showing high variability, which helps improve the model's generalization).

6 Experiments

Next, we present the experiments conducted using the MLMs, comparing their performance when trained with input consisting solely of words versus when syntax is explicitly infused into the input. Additionally, we analyze whether there are differences in training dynamics between these two approaches, following the approach described in §5. The code is at `https://github.com/Kuina-sama/syntax-infusion-gender-classification`.

Setup We train both single-task and multi-task learning models using three different MLMs: DistilBERT (Sanh et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). The motivation for using these models is to investigate if performance trends vary based on the model's capacity and strength. For instance, DistilBERT is generally less robust than BERT, which tends to underperform compared to RoBERTa. In the single-task setup, we train an independent model for each task. In the multi-task setup, we train a model that learns all three tasks simultaneously. Hyperparameters are described in Appendix B.

Metrics We use the macro F1-score as the main evaluation metric.[5] It is computed as $F1 = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$, where precision is the proportion of correctly predicted positive cases among all predicted positive cases, and recall is the proportion of correctly predicted positive cases among all actual positive cases.

6.1 Experimental results

Table 1 presents the results of single-task models trained to predict male and female categories within the *about* category.

Model	Infusion	Avg	M	F
BERT	no	84.09	83.72	84.46
	`abs`	85.25	85.20	85.29
	`rel`	**85.59**	85.57	85.60
	`pos`	84.86	84.90	84.82
	`brackets`	84.82	84.80	84.84
DistilBERT	no	83.92	83.52	84.31
	`abs`	83.93	83.85	84.01
	`rel`	83.84	83.69	83.99
	`pos`	84.35	84.19	84.51
	`brackets`	**83.97**	83.87	84.07
RoBERTa	no	86.18	85.97	86.38
	`abs`	85.84	85.63	86.05
	`rel`	86.26	86.13	86.39
	`pos`	85.71	85.48	85.94
	`brackets`	**86.39**	86.26	86.52

Table 1: F1-scores for the *about* dimension, including male and female categories.

This category was the easiest across all gender contextual dimensions. It can be noted that explicitly introducing syntax can lead to improvements compared to baseline models. In particular, the BERT model shows an improvement of approximately 1.5 points in the F1-score when employing relative positional encodings, with similar gains observed for other encodings and consistent improvements for both male and female categories. Meanwhile, the DistilBERT and RoBERTa models exhibit more subtle variations. Notably, the bracketing-based encoding enhances RoBERTa's performance by about 0.2 points. Overall, RoBERTa variants were the best performers, although the differences across various masked language models were generally not substantial.

Table 2 provides a corresponding overview of performance in the *to* dimension. This dimension was the harder for both models with and without explicit syntax integration.

Model	Infusion	Avg	M	F
BERT	no	49.90	42.30	57.49
	`abs`	51.54	43.96	59.12
	`rel`	52.26	45.65	58.86
	`pos`	46.16	32.16	60.16
	`brackets`	**57.15**	54.67	59.62
DistilBERT	no	49.09	37.83	60.35
	`abs`	47.85	34.23	61.47
	`rel`	54.33	49.46	59.19
	`pos`	**56.02**	49.83	62.21
	`brackets`	51.24	42.61	59.86
RoBERTa	no	41.80	17.61	65.99
	`abs`	**60.39**	54.70	66.08
	`rel`	55.59	49.58	61.59
	`pos`	39.35	14.32	64.37
	`brackets`	53.75	44.65	62.85

Table 2: F1-scores for the *to* dimension, including male and female categories.

Model	Infusion	F1	M	F
BERT	no	49.88	36.97	62.78
	`abs`	73.24	72.51	73.96
	`rel`	67.93	68.40	67.46
	`pos`	**74.48**	74.09	74.87
	`brackets`	74.24	72.85	75.63
DistilBERT	no	44.00	24.46	63.53
	`abs`	**67.77**	66.39	69.14
	`rel`	62.71	59.45	65.97
	`pos`	66.85	64.45	69.24
	`brackets`	62.13	59.83	64.42
RoBERTa	no	32.79	0.33	65.24
	`abs`	79.37	80.03	78.71
	`rel`	77.34	77.70	76.98
	`pos`	**79.99**	80.38	79.59
	`brackets`	76.94	77.37	76.51

Table 3: F1-scores for the *as* dimension, including male and female categories.

The positive impact of explicitly incorporating syntactic information into the models is more clear for this dimension. The baseline models struggled classifying samples from the male category. This issue was particularly pronounced in the RoBERTa-based model, echoing similar findings reported in the original release of the MD Gender (Dinan et al., 2020). BERT, whose baseline implementation obtained more balanced results, showed a 7-point improvement in F1-score when using the bracketing-based encoding. Interestingly, the DistilBERT model showed the best results when using the relative PoS-based encoding. For RoBERTa, the inclusion of absolute positional encoding improves performance, boosting the F1-score by a large margin. This suggests that incorporating syntactic information can enhance overall per-

[5] The F1-score provides a balanced measure that is particularly useful for imbalanced datasets.

formance while also addressing specific challenges in classifying certain gender labels—a limitation that has been previously noted in this model. Consequently, the model achieves more balanced performance across male and female categories, reducing bias. However, not all models benefit equally from this approach. Specifically, encodings based on part-of-speech information present challenges. Interestingly, this aligns with findings in dependency parsing, where using predicted POS tags (in this case, generated with Stanza) can empirically lead to lower performance compared to alternative encodings (Muñoz-Ortiz, Strzyz, and Vilares, 2021).

Finally, Table 3 presents the performance for the third contextual dimension, the *as* dimension. The single-task results show that adding explicit syntactic information to inputs improves overall performance and reduces gender biases, particularly improving male classification, where baseline models struggled. Performance boosts vary depending on the encoder but are significant in all cases. Notably, even for RoBERTa, despite the convergence issues observed in the syntax-free baseline,[6] the inclusion of syntactic information resulted in clear gains.

Table 4 presents the results for all contextual models trained as a single multi-task model across the *about*, *to*, and *as* dimensions. In this setup, incorporating syntactic information leads to small but consistent improvements. For BERT, absolute positional encoding improves performance in the *about* and *to* dimensions. Similarly, DistilBERT benefits from relative PoS-based encoding, and RoBERTa shows improved results with relative positional encoding, including a slight gain in the *as* dimension.

Overall, the improvements in this MTL setup are much smaller than those observed in single-task setups. One possible explanation is that the MTL configuration, with its shared parameters and combined loss function, inherently helps the model differentiate between the contextual dimensions (*about*, *to*, and *as*). This shared setup might introduce an inductive bias in the shared encoder, guiding it to learn syntactic dependencies as part of optimizing for the combined loss across the dimensions. This implicit differentiation could reduce the relative impact of explicitly incorporating syntactic information, as the model can already leverage task interdependence to improve performance.

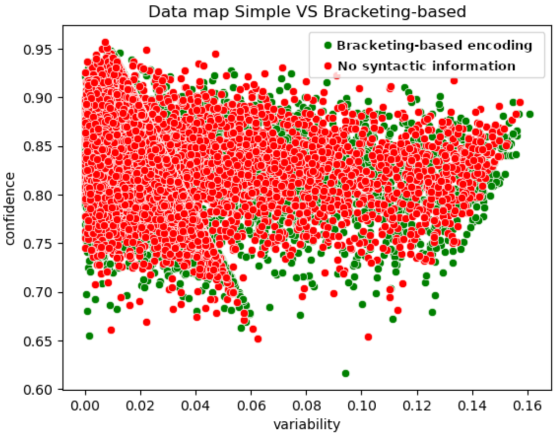


Figure 3: Data Maps: DistilBERT single-task on *about*, baseline vs. syntax-aware.

6.2 Analysis with Training Dynamics

We present additional insights into the task at hand, relying on training dynamics, as introduced in §5. To illustrate these dynamics, Figures 6, 7, and 8 in the Appendix present the dataset cartography generated by the DistilBERT model trained on the *as*, *about*, and *to* dimensions, respectively, in a single-task setup, plotting the confidence, variability, and correctness metrics.[7] For the *as* dimension, most samples exhibit low confidence, low variability, and low correctness values, with no identifiable *easy-to-learn* or *ambiguous* categories. In contrast, for the *about* dimension, most samples exhibit high confidence and low variability, placing them in the *easy-to-learn* category. For the *to* dimension, the *confidence* and *variability* metrics are notably lower than those for *about*, aligning with the low performance observed.

To better understand the effect of infusing syntax, Figures 3, 4, and 5 compare the training dynamics for the *about*, *to*, and *as* dimensions, respectively, highlighting the differences between models trained without and with explicit syntactic infusion. DistilBERT was used for these comparisons, along with the syntactic encoding that performed best.

[6] We made additional efforts to address these convergence issues but were unsuccessful.

[7] Our analysis focuses on single-task models, as multi-task models showed minimal differences between syntactic and non-infused models, with training dynamics yielding no significant results.

Model	Infusion	About			To			As		
		Avg	M	F	Avg	M	F	Avg	M	F
BERT	no	83.06	82.57	83.55	69.97	70.76	69.17	**72.96**	73.37	72.55
	`abs`	**84.10**	83.93	84.26	70.10	71.77	68.43	72.82	73.50	72.14
	`rel`	83.20	83.09	83.31	**70.84**	72.35	69.33	72.29	73.18	71.39
	`pos`	83.42	83.36	83.47	69.99	71.89	68.09	72.55	73.52	71.57
	`brackets`	83.46	83.31	83.60	69.26	71.04	67.47	72.47	73.20	71.73
DistilBERT	no	80.86	80.60	81.12	66.38	67.40	65.36	72.01	71.53	72.48
	`abs`	81.50	81.43	81.56	66.51	68.34	64.67	71.66	71.48	71.83
	`rel`	81.67	81.77	81.57	65.16	67.77	62.55	71.22	71.57	70.87
	`pos`	81.38	81.27	81.49	**66.81**	68.68	64.93	72.06	71.96	72.15
	`brackets`	**81.80**	81.63	81.96	66.17	68.18	64.15	**72.14**	72.11	72.17
RoBERTa	no	83.54	83.21	83.87	72.35	73.39	71.30	72.96	73.42	72.50
	`abs`	**84.78**	84.59	84.96	71.71	72.93	70.48	**73.42**	74.18	72.66
	`rel`	83.93	83.72	84.13	**73.02**	74.13	71.90	73.38	74.13	72.63
	`pos`	84.77	84.51	85.03	71.59	72.67	70.51	72.91	73.64	72.17
	`brackets`	84.47	84.23	84.71	72.13	73.18	71.07	73.24	73.96	72.52

Table 4: Results for contextual models trained as a single multi-task model across the *about*, *to*, and *as* dimensions.

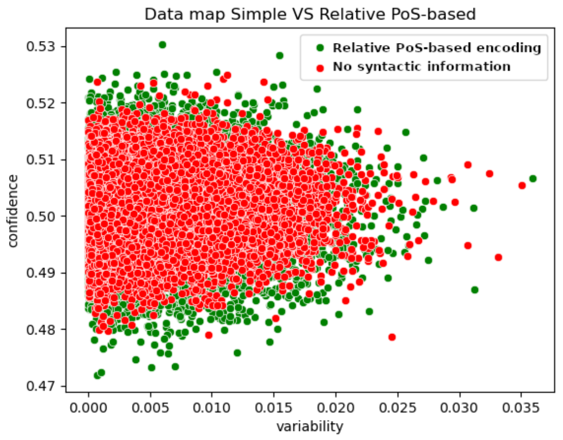


Figure 4: Data Maps: DistilBERT single-task on *to*, baseline vs. syntax-aware.

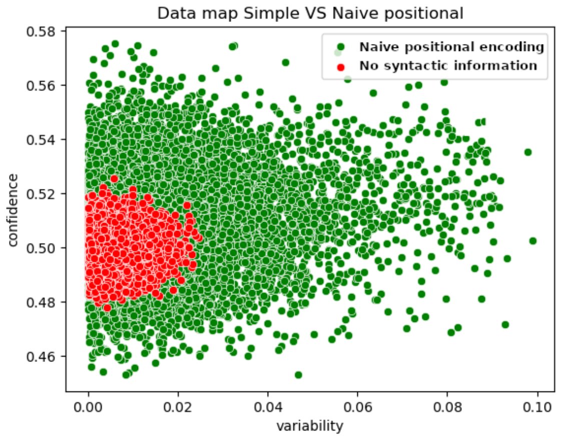


Figure 5: Data Maps: DistilBERT single-task on *about*, baseline vs. syntax-aware.

Due to the large number of data points, we randomly sampled a fraction of the points.

For the *about* dimension, we observed in §6.1 that including syntactic information barely modified performance. This is also reflected in the training dynamics (Figure 3), where including syntactic information only results in a slight improvement in variability. Similarly, changes in performance for the *to* dimension were modest, which is consistent with the small differences shown in training representations in Figure 4. Finally, Figure 5 shows that including syntactic information boosts both higher variability and higher confidence for the *as* dimension, helping the model differentiate the samples. This causes a large boost in F1-score compared to the baseline model, which did not converge.

7 Conclusion

We studied masked language models that implicitly encode syntax versus those augmented with explicit syntactic information for the task of multidimensional gender categorization into *about*, *to*, and *as* dimensions. By incorporating linearized dependency labels, we showed that explicitly enriched models could consistently outperform their counterparts in reducing gender mistakes and improving classification accuracy, particularly in single-task setups. However, these improvements largely diminished in multi-task learning. With respect to future work, it should assess syntactic infusion in multilingual settings and extend gender classification for greater inclusivity.

Acknowledgments

This work was supported by GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and by ERDF, EU; SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; Xunta de Galicia (ED431C 2024/02); Grant PRE2021-097001 funded by MICIU/AEI/10.13039/501100011033 and by ESF+ (predoctoral training grant associated to project PID2020-113230RB-C21); LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF; and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

References

Amini, A., T. Liu, and R. Cotterell. 2023. Hexatagging: Projective dependency parsing as tagging. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1453–1464, Toronto, Canada, July. Association for Computational Linguistics.

Asr, F. T., M. Mazraeh, A. Lopes, V. Gautam, J. Gonzales, P. Rao, and M. Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*, 16(1):e0245533.

Bai, J., Y. Wang, Y. Chen, Y. Yang, J. Bai, J. Yu, and Y. Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online, April. Association for Computational Linguistics.

Barnes, J., R. Kurtz, S. Oepen, L. Øvrelid, and E. Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online, August. Association for Computational Linguistics.

Bartl, M., A. Mandal, S. Leavy, and S. Little. 2024. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*.

Bartl, M., M. Nissim, and A. Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Basta, C., M. R. Costa-jussà, and J. A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA, July. Association for Computational Linguistics.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Berzak, Y., Y. Huang, A. Barbu, A. Korhonen, and B. Katz. 2016. Anchoring and agreement in syntactic annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2224, Austin, Texas, November. Association for Computational Linguistics.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. *Advances in neural information processing systems*, 29.

Bugliarello, E. and N. Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting* *of the Association for Computational Linguistics*, pages 1618–1627, Online, July. Association for Computational Linguistics.

Chen, H., M. Roth, and A. Falenska. 2024. What can go wrong in authorship profiling: Cross-domain analysis of gender and age prediction. In A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, and D. Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 150–166, Bangkok, Thailand, August. Association for Computational Linguistics.

Cho, I., Y. Jung, and J. Hockenmaier. 2023. SIR-ABSC: Incorporating syntax into RoBERTa-based sentiment analysis models with a special aggregator token. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8535–8550, Singapore, December. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dinan, E., A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November. Association for Computational Linguistics.

Fraser, K. and S. Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta, March. Association for Computational Linguistics.

Garimella, A., A. Amarnath, K. Kumar, A. P. Yalla, A. N, N. Chhaya, and B. V. Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online, August. Association for Computational Linguistics.

Garimella, A., C. Banea, D. Hovy, and R. Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy, July. Association for Computational Linguistics.

Garrido-Muñoz, I., A. Montejo-Ráez, and F. M. Santiago. 2022. Exploring gender bias in spanish deep learning models. In *SEPLN (Projects and Demonstrations)*, pages 44–47.

Ghanbarzadeh, S., Y. Huang, H. Palangi, R. Cruz Moreno, and H. Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada, July. Association for Computational Linguistics.

Glavaš, G. and I. Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online, April. Association for Computational Linguistics.

Gonen, H. and Y. Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hall Maudslay, R., H. Gonen, R. Cotterell, and S. Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China, November. Association for Computational Linguistics.

Han, D., P. Martínez-Gómez, Y. Miyao, K. Sudoh, and M. Nagata. 2013. Using unlabeled dependency parsing for prereordering for Chinese-to-Japanese statistical machine translation. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 25–33, Sofia, Bulgaria, August. Association for Computational Linguistics.

Harrison, S., E. Gualdoni, and G. Boleda. 2023. Run like a girl! sport-related gender bias in language and vision. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14093–14103, Toronto, Canada, July. Association for Computational Linguistics.

Hewitt, J. and C. D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Imran, M., O. Kellert, and C. Gómez-Rodríguez. 2024. A syntax-injected approach for faster and more accurate sentiment analysis. *arXiv preprint arXiv:2406.15163*.

Iwamoto, R., I. Yoshida, H. Kanayama, T. Ohko, and M. Muraoka. 2023. Incorporating syntactic knowledge into pretrained language model using optimization for overcoming catastrophic forgetting. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10981–10993, Singapore, December. Association for Computational Linguistics.

Johansson, R. and P. Nugues. 2008. The effect of syntactic representation on semantic role labeling. In D. Scott and H. Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 393–400, Manchester, UK, August. Coling 2008 Organizing Committee.

Kaneko, M., A. Imankulova, D. Bollegala, and N. Okazaki. 2022. Gender bias in masked language models for multiple languages. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States, July. Association for Computational Linguistics.

Kurita, K., N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August. Association for Computational Linguistics.

Li, Z., J. Cai, S. He, and H. Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Liu, T., X. Wang, C. Lv, R. Zhen, and G. Fu. 2020. Sentence matching with syntax- and semantics-aware BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3302–3312, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marjanovic, S., K. Stańczak, and I. Augenstein. 2022. Quantifying gender biases towards politicians on reddit. *PloS one*, 17(10):e0274317.

May, C., A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Muñoz-Ortiz, A., M. Strzyz, and D. Vilares. 2021. Not all linearizations are equally data-hungry in sequence labeling parsing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 978–988, Held Online, September. INCOMA Ltd.

Muñoz-Ortiz, A., D. Vilares, and C. Gómez-Rodríguez. 2023. Assessment of pre-trained models across languages and grammars. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–373, Nusa Dua, Bali, November. Association for Computational Linguistics.

Nadeem, M., A. Bethke, and S. Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August. Association for Computational Linguistics.

Nangia, N., C. Vania, R. Bhalerao, and S. R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November. Association for Computational Linguistics.

Newman, M. L., C. J. Groom, L. D. Handelman, and J. W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

Nivre, J. 2010. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152.

Perera, R. and P. Nand. 2016. Answer presentation in question answering over linked data using typed dependency subtree patterns. In *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*, pages 44–48, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Reddy, S., O. Täckström, M. Collins, T. Kwiatkowski, D. Das, M. Steedman, and M. Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

Ross, C., B. Katz, and A. Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online, June. Association for Computational Linguistics.

Rotman, G. and R. Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M., S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July. Association for Computational Linguistics.

Schwemmer, C., C. Knight, E. D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, and J. W. Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171.

Sheng, E., K.-W. Chang, P. Natarajan, and N. Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online, November. Association for Computational Linguistics.

Sheng, E., K.-W. Chang, P. Natarajan, and N. Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November. Association for Computational Linguistics.

Sobhani, N. and S. Delany. 2024. Towards fairer NLP models: Handling gender bias in classification tasks. In A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, and D. Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 167–178, Bangkok, Thailand, August. Association for Computational Linguistics.

Stanczak, K. and I. Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168.*

Stanovsky, G., N. A. Smith, and L. Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Strzyz, M., D. Vilares, and C. Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 717–723, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Strzyz, M., D. Vilares, and C. Gómez-Rodríguez. 2020. Bracketing encodings for 2-planar dependency parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2472–2484, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Sun, T., A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Swayamdipta, S., R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November. Association for Computational Linguistics.

Tal, Y., I. Magar, and R. Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, and H. Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington, July. Association for Computational Linguistics.

Thakur, V. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162.*

Tian, Y., G. Chen, and Y. Song. 2021. Enhancing aspect-level sentiment analysis with word dependencies. In *Proceedings* *of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3726–3739, Online, April. Association for Computational Linguistics.

Vashishtha, A., K. Ahuja, and S. Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada, July. Association for Computational Linguistics.

Waldis, A., Y. Perlitz, L. Choshen, Y. Hou, and I. Gurevych. 2024. Holmes: Benchmark the linguistic competence of language models. *arXiv preprint arXiv:2404.18923*.

Wang, T., J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319.

Wang, W., H. Bai, J.-t. Huang, Y. Wan, Y. Yuan, H. Qiu, N. Peng, and M. Lyu. 2024. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3781–3789.

Webster, K., M. Recasens, V. Axelrod, and J. Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

You, Z., H. Lee, S. Mishra, S. Jeoung, A. Mishra, J. Kim, and J. Diesner. 2024. Beyond binary gender labels: Revealing gender bias in LLMs through gender-neutral name predictions. In A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, and D. Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 255–268, Bangkok, Thailand, August. Association for Computational Linguistics.

Zhang, M., Z. Li, G. Fu, and M. Zhang. 2019. Syntax-enhanced neural machine translation with syntax-aware word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1151–1161, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zheng, J., F. Fan, and J. Li. 2024. Incorporating lexical and syntactic knowledge for unsupervised cross-lingual transfer. *arXiv preprint arXiv:2404.16627*.

A Additional results for training dynamics

Figure 6 presents the data map generated during the training of DistilBERT in a single-task setup for the *about* dimension. A large portion of the samples exhibits high confidence and low variability, placing them in the easy-to-learn category. However, the figure also highlights areas for improvement, as a notable number of samples display low confidence, low variability, and low correctness values. Specifically, samples represented in blue, orange, and green indicate instances where the correctness rate falls below 0.5. These samples may be ambiguous or difficult to learn, suggesting the potential to explore better representations for these inputs to facilitate their learning.

Figure 7 illustrates the representation of the training samples obtained after training the DistilBERT model on the *as* dimension without incorporating explicit grammatical information. In this representation, the *variability* parameter exhibits very low values, while the *confidence* level achieved is moderate. This indicates that no samples were identified during this training process that could be categorized as *easy-to-learn* or *ambiguous*. The absence of these two regions, which are crucial for effective model learning, might explain the low performance observed, with an average F1 score of only 0.44 when evaluating this model.

Figure 8 presents the data map generated during the training of DistilBERT in a single-task setup for the *to* dimension. In this representation, the values for the *confidence* and *variability* metrics are notably lower than for the *about*. This phenomenon aligns with the low performance observed in the single-task models trained on the *to* dimension. Despite efforts to improve these results through hyperparameter adjustments, no significant improvements were achieved.

B Model Training Hyperparameters

Parameter	**Value**
Learning Rate	1×10^{-16}
Number of Epochs	100
Weight Decay	0.1
Optimizer	AdamW
Loss Function	Cross Entropy Loss
Dropout Rate	0.1
Encoder Dimension	768
LSTM Hidden Dimension	128
Embedding Dimension	100

Table 5: Hyperparameters used to train the models employed in this work

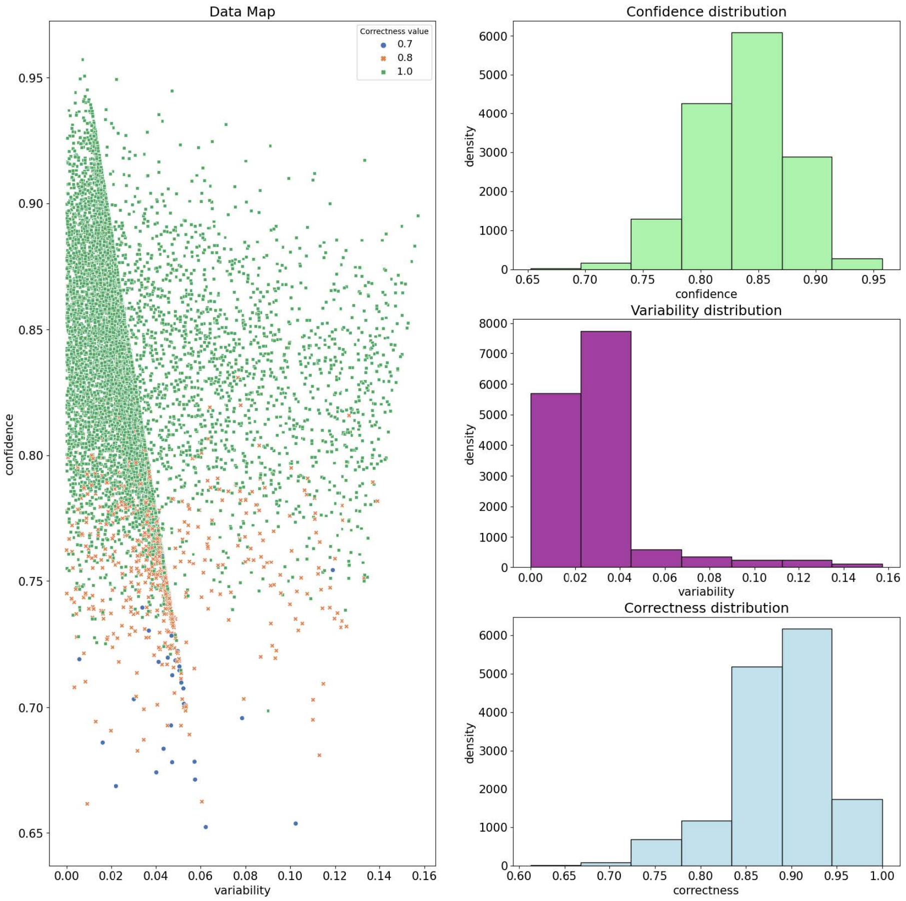


Figure 6: Data mapping of the ConvAI2 training dataset used to train a DistilBERT baseline model in a single-task setup for the *about* dimension.

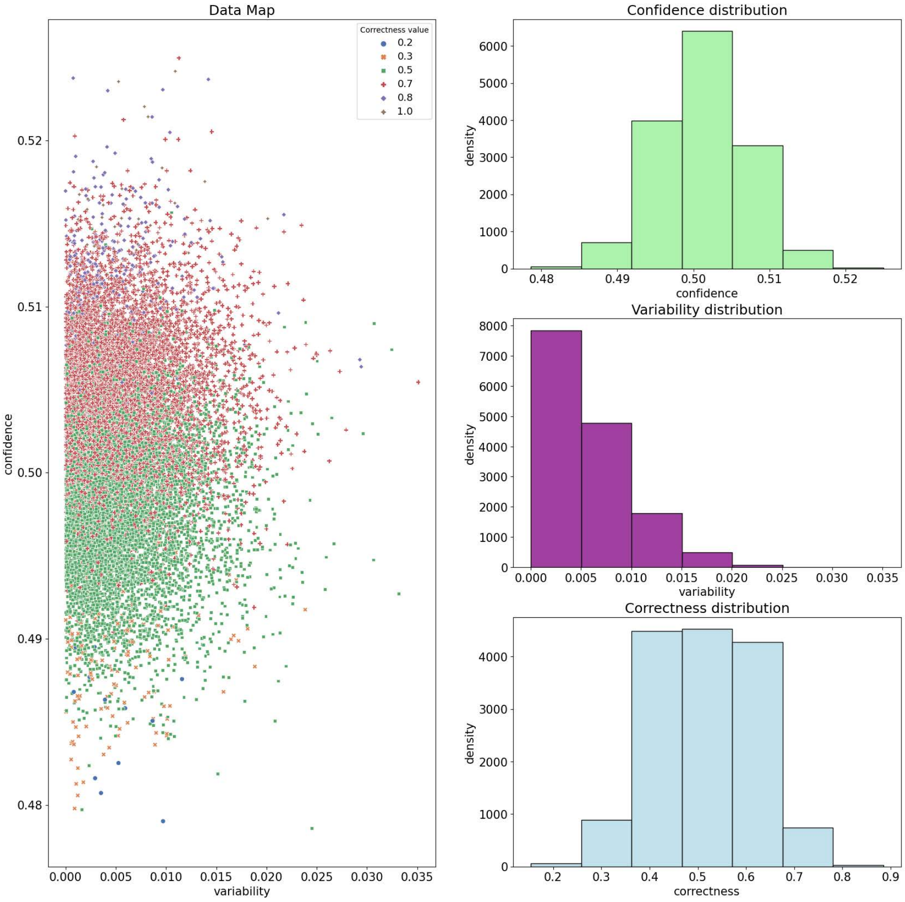


Figure 7: Data mapping of the ConvAI2 training dataset used to train a DistilBERT baseline model in a single-task setup for the *as* dimension.

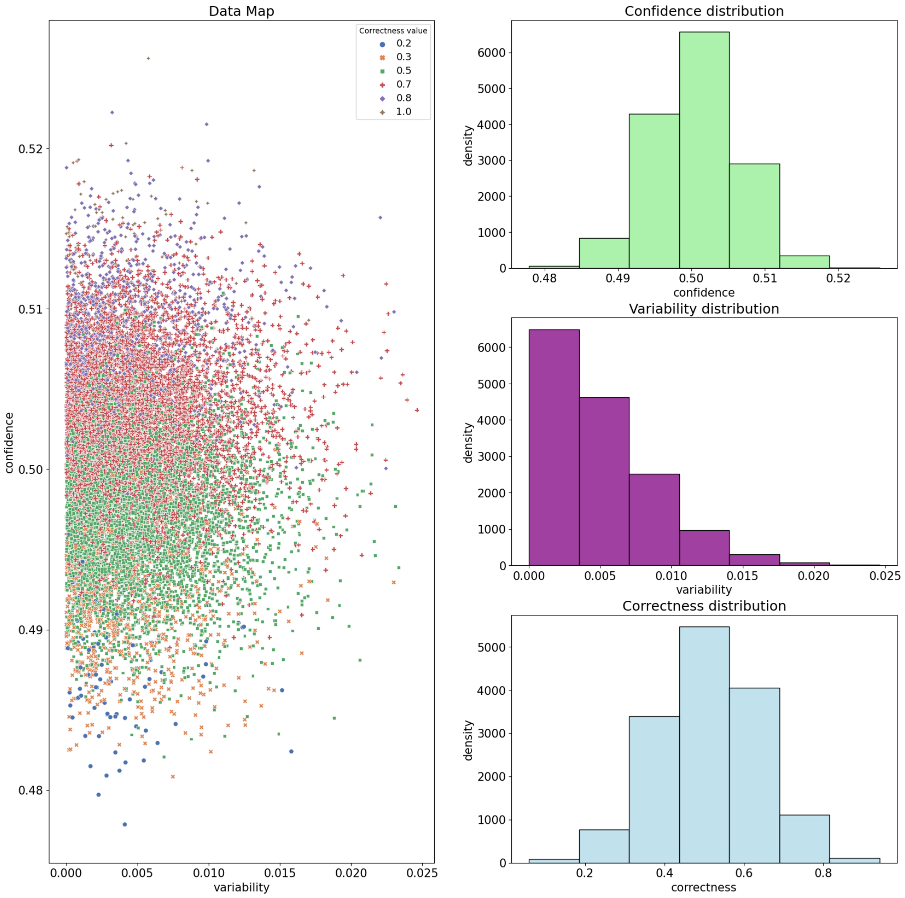


Figure 8: Data mapping of the ConvAI2 training dataset used to train a DistilBERT baseline model in a single-task setup for the *to* dimension.