Spanish Fashion NER: A Reproducible Framework with Fine-Grained Taxonomy and LLM Evaluation

Reconocimiento de Entidades en Moda en Español: una Metodología Reproducible con Taxonomía Detallada y Evaluación mediante Modelos de Lenguaje

María Villalba-Osés,[1] **Juan Pablo Consuegra-Ayala**[1,2] **Manuel Palomar,**[1,2,3]
[1] Digital Intelligence Centre (CENID), University of Alicante
[2] Department of Language and Computing Systems, University of Alicante
[3]Digital Intelligence Center (CENID), University of Alicante {maria.villalba, juan.consuegra}@ua.es
mpalomar@dlsi.ua.es

Abstract: The fashion domain generates large amounts of textual data, but transforming this unstructured content into structured information remains challenging, and Named Entity Recognition (NER) resources, especially beyond English, remain limited. This paper proposes a reproducible methodology for building domain-specific NER resources. As a case study, we introduce FashionNLP++, a Spanish resource for fashion NER, consisting of a fine-grained taxonomy and an annotated evaluation dataset. We evaluate several state-of-the-art Large Language Models (LLMs) in a zero-shot setting. Results show that, while LLMs achieve strong performance, their outputs vary in precision, consistency, and compliance with annotation constraints. All resources are publicly available to support research in domain-specific NLP and under-resourced languages.
Keywords: Moda, Reconocimiento de Entidades Nombradas, Procesamiento de Lenguaje Natural

Resumen: El dominio de la moda genera grandes cantidades de datos textuales, pero transformarlos en información estructurada sigue siendo un desafío, y los recursos para reconocimiento de entidades nombradas (NER), especialmente fuera del inglés, son limitados. Este trabajo propone una metodología reproducible para construir recursos NER específicos de dominio. Como caso de estudio, se presenta FashionNLP++, un recurso en español para NER en moda, con una taxonomía de grano fino y un conjunto de datos anotado, evaluado con modelos de lenguaje de gran tamaño (LLMs) en zero-shot. Los resultados muestran que, aunque los LLMs alcanzan buen rendimiento, presentan variaciones en precisión, consistencia y cumplimiento del formato. Los recursos se publican para apoyar la investigación en PLN.
Palabras clave: Fashion, Named Entity Recognition, Natural Language Processing

1 Introduction

The fashion industry generates large amounts of textual content across sources such as e-commerce platforms, magazines, and social media. These texts contain detailed descriptions of garments, materials, colors, styles, and accessories, making them a potentially valuable source of information for applications such as trend analysis (Sleiman, Tran, and Thomassey, 2022; Shen, 2023; Kim and Park, 2023), recommendation systems (Chakraborty et al., 2021; Liu et al., 2024), and semantic search (An and Park, 2020; Moro, Salvatori, and Frisoni, 2023). However, despite the abundance of raw textual data, the availability of structured linguistic resources for research in Natural Language Processing (NLP) within the fashion domain remains limited.

In particular, the development of domain-specific resources such as annotated corpora, taxonomies, and benchmarks (Deldjoo et al., 2023) for tasks like Named Entity Recognition (NER) is still scarce. Existing research on AI in fashion has largely focused on vi-

sual or multimodal analysis (Cheng et al., 2021), while textual resources specifically designed for NLP research have received comparatively less attention. This limitation becomes even more pronounced for languages other than English, such as Spanish, where annotated datasets and domain-specific linguistic resources for fashion are still largely unavailable.

NER is a key NLP task for structuring information in text by identifying and categorizing relevant entities (Pakhale, 2023). In the context of fashion, NER could enable the automatic extraction of domain-specific information such as garment types, materials, colors, etc. However, the absence of well-defined taxonomies and annotated textual datasets in this domain represents a significant challenge for training and evaluating NER systems. Specifically, there is a lack of high-quality annotated corpora and standardized taxonomies publicly available that are specifically tailored to the fashion industry, particularly for languages other than English, which further complicates reproducibility and comparative evaluation across studies.

To address these limitations, this paper aims to design and validate a methodology for constructing domain-specific resources in Spanish for NER in the fashion domain. The proposed approach includes the design of a fashion-oriented taxonomy of entities, the creation of an annotated textual dataset for evaluation, and the evaluation of several state-of-the-art Large Language Models (LLMs) on the resulting NER task. Compared to other works, the proposal of this paper provides a reproducible framework that integrates taxonomy design, corpus construction, and the evaluation of the resulting resources through state-of-the-art LLMs, with a specific focus on under-resourced languages such as Spanish.

The specific contributions of this research are as follows:

- The proposal of a reproducible methodology for building domain-specific NER resources in low-resource settings, validated through its application to the fashion domain and transferable to other specialized areas.
- The definition of a domain-specific taxonomy of fashion entities, extending existing resources (such as *FashionNLP v2*), and grounded in both fashion literature and empirical analysis of textual data.
- The construction of a Spanish annotated corpus for NER in the fashion domain, obtained through the adaptation and re-annotation of existing resources.
- A qualitative evaluation of several state-of-the-art LLMs on the proposed dataset, providing insights into their capabilities and limitations in recognizing domain-specific fashion entities.

The remainder of this paper is organized as follows: Section 2 reviews previous studies on NER in the fashion domain and existing linguistic resources for domain-specific NLP. Section 3 describes the proposed methodology for taxonomy design, corpus construction, and annotation. Section 4 presents the evaluation of the generated resources using state-of-the-art LLMs on the NER task, along with the corresponding results and discussion. Finally, Section 5 summarizes the main conclusions and outlines directions for future work.

2 Related Work

While NLP techniques have increasingly been applied to analyze textual data in the fashion domain, research in this area remains relatively fragmented. The following sections review existing work on NLP applications in fashion, advances in entity recognition approaches, and the availability of datasets and resources supporting research in this domain.

2.1 NLP in the Fashion Domain

NLP techniques have increasingly been applied to the fashion domain to analyze textual information from resources such as product descriptions (Harrag, Touameur, and Behilil, 2025; Bist, Gurbaxani, and Gupta, 2024), fashion editorials (Sleiman, Tran, and Thomassey, 2022; Li and Leonas, 2021; Beheshti-Kashi et al., 2015), and user-generated content (e.g., images, reviews and social media) (Joseph, Lora, and others, 2024; Yuan and Lam, 2022; Villalba-Osés, Consuegra-Ayala, and Palomar, 2026). These textual sources have been used to support various applications, including fashion recommendation systems (Chakraborty et al., 2021; Liu et al., 2024), trend analysis (Sleiman, Tran, and Thomassey, 2022; Shen, 2023; Kim and Park, 2023), and product categorization (Norman et al., 2018; Peng et al., 2022).

Across recent studies, common NLP tasks include text classification (e.g., predicting categories or styles), sentiment analysis (e.g., modeling customer opinions), and information extraction to transform unstructured descriptions into structured attributes, such as Named Entity Recognition, which is commonly used to extract and structure fashion-related information from textual descriptions, enabling its use in applications such as product indexing, recommendation systems, and trend analysis.

2.2 Named Entity Recognition in Fashion

NER is a fundamental task in NLP that aims to detect and classify relevant entities mentioned in textual data (Pakhale, 2023). Traditionally, NER systems were developed using rule-based approaches and statistical models such as Conditional Random Fields (Khan et al., 2022). In recent years, however, neural architectures based on deep learning ——particularly transformer-based models—— have significantly improved performance in entity recognition tasks (Zaratiana et al., 2024; Lothritz et al., 2020; Abdullah et al., 2024; Kaur et al., 2024).

More recently, LLMs have also been explored for NER (Hu et al., 2024; Xie et al., 2024; Majdik et al., 2024; Stefanelli, Maggini, and Rigutini, 2025; Wang et al., 2025), either through fine-tuning or prompt-based approaches. Despite these advances, the effectiveness of NER systems strongly depends on the availability of annotated datasets and clearly defined entity categories, which remain limited in many specialized domains (Jehangir, Radhakrishnan, and Agarwal, 2023).

In the fashion domain, NER plays a key role in structuring descriptive information (Mousterou, 2024a), enabling the transformation of unstructured textual data into interpretable attributes that can be leveraged in downstream applications such as recommendation, search, and trend analysis (Kashilani, Damahe, and Thakur, 2018; Liu et al., 2024; Shen, 2023; Kim and Park, 2023; Norman et al., 2018). However, the development and evaluation of such systems remain constrained by the limited availability of domain-specific annotated datasets and well-defined entity taxonomies (Zou et al., 2019).

2.3 Fashion Resources for NLP tasks

Several datasets and resources have been developed to support research in the fashion domain. However, a large portion of these resources focuses primarily on visual data, such as clothing images used for tasks including visual recognition tasks such as fashion item classification (Kashilani, Damahe, and Thakur, 2018; Kadam, Adamuthe, and Patil, 2020), attribute recognition (Parekh et al., 2022; Ge et al., 2019), and fashion (Moro, Salvatori, and Frisoni, 2023; Murtaza et al., 2025). Many of these datasets combine images with textual metadata, enabling multimodal approaches to fashion analysis.

Despite the existence of some studies addressing textual data in the fashion domain (Sleiman, Tran, and Thomassey, 2022), such as FashionNLP v2 (eXascale Infolab, 2024), which provides linguistic resources and tools for processing fashion-related textual data, and NER-Luxury (Mousterou, 2024a), which focuses on named entity recognition for luxury fashion brands, resources specifically designed for systematic textual analysis remain limited. In particular, annotated corpora suitable for NLP tasks such as NER are still scarce. This scarcity is even more pronounced for languages other than English, where domain-specific datasets and linguistic resources are largely unavailable.

Furthermore, existing resources often lack clearly defined taxonomies that capture the semantic categories commonly used to describe fashion items and attributes. This limitation makes it difficult to consistently extract and structure information from fashion-related texts. For instance, existing resources such as *FashionNLP v2* (eXascale Infolab, 2024) focus on a predefined set of attributes, which may not fully capture the diversity and granularity of fashion descriptions, particularly in aspects such as fabric, style, context, or nuanced garment details. Similarly, domain-specific datasets like NER-Luxury (Mousterou, 2024b) are often restricted to specific segments (e.g., luxury brands), limiting their generalizability to broader fashion contexts. This lack of coverage becomes critical in fashion analysis, where visual and functional attributes play a central role in capturing trends, user preferences, and the semantic richness of fashion descriptions.

2.4 Our Outlook

Compared to existing work, this paper proposes a reproducible methodology for the development of domain-specific NER resources in the fashion domain. While prior studies have either focused on general-purpose NLP techniques or on specific datasets with limited scope, our approach integrates taxonomy design, corpus construction, and systematic evaluation using state-of-the-art LLMs on the resulting NER task.

In addition, unlike existing resources that often rely on predefined or coarse-grained attribute sets, the proposed taxonomy captures a broader and more fine-grained range of fashion-related entities, including garments, shoes, accessories, fabric, colors, style, context and garment detail. This enables a richer and more structured representation of fashion descriptions.

Furthermore, this work specifically addresses the lack of resources for languages other than English by introducing FashionNLP++, a Spanish annotated dataset, contributing to the development of NLP resources in under-resourced settings. Finally, the proposed resources are systematically evaluated using state-of-the-art LLMs on the NER task, providing insights into their effectiveness and applicability.

3 Method

This section describes a reproducible methodology for constructing domain-specific resources for NER in the fashion domain. The proposed approach is designed to be generalizable across datasets and languages, and is structured into three main stages: (i) the design of a domain-specific taxonomy, (ii) the construction or adaptation of a textual corpus, and (iii) the annotation process.

In this work, the methodology is instantiated using the textual data provided by the *FashionNLP v2* resource (eXascale Infolab, 2024) as the initial corpus, resulting in a domain-specific resource for evaluation. However, the process described is dataset-agnostic and can be applied to any domain-relevant textual collection. The resources developed in this work, including the FashionNLP++ dataset, the proposed taxonomy, and the evaluation prompt, are publicly available.[1]

[1] `https://doi.org/10.5281/zenodo.20625537`

3.1 Taxonomy Design

The definition of an appropriate entity taxonomy is a key step in the development of domain-specific resources for NER. This stage involves identifying the semantic dimensions that characterize how concepts are described within a given domain and organizing them into a structured set of annotation categories.

The taxonomy construction process consists of the following steps:

1. To define the taxonomy, relevant domain **knowledge sources** can be considered, including specialized literature, existing resources, and domain-specific textual data. These sources provide information about the attributes, components, and descriptors that are commonly used to refer to entities in natural language, and therefore help determine which semantic dimensions should be explicitly represented in the annotation scheme.
2. Based on this information, a taxonomy can be constructed to capture the main **semantic elements** present in domain-related textual descriptions. The objective is to provide a structured representation of the attributes and properties that are most relevant for the target domain, while keeping the taxonomy interpretable and applicable to annotation tasks.
3. The taxonomy may also be adapted to the **target language** in order to ensure that the entity categories align with the terminology commonly used in that linguistic context. This is particularly important in under-resourced settings, where domain-specific resources are often limited or unavailable.

In this work, the taxonomy was defined for the fashion domain in Spanish. To support its design, we relied on established literature in fashion studies and design, including Fashion Design: The Complete Guide" (Hopkins, 2022) and The Fundamentals of Fashion Design" (Sorger and Udale, 2006). These sources describe garments through multiple semantic dimensions, such as clothes, color, fabric, style, context, garment detail and accessories, which are central to how fashion items are represented in textual descriptions.

Based on these dimensions, the taxonomy was defined through a selection and adaptation process oriented to the NER task. The

process was conducted by a researcher with a background in NLP and AI applied to fashion, as well as expertise in fashion design, ensuring that the resulting categories were domain-relevant and suitable for information extraction.

Following this process, the final taxonomy consists of a set of entity categories that capture the main semantic components of fashion-related discourse in Spanish (see Table 1).

3.2 Corpus Construction

The next step in the proposed methodology consists of constructing a textual corpus that serves as the basis for evaluating NER systems in the target domain. This process involves selecting a domain-relevant textual resource, preprocessing the textual data, and adapting it to the target language and annotation scheme. The following steps were followed to construct the corpus:

1. First, a preprocessing step may be performed to clean and standardize the textual data. This can include operations such as removing existing annotation labels, normalizing text formats, or filtering irrelevant content, depending on the characteristics of the source corpus.
2. Next, the corpus may be adapted to the target language when necessary. This step can involve automatic translation, followed by manual revision to ensure the preservation of semantic content, particularly for domain-specific terminology.
3. Finally, the corpus is organized according to the requirements of the task. Depending on the intended use, this may involve defining one or more dataset splits (e.g., training, validation, or evaluation). The resulting corpus constitutes the base textual resource that is subsequently annotated following the proposed taxonomy.

In this work, the proposed methodology was instantiated to construct our resource, which consists of a Spanish-language corpus. Since the initial data was not available in the target language, it was adapted following the described pipeline. First, the original FashionNLP v2 annotations were removed to obtain the raw text. This decision was motivated by the fact that a direct translation of the annotated corpus did not preserve the alignment between entities and tokens in Spanish. The resulting raw corpus was then translated using GPT-4o and manually revised to preserve domain-specific semantics. Finally, the translated corpus was re-annotated according to the proposed taxonomy, serving as the basis for the annotation process. The resulting corpus was organized as a single evaluation split, which was subsequently used in all experiments reported in this work.

3.3 Annotation Procedure

The next step in the proposed methodology consists of annotating the corpus according to the defined entity taxonomy and a standard labeling scheme for NER. This step aims to generate an annotated dataset suitable for evaluating NER models in the target domain.

The following algorithmic procedure describes the proposed annotation process in a generic manner. This formulation is intended to be applicable to different domains and languages, while the specific instantiation adopted in this work is described below. The resulting annotated corpus follows a standard NER labeling format and is aligned with the proposed taxonomy, making it suitable for evaluating domain-specific NER models.

1. First, an annotation scheme is selected. In many NER settings, this involves assigning a label to each token in the text according to a predefined format, such as the B-I-O scheme, which distinguishes between the beginning of an entity, tokens inside an entity, and tokens outside any entity.
2. Next, entity labels are assigned based on the defined taxonomy. Each token in the corpus is annotated according to its corresponding entity type when applicable, ensuring alignment between the textual data and the semantic categories defined in the taxonomy.
3. To facilitate the annotation process, a semi-automatic approach may be adopted. This typically involves generating preliminary annotations using automated methods, followed by manual revision to correct errors and refine entity boundaries. This strategy balances efficiency and annotation quality.
4. Finally, a validation step can be performed to ensure consistency across the

Label	Description	Examples
Ropa (clothes)	*Main clothing items worn on the body*	dress, jacket, shirt, skirt
Zapatos (shoes)	*Footwear items*	boots, sneakers, sandals
Accesorios (accessories)	*Complementary fashion items used with clothing*	belt, necklace, handbag
Color (colors)	*Color attributes describing fashion items*	red, black, navy blue
Contexto (context)	*Situational or contextual references related to clothing use*	festival, beach, office
Estilo (style)	*Stylistic descriptors of fashion items*	vintage, casual, minimalist
Detalle de la prenda (garment detail)	*Specific parts or decorative elements of garments*	sleeves, collar, buttons
Tela (fabric)	*Materials used in garments*	denim, silk, cotton
Marcas (brands)	*Fashion brands or labels*	Nike, Gucci, Zara

Table 1: Proposed taxonomy for Spanish fashion NER.

dataset. This may include reviewing ambiguous cases, verifying multi-word expressions, and ensuring that annotation guidelines are applied uniformly throughout the corpus.

In this work, this procedure was applied to annotate the constructed corpus using the B-I-O annotation scheme (Beginning, Inside, Outside) (Ramshaw and Marcus, 1995) and the defined taxonomy. First, preliminary annotations were automatically generated using ChatGPT based on the proposed taxonomy. These annotations were then imported into the INCEpTION annotation platform and manually reviewed and corrected to ensure consistency with the annotation guidelines. The resulting resource constitutes a Spanish-language evaluation dataset for fashion NER.

4 Experimentation

This section presents the experimental analysis of the proposed resource. First, Section 4.1 describes the experimental setup, including the defined evaluation scenarios. Then, Section 4.2 presents the main statistics of the annotated dataset, including its structure and entity distribution. Finally, Section 4.3 reports the evaluation results obtained by LLMs on the NER task in the fashion domain.

4.1 Experimental Setup

To assess the effectiveness of the proposed resource and methodology, two evaluation scenarios are defined. Each scenario addresses a different aspect of the contribution: (i) the analysis of the constructed dataset and its properties, and (ii) the evaluation of its usefulness for NER tasks using large language models.

Scenario 1: Dataset Statistics The objective of this scenario is to analyze the characteristics of the constructed resource and compare it with the original data source. This includes examining the distribution of entity categories, the coverage of the proposed taxonomy, and the overall composition of the dataset.

Scenario 2: Model Evaluation The objective of this scenario is to evaluate the performance of large language models on the NER task using the constructed dataset. In this setting, models are provided with the defined taxonomy and are required to extract entities from the test corpus. Beyond assessing the constructed resource, this scenario also provides a comparative analysis of large language models, offering insights into their suitability for domain-specific NER tasks.

The evaluation is conducted from two complementary perspectives. First, a quantitative analysis is performed using standard NER metrics, including precision, recall, and F1-score. Second, a qualitative analysis is carried out to assess the practical applicability of large language models for domain-specific NER tasks. In particular, the analysis considers ease of use, first-pass annotation quality, compliance with the BIO annotation format, adherence to the defined taxonomy, and

Metric	Value
Number of sentences	1,872
Number of tokens	62,200
Min. sentence length (tokens)	3
Max. sentence length (tokens)	195
Avg. sentence length (tokens)	33.23
Std. deviation	19.48

Table 2: Dataset statistics.

consistency across outputs.

4.2 Scenario 1: Dataset Statistics

This section presents the main statistics of the constructed dataset. The analysis focuses on the distribution of entity categories, the overall composition of the corpus, and the coverage of the proposed taxonomy. Specifically, the analysis includes corpus-level statistics such as the number of sentences and tokens, sentence length characteristics, and the distribution of entity annotations across categories. Additionally, a comparison with the base taxonomy is performed to analyze how the proposed taxonomy extends and enriches the semantic representation of fashion-related concepts.

4.2.1 Results

Table 2 summarizes the main statistical properties of the constructed dataset, including the number of sentences and annotated tokens, as well as sentence length statistics measured in tokens. The dataset contains 4,943 annotated entities. Multi-word entities account for 14.84% of all annotations, with an average entity length of 1.20 tokens and a maximum entity length of 5 tokens. No nested entities are present in the corpus.

Figure 1 shows a comparison between the base taxonomy (FashionNLP v2) and the proposed taxonomy (FashionNLP ++) in terms of the number of entity categories. The proposed taxonomy introduces additional categories that allow for a more fine-grained representation of fashion-related attributes.

Table 3 presents the distribution of entity annotations, comparing the annotation scheme of the base dataset with the fine-grained categories introduced in the proposed taxonomy. Although FashionNLP v2 organizes fashion concepts into categories such as clothes, shoes, and accessories, all entities in the annotated corpus are labeled using the single ITEM category. In contrast, FashionNLP++ adopts a fine-grained annotation scheme that explicitly distinguishes between different types of fashion-related entities.

Entity	FashionNLP v2	FashionNLP++
ITEM	683	–
ROPA	–	1848
ACCESORIOS	–	562
ZAPATOS	–	352
COLOR	–	527
TELA	–	485
MARCAS	–	451
CONTEXTO	–	309
ESTILO	–	248
DETALLE	–	161

Table 3: Comparison of entity distribution between the base and proposed datasets.

4.2.2 Discussion

As shown in Table 2, the constructed dataset exhibits a relatively large size in terms of both number of sentences and tokens, with a wide variability in sentence length. The high standard deviation indicates that the dataset includes both short and long descriptive sentences, which is consistent with the heterogeneous nature of fashion-related texts.

Figure 1 highlights the extension of the base taxonomy by introducing additional entity categories beyond garments, footwear, and accessories. This expansion enables a richer semantic representation by incorporating attributes such as color, context, style, garment detail, material and brands, which are essential for capturing the multifaceted nature of fashion descriptions.

Furthermore, as shown in Table 3, the base dataset relies on a single-category labeling scheme (ITEM), whereas the proposed dataset distributes annotations across multiple semantic categories. This transition reflects a shift from a general representation to a more detailed and structured annotation scheme.

In terms of entity distribution, *ROPA* is the most frequent category, followed by *ACCESORIOS* and *MARCAS*, which aligns with the descriptive focus of fashion texts on garments and their visual attributes. In contrast, categories such as *DETALLE_DE_LA_PRENDA* and *CONTEXTO* appear less frequently, as they capture more specific or nuanced characteristics.

Overall, the results show that the proposed dataset provides a more balanced and informative representation of fashion-related concepts, which is expected to benefit downstream tasks such as entity recognition and trend analysis.

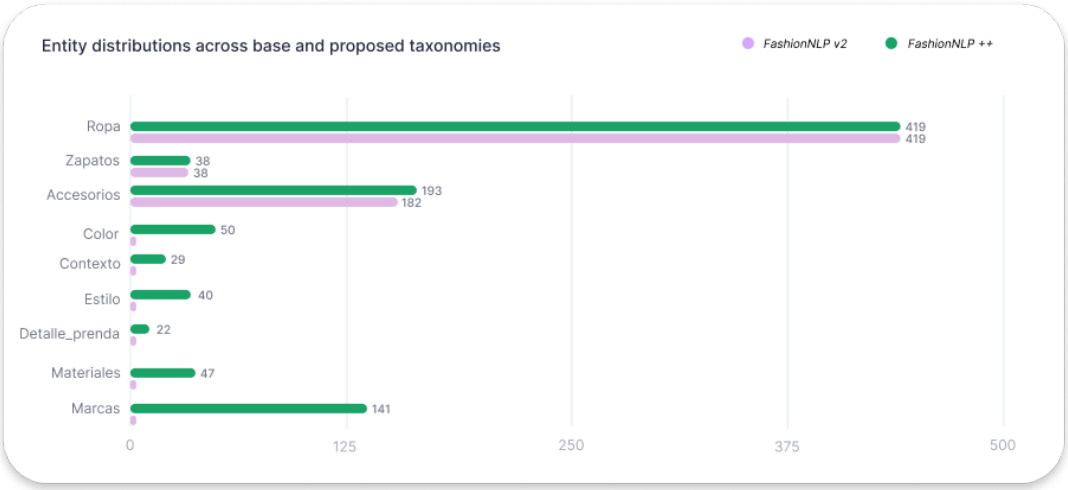

Figure 1: Comparison between the base taxonomy and the proposed extended taxonomy in terms of entity coverage.

4.3 Scenario 2: Model Evaluation

This section evaluates the performance of large language models, particularly GPT-4o (ChatGPT), Gemini 2.5 Pro, and Claude 4 Sonnet, on the NER task using the constructed dataset. All experiments were conducted using the default parameters provided by each platform, with temperature set to 0 to minimize output variability. The goal is to assess both the effectiveness of the resource for entity extraction and the capability of different models to handle domain-specific NER tasks when provided with a predefined taxonomy. To ensure reproducibility, the prompt used to interact with the models is provided in Appendix A

The evaluation is conducted from both quantitative and qualitative perspectives, allowing for a comprehensive analysis of model behavior. The quantitative evaluation is based on standard NER metrics, including precision, recall, and F1-score. In addition, a qualitative analysis is performed considering ease of use, first-pass annotation quality, compliance with the BIO annotation format, adherence to the defined taxonomy, and consistency across outputs.

4.3.1 Results

Table 4 presents the quantitative results obtained by the evaluated models in terms of precision, recall, and F1-score. The results show differences in performance across models, highlighting variations in their ability to accurately identify and classify entities.

Table 5 shows the performance of the models across different entity categories.

Model	Precision	Recall	F1-score
ChatGPT	0.8563	0.8437	0.8500
Gemini	0.3790	0.7582	0.5054
Claude	0.7822	0.7133	0.7462

Table 4: Performance comparison of models against the gold standard.

Table 6 presents the qualitative comparison of the evaluated models across the selected dimensions. Specifically, ease of use refers to the simplicity of obtaining usable annotations; first-pass annotation quality evaluates the quality of the annotations before manual correction; format compliance measures adherence to the BIO annotation scheme; taxonomy adherence assesses compliance with the proposed entity categories; and consistency evaluates the stability of the generated annotations across outputs.

4.3.2 Discussion

As shown in Table 4, ChatGPT achieves the highest overall performance, obtaining the best balance between precision and recall, which results in the highest F1-score. Claude also demonstrates strong performance, although with slightly lower recall, leading to a moderate decrease in F1-score. In contrast, Gemini exhibits a different behavior, characterized by relatively high recall but substantially lower precision, indicating a tendency to over-generate entities and produce a higher number of false positives.

As shown in Table 5, this behavior is con-

	ChatGPT			Gemini			Claude		
Label	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
ACCESORIOS	0.9175	0.8099	0.8604	0.5722	0.7673	0.6555	0.9121	0.6270	0.7432
COLOR	0.9781	0.8482	0.9085	0.4601	0.7002	0.5553	0.7006	0.7192	0.7097
CONTEXTO	0.9134	0.9871	0.9488	0.2679	0.7258	0.3913	0.9049	0.9516	0.9277
DETALLE	0.7358	0.7222	0.7290	0.0409	0.4259	0.0746	0.8030	0.6543	0.7211
ESTILO	0.7220	0.9113	0.8057	0.1045	0.4476	0.1695	0.4596	0.6653	0.5437
MARCAS	0.9442	0.8620	0.9012	0.3168	0.8938	0.4678	0.9043	0.8429	0.8725
ROPA	0.7978	0.8777	0.8359	0.6483	0.8387	0.7313	0.7473	0.6369	0.6877
TELA	0.8652	0.7546	0.8062	0.4294	0.6021	0.5013	0.7785	0.7031	0.7389
ZAPATOS	0.9919	0.6932	0.8161	0.7649	0.8409	0.8011	0.9940	0.9347	0.9634

Table 5: Per-label precision, recall, and F1-score for each evaluated model.

Model	**Ease**	**First-pass**	**Format**	**Taxonomy**	**Consistency**
ChatGPT	✓	✓	✓	✓	✓
Gemini	∼	✗	✗	∼	✗
Claude	✓	∼	✓	✓	∼

Table 6: Qualitative comparison of models across practical evaluation dimensions. ✓ indicates high performance, ∼ indicates moderate performance, and ✗ indicates low performance.

sistent across entity categories. ChatGPT maintains stable and high performance across most labels, particularly in *COLOR*, *CONTEXTO*, and *MARCAS*, where both precision and recall are consistently high. Claude shows competitive results in several categories, although with greater variability, especially in *ESTILO* and *ROPA*. Gemini, on the other hand, presents strong recall across many categories but significantly lower precision, especially in *DETALLE* and *ESTILO*, suggesting difficulties in accurately distinguishing fine-grained or more abstract entity types.

Finally, as shown in Table 6, the qualitative evaluation reinforces the quantitative findings. ChatGPT demonstrates consistently strong performance across all evaluated dimensions, including ease of use, format compliance, and taxonomy adherence, indicating its robustness for domain-specific NER tasks. Claude also shows reliable behavior, particularly in terms of format compliance and taxonomy adherence, although with slightly lower consistency and first-pass annotation quality. In contrast, Gemini presents several practical limitations, including lower format compliance and reduced consistency, which may require additional post-processing or prompt refinement to obtain reliable outputs.

5 Conclusion

This paper presents a reproducible methodology for the construction of domain-specific resources for NER in the fashion domain. The proposed approach, structured around taxonomy design, corpus construction, and annotation, is designed to be generalizable across datasets and languages. As a concrete instantiation, we introduce a Spanish-language resource for fashion NER, including a domain-specific taxonomy and an annotated evaluation dataset. The experimental analysis demonstrates that the constructed resource provides a suitable benchmark for evaluating large language models in domain-specific NER tasks, revealing variability across models in terms of precision, consistency, and adherence to annotation constraints.

The main findings of this work are as follows:

- The proposed methodology was successfully applied to construct a domain-specific NER resource in the fashion domain.
- Two valuable resources for Spanish were developed: a fine-grained taxonomy and a manually curated annotated corpus for evaluation.
- Three state-of-the-art LLMs were evaluated on the proposed dataset, with Chat-

GPT achieving the best quantitative performance.

- A qualitative analysis highlighted the strengths and limitations of LLMs for fashion NER, particularly in terms of consistency and handling fine-grained entity distinctions.

Overall, this work contributes both a reusable methodology and a domain-specific resource, addressing the scarcity of annotated data in non-English settings and supporting future research in fashion-related NLP tasks.

5.1 Future Work

Future work will focus on extending the proposed methodology to other domains and languages, as well as exploring strategies to improve the robustness and reliability of LLMs in domain-specific NER scenarios. Additionally, future research will explore the construction of resources from native Spanish fashion texts in order to further improve linguistic naturalness and reduce potential translation-related biases. Future studies will also investigate the potential impact of data contamination when evaluating closed LLMs on translated resources. More importantly, the proposed resource will be applied to analyze fashion trends in reference sources such as fashion magazines (e.g., Vogue), enabling the extraction of structured insights from real-world fashion content.

Acknowledgements

This research has been funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the project: NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/021). It has also been funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by the EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

A Prompt Used for LLM Evaluation

To ensure reproducibility, this appendix provides the prompt used to interact with the evaluated LLMs for the NER task.

The prompt instructs the models to extract entities from the input text according to the predefined taxonomy and to return the output following the BIO annotation scheme. The same prompt was used across all evaluated models.

```
< You are an expert in Named
    Entity Recognition in the
    fashion domain.
You are provided with a taxonomy
    file. You must strictly follow
     this taxonomy when annotating
    .
Task:
Annotate the following text using
     the BIO tagging scheme.
Use ONLY these entity categories:
ROPA, ZAPATOS, ACCESORIOS, COLOR,
     CONTEXTO, ESTILO,
    DETALLE_DE_LA_PRENDA, TELA,
    MARCAS
Instructions:
- Label each token with one tag
- Use B- for beginning, I- for
    inside, O otherwise
- Preserve the exact token order
- Do not modify the text
- Do not add explanations
- Return ONLY the annotated text
    in BIO format
Text: >
```

References

Abdullah, A. A., S. H. Abdulla, D. M. Toufiq, H. S. Maghdid, T. A. Rashid, P. F. Farho, S. S. Sabr, A. H. Taher, D. S. Hamad, H. Veisi, et al. 2024. Ner-roberta: Fine-tuning roberta for named entity recognition (ner) within low-resource languages. *arXiv preprint arXiv:2412.15252*.

An, H. and M. Park. 2020. Approaching fashion design trend applications using text mining and semantic network analysis. *Fashion and Textiles*, 7(1):34.

Beheshti-Kashi, S., M. Lutjen, L. Stoever, and K.-D. Thoben. 2015. Trendfashion-a framework for the identification of fashion trends. In *INFORMATIK 2015*, pages 1195–1205. Gesellschaft für Informatik eV.

Bist, Y., P. Gurbaxani, and N. Gupta. 2024. Production classification in e-commerce based on product descriptions with natural language processing (nlp) and machine learning models. In *International Conference on Cognitive Computing and Cyber Physical Systems*, pages 39–46. Springer.

Chakraborty, S., M. S. Hoque, N. Rahman Jeem, M. C. Biswas, D. Bardhan, and E. Lobaton. 2021. Fashion recommendation systems, models and methods: A review. In *Informatics*, volume 8, page 49. MDPI.

Cheng, W.-H., S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu. 2021. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41.

Deldjoo, Y., F. Nazary, A. Ramisa, J. Mcauley, G. Pellegrini, A. Bellogin, and T. D. Noia. 2023. A review of modern fashion recommender systems. *ACM Computing Surveys*, 56(4):1–37.

eXascale Infolab. 2024. Fashionnlp v2. GitHub repository. Accessed: June 10, 2026.

Ge, Y., R. Zhang, X. Wang, X. Tang, and P. Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345.

Harrag, F., O. Touameur, and H. Behilil. 2025. Hybrid nlp model for automating fashion product descriptions: integrating transformers and word embeddings. *Performance Measurement and Metrics*, 26(2):88–104.

Hopkins, J. 2022. *Fashion Design: The Complete Guide*. Bloomsbury Publishing.

Hu, Y., Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Jehangir, B., S. Radhakrishnan, and R. Agarwal. 2023. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Joseph, V., C. P. Lora, et al. 2024. Exploring the application of natural language processing for social media sentiment analysis. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–6. IEEE.

Kadam, S. S., A. C. Adamuthe, and A. B. Patil. 2020. Cnn model for image classification on mnist and fashion-mnist dataset. *Journal of scientific research*, 64(2):374–384.

Kashilani, D., L. B. Damahe, and N. V. Thakur. 2018. An overview of image recognition and retrieval of clothing items. In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pages 1–6. IEEE.

Kaur, N., A. Saha, M. Swami, M. Singh, and R. Dalal. 2024. Bert-ner: A transformer-based approach for named entity recognition. In *2024 15th international conference on computing communication and networking technologies (ICCCNT)*, pages 1–7. IEEE.

Khan, W., A. Daud, K. Shahzad, T. Amjad, A. Banjar, and H. Fasihuddin. 2022. Named entity recognition using conditional random fields. *Applied Sciences*, 12(13):6391.

Kim, H. and M. Park. 2023. Discovering fashion industry trends in the online news by applying text mining and time series regression analysis. *Heliyon*, 9(7).

Li, J. and K. K. Leonas. 2021. Sustainability topic trends in the textile and apparel industry: a text mining-based magazine article analysis. *Journal of Fashion Marketing and Management*, 26(1):67–87, 04.

Liu, Q., J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, and J. Tang. 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17.

Lothritz, C., K. Allix, L. Veiber, T. F. Bissyandé, and J. Klein. 2020. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In *Proceedings of the 28th international conference on computational linguistics*, pages 3750–3760.

Majdik, Z. P., S. S. Graham, J. C. Shiva Edward, S. N. Rodriguez, M. S. Karnes, J. T. Jensen, J. B. Barbour, and J. F. Rousseau. 2024. Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study. *Jmir ai*, 3:e52095.

Moro, G., S. Salvatori, and G. Frisoni. 2023. Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval. *Neurocomputing*, 538:126196.

Mousterou, A. 2024a. Ner-luxury: Named entity recognition for the fashion and luxury domain.

Mousterou, A. 2024b. Ner-luxury: Named entity recognition for the fashion and luxury domain. *arXiv preprint arXiv:2409.15804*.

Murtaza, M., M. Fayyaz, M. Yasmin, M. Anwar, K. N. Qureshi, and U. A. Raza. 2025. A large-scale dataset and robust multifeature representation with maximum correlation-based feature fusion and matching for apparel image retrieval. *Expert Systems*, 42(9):e70097.

Norman, K., Z. Li, Y.-T. Oh, G. Golwala, S. Sundaram, and J. Allebach. 2018. Application of natural language processing to an online fashion marketplace. *Electronic Imaging*, 30:1–5.

Pakhale, K. 2023. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*.

Parekh, V., S. Mathur, S. Biswas, and K. Shaik. 2022. Vper: Visual product entity recognition for fashion-wear. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 270–274.

Peng, D., R. Liu, J. Lu, and S. Zhang. 2022. Unsupervised multi-modal modeling of fashion styles with visual attributes. *Applied Soft Computing*, 115:108214.

Ramshaw, L. and M. Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Shen, Z. 2023. Mining sustainable fashion e-commerce: social media texts and consumer behaviors. *Electronic Commerce Research*, 23(2):949–971.

Sleiman, R., K.-P. Tran, and S. Thomassey. 2022. Natural language processing for fashion trends detection. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE.

Sorger, R. and J. Udale. 2006. *The Fundamentals of Fashion Design*. AVA Publishing.

Stefanelli, M., M. Maggini, and L. Rigutini. 2025. Comparative analysis of token classification and zero-shot llm approaches for named entity recognition in the business domain. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Villalba-Osés, M., J. P. Consuegra-Ayala, and M. Palomar. 2026. Understanding gender bias in text-to-image models through quantitative and interpretable analysis: A fashion case study. *Expert Systems*, 43(4):e70232.

Wang, S., X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, and C. Guo. 2025. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.

Xie, T., Q. Li, Y. Zhang, Z. Liu, and H. Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593.

Yuan, Y. and W. Lam. 2022. Sentiment analysis of fashion related posts in social media. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1310–1318.

Zaratiana, U., N. Tomeh, P. Holat, and T. Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Zou, X., X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao. 2019. Fashionai: A hierarchical dataset for fashion understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0.