A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

REBUILD PRODUCT CLASSIFICATION

Leveraging Dynamic Masked Softmax and Shared Hidden Layers for Hierarchical Text-Based Product Classification with BERT

Lotte Groß (51403)

Work project carried out under the supervision of:

Prof. Qiwei Han

19/01/2024

Abstract (Shared Part)
Using state-of-the-art models for text (BERT-based) and image (ResNet, VGG16, ViT) analysis, this study develops a multimodal approach that leverages the collective strengths of both domains. Our results show that the combination of knowledge from text and image domains leads to the best classification framework, which achieves a remarkable macro-F1 score of 98% at all levels. This innovative approach significantly improves classification accuracy and efficiency in e-commerce. In addition, the study explores self-supervised learning and introduces a detailed taxonomy that provides comprehensive insights. This research highlights the superiority of a synergistic multimodal strategy to improve product understanding.

Abstract (Individual Part)
This study explores the transformative impact of BERT and its variants, particularly RoBERTa, on hierarchical multi-class product classification. Leveraging the bidirectional nature of BERT, the research evaluates flat and hierarchical model architectures, revealing RoBERTa's superiority due to its nuanced understanding of diverse language styles in product titles. The hierarchical model, incorporating dynamic masked softmax, achieves a remarkable 96% accuracy in layer 2, showcasing efficient category handling. Despite longer training times, the innovative approach mitigates error propagation. The study emphasizes the trade-off between computational cost and interpretability, providing insights for future NLP research.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1 Introduction

In the ever-evolving e-commerce landscape, where giants such as Amazon, eBay, Taobao, and Rakuten manage an enormous number of products, the careful organization and classification of goods becomes the foundation for an improved user experience and seamless navigation (Zahavy et al. 2018). To understand the scale of this endeavor, consider the size of Rakuten's e-commerce platform, which, as of May 2016, hosted 186 million active products from 43,363 different merchants (Cevahir and Murakami 2016). The rapid growth of the industry requires efficient solutions, as manually categorizing this vast inventory would be a time-consuming and costly task, which highlights the urgency of sophisticated automated product classification systems (Bi, Wang, and Fan 2020).

However, tackling this task presents various challenges, such as the sparse distribution of products across multiple categories and skewed data distributions. Additionally, the diverse lengths of product titles and descriptions further complicate the classification process. Notably, despite the abundance of product data, ensuring accurate pairings between current product titles and assigned categories remains elusive (Cevahir and Murakami 2016). Finally, the persistent issue remains that different platforms employ varying product taxonomies, complicating a standardized and automated approach. Additionally, while these taxonomies provide a basic framework for hierarchical product categorization, they often fall short in adaptability and detail. For example, the hierarchical depth of the Google Product Taxonomy shoe category is limited to the second level, compared to the fifth level in other fashion categories. This inconsistency indicates a gap in the taxonomy's ability to encompass diverse products and provide granular categorizations across categories.

For GRIPS Intelligence, refining product classification is critical to facilitate strategic decision-making for its clients. Streamlining the product categorization would enable seamless

comparison of offerings with competitors, accurate decision-making, and market-driven product launches. This advancement not only overcomes challenges such as inconsistent taxonomies and sparse data but also significantly improves the overall competitiveness of GRIPS Intelligence's clients.

Besides classification based on textual descriptions or product titles, researchers increasingly rely on computer vision techniques based on product images to solve the classification problem (Yu et al. 2017). This paper adopts this approach by conducting separate text- and image-based automated product classification assessments and by exploring a multi-modal approach. Additionally, we aim to provide insights into the effectiveness of these methods within a multi-leveled context. Transitioning to our research questions, we seek to answer the following:

1. *How do text and image modalities individually impact hierarchical product classification?*
2. *Does combining textual and visual features in a multi-modal approach outperform methods that rely solely on one modality in e-commerce product classification?*
3. *Can self- and unsupervised learning techniques be a solution for dealing with label ambiguities and inconsistent product taxonomies?*

Guided by the above research questions, our hypothesis posits that a multi-model approach, integrating both text and image features, will yield superior results in hierarchical product classification compared to individual modalities. The integration of textual and visual information is expected to capture a more comprehensive understanding of product characteristics, leading to better classification accuracy. Furthermore, we hypothesize that integrating self-supervised learning is viable when dealing with untrustworthy labels. Last but not least, we think that the application of unsupervised learning accompanied by sophisticated dimensionality reduction for high-dimensional data can develop and enrich inconsistent product

taxonomies.

The paper unfolds in the following structure: *Chapter 2* briefly overviews relevant literature on product categorization and explores various approaches. In *Chapter 3*, the dataset is introduced, with a thorough exploration of the specific attributes associated with each product. In *Chapter 4,* a comprehensive exploratory data analysis is conducted, revealing relevant patterns and findings from the dataset. *Chapter 5* explains the methodology adopted and provides a detailed account of the models and modalities used. This chapter forms the basis for *Chapter 6,* in which the research findings are systematically presented and discussed. The concluding *Chapter 7* concludes the study with a summary of the key findings, accompanied by a discussion in which all limitations inherent in the research methodology are acknowledged and addressed. The subsequent papers, labeled A through D, delve into distinct aspects of the research. *Paper A* investigates text-based classification and the impact of dynamic masking, followed by an analysis of various image-based methods in *Paper B*. *Paper C* integrates the findings from the preceding papers through a multi-modal approach. Finally, self-supervised learning is employed in *Paper D*, and an innovative concept for hierarchy refinement is introduced.

2 Related works

In this chapter, we conduct a comprehensive review of existing literature in the domain of product categorization, exploring both flat and hierarchical structures to understand the effectiveness of various methodologies. Additionally, we delve into the integration of vision-language models for product classification, examining how the fusion of visual and textual information contributes to advancements in accuracy and efficiency across diverse product domains.

2.1 Flat & Hierarchical Product Categorization

The organization of product catalogs in clearly pre-defined hierarchies or taxonomies is common practice in the field of e-commerce. These structured classifications facilitate product search and navigation. Examples of such taxonomies are the Wordnet hierarchy, Google's product taxonomy, and the Open Directory Project (ODP) (Krishnan and Amarthaluri 2019). The challenge of multi-class classification within these taxonomies has been the subject of research for some time due to the urge for automatic classification. Over the years, various strategies have been researched and implemented using textual as well as visual features (Bergamaschi, Guerra, and Vincini 2002; Hasson et al. 2021; Gupta et al. 2016; Shen et al. 2011; Cevahir and Murakami 2016). Among these, two prominent approaches have emerged as standard procedures: flat single-level classifiers, which classify products at their lowest hierarchy level in a single step, and hierarchical multi-level classifiers, which categorize products through a series of steps that follow the structure of the taxonomy.

Shen et al. (2011) from eBay Research lab proposed using domain-specific feature generation and modeling techniques to improve the classification accuracy of lower levels in a hierarchical taxonomy. Their innovative approach developed features that captured a wide range of rich domain knowledge and linguistic cues. These features were then fed into a Support Vector Machines (SVM) based model to distinguish several confusing category groups to solve the problem of a performance bottleneck of a live system used by eBay at the time.

In their study, Kozareva (2015) from Yahoo! Labs introduced an automatic mechanism for product categorization that assigns correct categories to products based solely on their titles. The mechanism operates within a complex taxonomy of 319 categories structured across six levels. A comprehensive empirical evaluation was conducted using a dataset of 445,408 product titles. Comparing various algorithms, the most effective system achieved a notable f-score of 88%. This indicates high precision in classifying products into specific categories compared to

human labeling.

A year later, Gupta et al. (2016) proposed a new distributional semantics representation for product description vector formation. Their work focused on a two-level ensemble approach that leverages path-wise, node-wise, and depth-wise classifiers corresponding to the taxonomy tree, effectively reducing error in the final product classification task. This approach has demonstrated its effectiveness on datasets from a leading e-commerce platform, achieving improved results over previous methods.

While many approaches to solving the category recognition task in e-commerce use traditional machine learning methods such as SVM, deep learning algorithms have also been applied to solve this task in recent years (Hasson et al. 2021). Building on prior work, Cevahir and Murakami's (2016) work on an automatic classification tool for titles and descriptions of products marks another significant stride in this domain. They combined deep belief nets and deep autoencoders and incorporated a selective reconstruction approach during the training phase. This approach was specifically designed to manage large, sparse feature vectors, a crucial aspect given the scale of their training, which included around 150 million products categorized in a taxonomy tree with up to five levels and 28,338 leaf categories.

Further adding to the comparative analysis in this field, Krishnan & Amarthaluri (2019) explore the efficacy of flat models against hierarchical models. Their research reveals scenarios where flat models exhibit superior performance. They propose two deep learning-based models for extracting features from unstructured product data, creating a unique product signature. Their approach, which elegantly combines structured and unstructured data, proves robust against the challenges of varying attribute orders and categories. They also showed that with a large amount of data, Deep Learning models significantly outperform traditional Machine Learning techniques, where multi-CNNs exhibit the advantage of being parallelizable.

Gao et al. (2020) contributed significantly with their Deep Hierarchical Classification

framework, which uniquely integrates multi-scale hierarchical information into neural networks. This approach, along with a novel loss function, has proven to outperform existing methods in accuracy by effectively penalizing hierarchical prediction errors.

Similarly, Hasson et al. (2021) introduce CatReComm, an interactive, real-time system that provides category recommendations in various e-commerce scenarios. This system, which focuses on listing and search-query category recognition, is pioneering in using a convolutional sequence-to-sequence approach for category recognition, showcasing its utility in an end-to-end scenario.

To summarise, these studies represent the state of the art in product classification research. The advances in this area are varied and robust - from deep hierarchical structures and semantic representations to real-time systems and neural modeling-based tools. Each paper uniquely addresses the intricacies of product categorization and significantly enhances the capabilities of classification systems in e-commerce and related industries.

2.2 Product Classification with Vision-Language Models

While Cevahir & Murakami's (2016) work focused primarily on textual content for product classification, their initial evaluations indicated the potential of image data to improve system performance, despite the challenges posed by the significant processing time required for images. This suggests that the evolution of product classification in e-commerce towards integrating visual components alongside text analysis offers promising opportunities for more accurate categorization, albeit with computational efficiency in mind.

Thus, Kannan et al. (2011) delved into combining text and image signals for product classification, addressing the challenge posed by brief and overlapping textual descriptions across categories. They introduced the Confusion Driven Probabilistic Fusion++ (CDPF++) algorithm, which enhances text classifiers by focusing on image classifiers for confusing

categories and adapts to vocabulary changes using unlabeled data. Their research conducted on datasets from Bing Shopping demonstrated a significant improvement in precision and recall over text-only classifiers, highlighting the efficacy of integrating image data in product categorization. Since then, multiple implementations have utilized text and image information to categorize products more precisely.

While Kannan et al. (2011) improved their text-based product classification by incorporating image data, Kalva, Enembreck, and Koerich (2007) approached the challenge oppositely, enhancing their image classification by integrating contextual textual information. Their methodology highlighted the benefits of integrating independent classifiers for text and images, reinforcing that multi-modal approaches can lead to meaningful improvements in product classification.

In the specific context of the fashion industry, Yu et al. (2017) introduced an approach that utilized textual metadata for detecting the main product in fashion images. Based on a CNN, their best approach learns a joint embedding of object proposals and textual metadata using compact representations of bounding boxes extracted from frozen layers of a pre-trained network. Their method, tested on a large-scale dataset from eight e-commerce sites, outperformed existing baselines and underscored the value of text in enhancing image-based classification in niche e-commerce sectors.

Further expanding on multi-modal integrations, Zahavy et al. (2018) explored a decision-level fusion approach that combined text and image neural network classifiers. Their research on a dataset from Walmart.com confirmed that multi-modal networks outperform single-modality models. In addition, they emphasized the practical applications for e-commerce and pointed out the adaptability of their approach to other modalities such as audio, video, and physical sensors.

The field then witnessed a substantial leap with the SIGIR'20 e-commerce workshop. In this

event, as described by Amoualian et al. (2021), a challenging dataset from Rakuten France, comprising approximately 99,000 products with titles, descriptions, images, and 27 corresponding distinct product categories, was introduced. This dataset formed the basis for various studies that pushed the boundaries of multi-modal product classification. The workshop highlighted the effectiveness of various bi-modal fusion techniques, ranging from simple decision-level late fusion to more complex co-attention methods. Notably, the study by Bi, Wang, and Fan (2020) demonstrated the effectiveness of a multi-modal late fusion approach, which won the challenge with a macro-F1 score of 0.9144, showcasing the practical benefits of such methods in real-world scenarios.

In summary, these studies collectively illustrate the evolving landscape of product classification in e-commerce, with a clear trend towards the fusion of visual and textual information. Each contribution, whether in general e-commerce or specific sectors like fashion, underscores the potential of multi-modal models in enhancing classification accuracy and efficiency.

3 Data

To achieve successful outcomes in Natural Language and image processing, it is imperative to implement preparatory measures, specifically tailored to the dataset being analyzed. These measures are carefully designed to enhance the accuracy and efficacy of the subsequent analytical processes. In the following sections, we will introduce the dataset in use, presenting its characteristics through descriptive and exploratory analyses. Furthermore, we will delineate the key preparatory steps required for a seamless implementation of machine learning models, ensuring that they yield reliable and insightful results.

3.1 Dataset

The startup Grips, a leading global transaction intelligence provider, provided the dataset used for our analysis. Grips tracks sales of more than 40 million products across 60 thousand online retailers, offering global insights into everything from market behavior to the performance of a single product on a given website. The data is stored in a tabular fashion, with the raw dataset consisting of 11 features for 26,258,253 data points. The data points are products scraped from various websites, where each data point entails the features: product URL and title, brand, and domain. Those features were enhanced by a third-party company that scraped additional features like price, GTIN, image URL, and SKU, amongst others (see the complete list of features in *Appendix I*). Additionally, the data was enriched by a product categorization based on the Google Product Taxonomy.

3.2 Data Preparation

To cut down the significant data size, we opted for a subset of the data based on the most coarse hierarchical category level. Among the 21 available categories at Level 1, we filtered for „Apparel & Accessories“ due to its richness and diversity in data. Additionally, this category was chosen because Grips Intelligence urged the application of unsupervised learning techniques to determine additional categories at finer levels, specifically for the footwear category. The choice was also influenced by the company's observation that the „Apparel and Accessories“ category demonstrated the highest levels of predictive performance and accuracy, offering clean, reliable data crucial for the precise training of models. Categorizing fashion products in e-commerce is a recognized challenge for numerous retailers, such as Shopify; some foundational work has already been conducted in this domain, providing a basis for further developing and extending this research area.

To obtain reliable training data, we filtered our dataset specifically for the most prominent

fashion platforms, as they are most likely to have well-maintained and clean product titles and high-quality images that can be leveraged for image classification methods. We chose the largest 40 retailers (including C&A, Bon Prix, and Peek&Cloppenburg, named as relevant by Grips) that were known for their relatively pure „Apparel and Accessories“ assortment (list in *Appendix II*). We downloaded the respective images if available and discarded the datapoints with unidentifiable images. Other preprocessing measures we took at this stage were data cleaning measures, dropping observations with missing data to ensure data quality as well as only keeping features with relevance for our product classification task (dropping „GTIN“, „Currency“, „SKU“, „Offer ID“, „Batch ID“). Finally, to obtain more flexibility in the architecture of our hierarchical classifaction model, the values in the „category“ column were split into each respective hierarchical level, generating four new columns („cat_level_1“, „cat_level_2“, „cat_level_3“, „cat_level_4“). After those preprocessing steps, our final generated data frame includes 271.700 products.

For the objectives of our individual work, we further created more narrowed-down subsets, focusing on individual platforms like Zalando. This approach allowed for a more granular analysis and for an easier test of the models. However, it is important to note that all these subsets went through the same pre-processing pipeline to ensure consistency in our methodology and reliability in our findings.

4 Explorative Data Analysis

To lay the foundation for effective classification tasks and gain a comprehensive understanding of our subset dataset, we began with a comprehensive data exploration and descriptive analysis (EDA) to make informed and accurate modeling decisions. As described previously, we assessed data quality by treating missing values in the classification task-relevant features, including product title, brand, category, image, and URL.

4.1 Hierarchical Analysis

The product categorization used in our dataset is based on Google Taxonomy and is structured in a multi-layered hierarchical fashion. The following overview (see Figure 1) aims to illustrate how the data is organized hierarchically. It is important to note that this diagram only shows a brief sample of the possible categorical paths, and the actual range of categories is much more comprehensive.

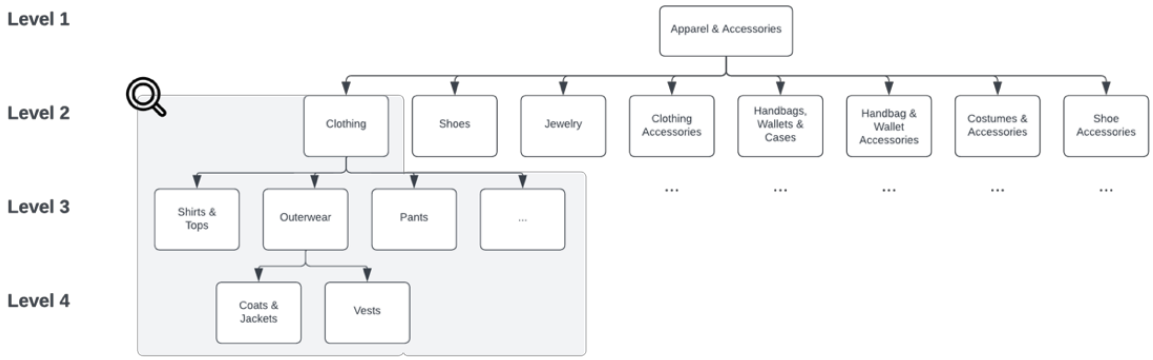


Figure 1: Overview of Google's Taxonomy structure for Apparel & Accessories

At the base of the hierarchical structure is the foundational first-level category, „Apparel & Accessories". This category, already filtered, marks the initial phase of our analysis, encompassing a broad spectrum of fashion-related products and representing the most comprehensive level of categorization in our dataset.

Our analysis reveals that the average hierarchy depth within the „Apparel & Accessories" category is three, with 67.3% of all products being categorized up to that level. The depth ranges from a minimum of two to a maximum of five levels. However, it is noteworthy that less than 7% of all products extend to Level 5, leading us to focus primarily on the first four levels in our analysis. As we delve into the second level of our hierarchical structure, the dataset is divided into eight main categories, including „Clothing", „Shoes", and „Jewelry". These serve as parent categories for further detailed classification. For instance, under the „Clothing" category at the second level, we find third-level subcategories such as „T-shirts and Tops", „Dresses", and

„Pants". However, it is crucial to understand the depth variation among these second-level categories. For example, „Shoes" consistently presents a depth of two, with no further subcategories, highlighting an apparent shortcoming of the Google Product Taxonomy. Similarly, categories like „Handbags", „Wallets & Cases" and „Handbag & Wallet Accessories" have a consistent depth at Level 3. In response to this, an in-depth analysis and potential expansion of the classification system are explored in *Paper D*. In contrast, „Clothing" exhibits a more complex structure, extending up to five levels, with an average depth of 3.19. Furthermore, when examining the top 10 categories with the most distinct combinations after the second level, „Clothing" leads significantly with 53 distinct combinations, followed by „Clothing Accessories" and „Jewelry" (*Appendix II.I*). This indicates a higher level of categorization complexity and variety within these areas. These statistics highlight each category's diverse complexity, shaping our categorization and analysis approach. The detailed findings on analysis categories can be found in *Table 1*.

	Hierarchy Depth Statistics by Cat_level_2 Category		
Category	Average Depth	Maximum Depth	Minimum Depth
Shoes	2,00	2	2
Shoe Accessories	2,81	3	2
Handbag & Wallet Accessories	3,00	3	3
Handbags, Wallets & Cases	3,00	3	3
Jewelry	3,01	4	2
Clothing Accessories	3,03	4	3
Costumes & Accessories	3,05	4	3
Clothing	3,19	5	2

Table 1: Hierarchy Depth Statistics by Category on Level 2

4.2 Categorical Analysis

The following sections present an overview of the distribution and potential categories across all four hierarchical levels of our fashion dataset. Please note that the plots for Level 3

and Level 4 are presented on a logarithmic scale, which is particularly useful given our data's high range and varying absolute numbers. This approach ensures that all categories are effectively visualized and comparable, even with substantial value disparities.

Starting with the second level of our hierarchical structure, we observe distinct patterns in the class distribution, as depicted in *Figure 2*. Here, the absolute number of observations for each category is displayed above the bars, while the proportion of labels within each category is represented along the y-axis.

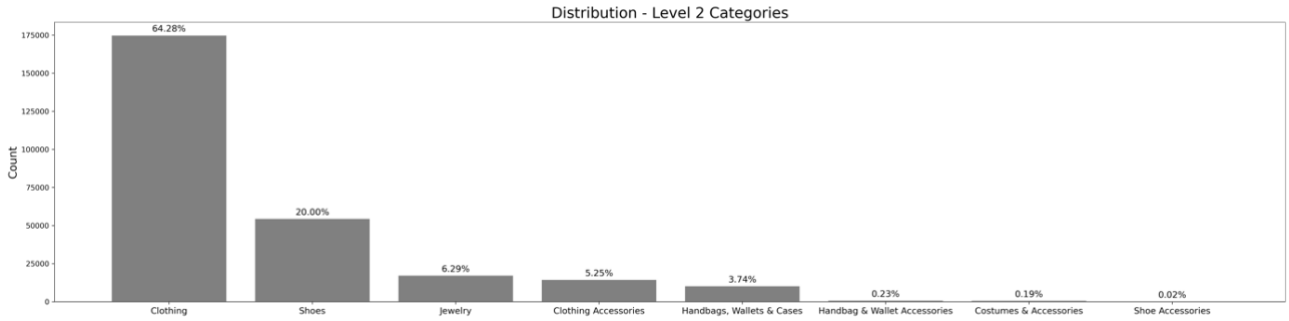


Figure 2: Class Distribution for Categories on Level 2

A detailed examination of this distribution highlights a clear trend. „Shoes“ and „Clothing“ emerge as the most prominent categories, containing the majority of products in our dataset. Combined, these two categories account for over 80% of all products, underscoring their dominance. In strong contrast, the „Shoe Accessories“ category is significantly smaller, encompassing fewer than 50 products. This disparity and the resulting class imbalance is not limited to the second level but also profoundly influences the entire hierarchical structure. The predominance of „Shoes“ and „Clothing“ shapes the overall distribution and categorization patterns, indicating the areas of product concentration and potential under-representation in our dataset.

Advancing from the second to the third level in our dataset's hierarchy introduces a higher degree of granularity, as shown in *Figure 3*. At this stage, most broader second-level parent categories are further subdivided into a multitude of specific subcategories. This detailed

segmentation results in 56 distinct subcategories, offering an in-depth perspective of the product types within each category.

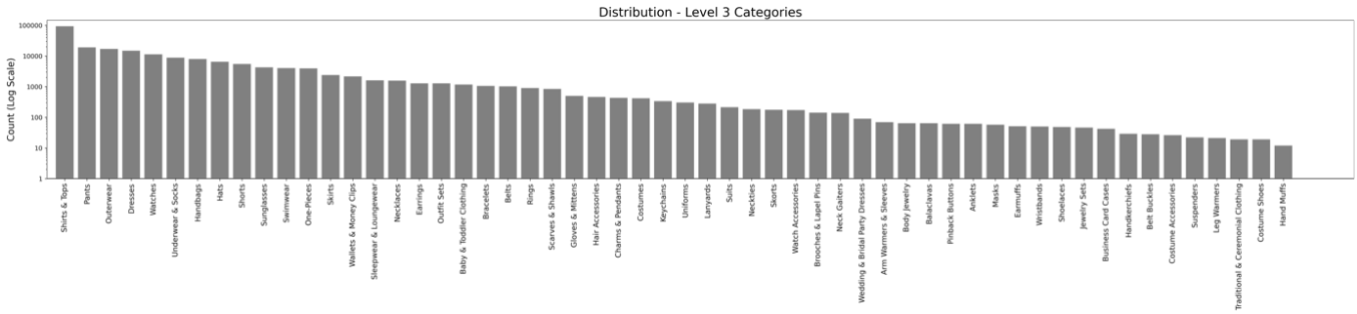


Figure 3: Class distribution for Categories on Level 3 in logaritmic scale

Figure 3 highlights that a considerable number of third-level categories are associated with the second-level „Clothing“ category. Within this, we observe specific subcategories like „Shirts & Tops“, the most prominent, followed by „Pants“ and „Outerwear“. This variety reveals the „Clothing“ segment’s complexity and diversity. The graph also distinctly shows „Shoes“ as a standalone second-level category, lacking further subdivisions into shoe-related subcategories. This distribution at Level 3 underscores the class imbalance stemming from the dominant „Clothing“ category, with its diverse and distinctive subcategories highlighting the uneven distribution in our dataset.

Additionally, in the long-tail segment, we observe several Level 3 categories connected to other Level 2 categories, such as „Watches“, „Handbags“, and „Belts“. While these are less common compared to the clothing-related subcategories, they hold significance within their respective Level 2 groups. As we explore the fourth hierarchy level of our dataset (*Figure 4*), which includes 42 potential subcategories, the data displayed on the y-axis reveals a significant detail: a substantial portion of products do not reach this level of detailed classification. This pattern indicates a trend where the depth of product categorization decreases, with many products not being classified beyond the initial levels.

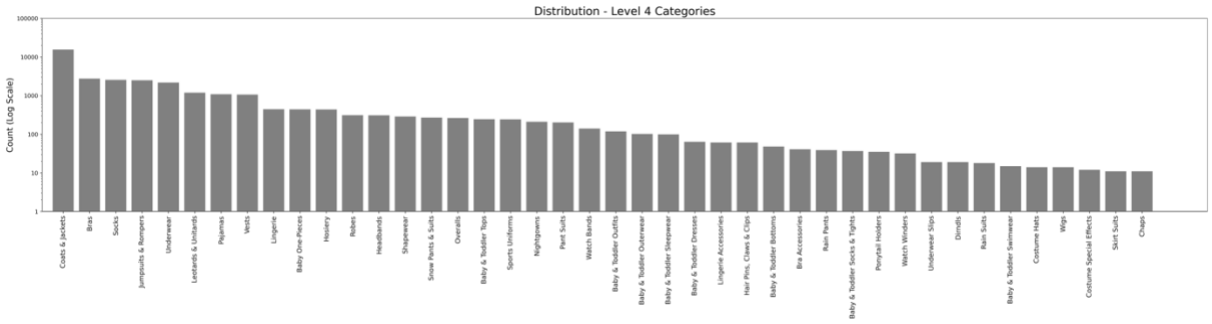


Figure 4: Class distribution for Categories on Level 4 in logaritmic scale

Among the products that do reach this level, many continue to fall under the broader Level 2 „Clothing“ category. Notably, „Coats & Jackets“ emerges as the most observed subcategory at Level 4, aligning with „Outerwear“, one of the top three categories at Level 3. This indicates that while the depth of classification in categories like „Outerwear“ is considerable, the overall trend across the dataset shows a tapering in the classification depth, with fewer products being categorized into the most granular Level 4 subcategories. In summary, the predominance of clothing-related subcategories across all hierarchy levels highlights the depth and detail of their categorization. This trend not only reflects the comprehensive classification of clothing products but also mirrors the class imbalance originating from the Level 2 categories.

4.3 Product Title Analysis

Every product in our dataset is identified by a unique product title, a key component essential for the text-based algorithms that we will utilize throughout this study. The distribution of these product title lengths, illustrated in *Figure 5*, offers valuable insights into their typical semantic structure and composition.

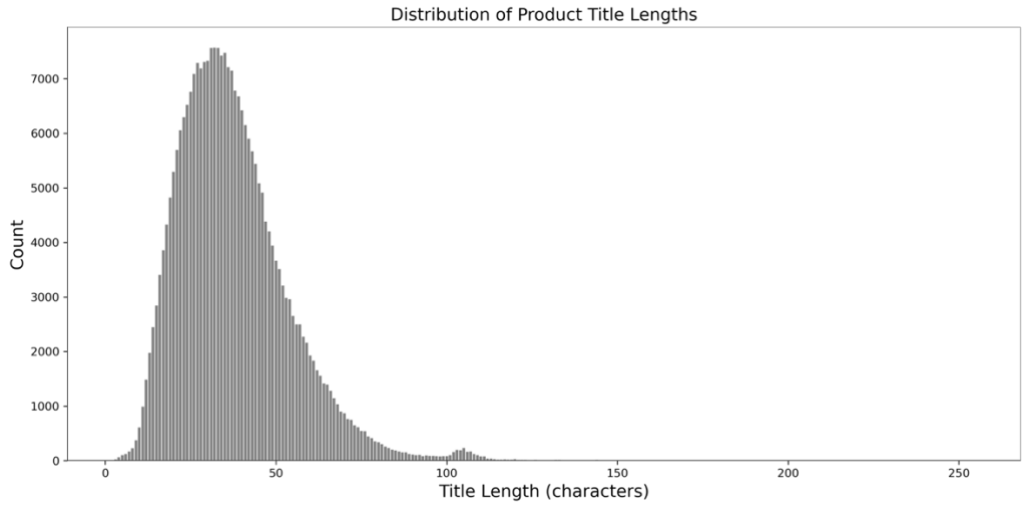


Figure 5: Distribution of Product Title Lengths

The distribution shows the range of lengths and emphasizes the importance of accurate and informative product descriptions. The shortest title consists of only two characters („On“) and suggests a minimalist or highly abbreviated labeling approach. On the other hand, the most extended product title comprises 255 characters („Funny dog owner shirts, funny dog shirt sayings, funny dog shirts for humans, funny dog shirts for humans, dog lover t-shirts, dog shirts with print for women, cute dog shirts for dogs, shirts with dogs on them, how to chill like a dog, funny dog t-shirt for kids“). This extreme verbosity likely results in an overly detailed and potentially superfluous description. Such outliers in title lengths emphasize the importance of accurate and informative product descriptions. A closer look at the graph shows that most product titles are grouped around an average length of 37 characters. This length appears to offer an optimal balance, providing enough detail to convey the essential product information while remaining succinct enough for quick comprehension.

When examining the most frequently used words in the product titles (*Figure 6*), a noteworthy observation is the presence of gender-specific terms like „man“ or „woman“ among the most frequently mentioned words. This suggests that many product titles are tailored to a specific gender demographic. Additionally, we notice that many frequently used words are directly linked to the „Clothing“ category at the second level, with terms such as „dress“,

„jacket“, and „top“ being prominent. This linguistic pattern echoes the earlier mentioned category imbalance, reinforcing the dominance of „Clothing“ in our dataset, as illustrated in *Figure 2*.

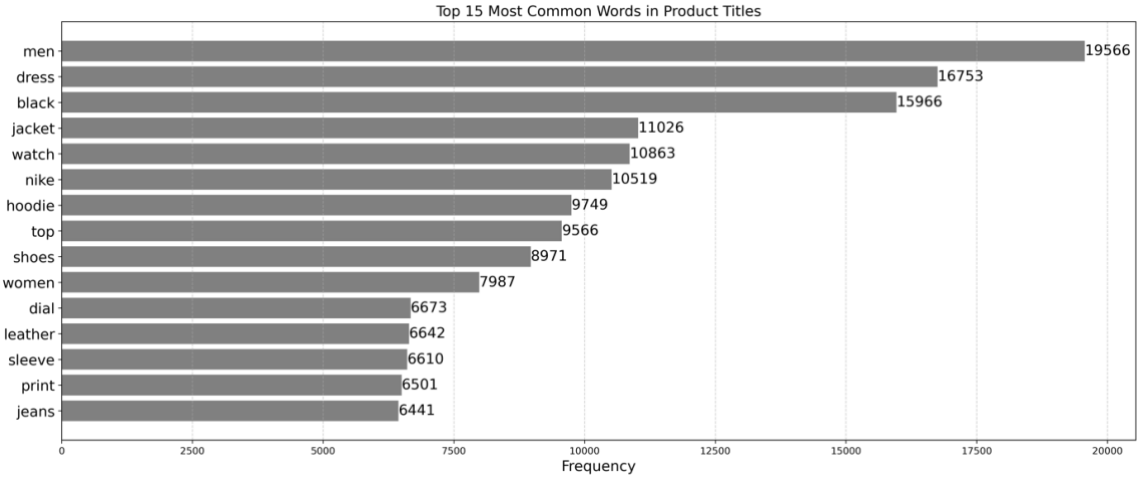


Figure 6: Distribution of the 15 most common words in Product Titles

4.4 Brand Analysis

In addition to our product descriptions, we also have information about the brand of each product. The visual representation in *Figure 7* showcases the top 10 brands in our dataset based on their product offerings, along with the distribution of offered products across Level 2 categories.

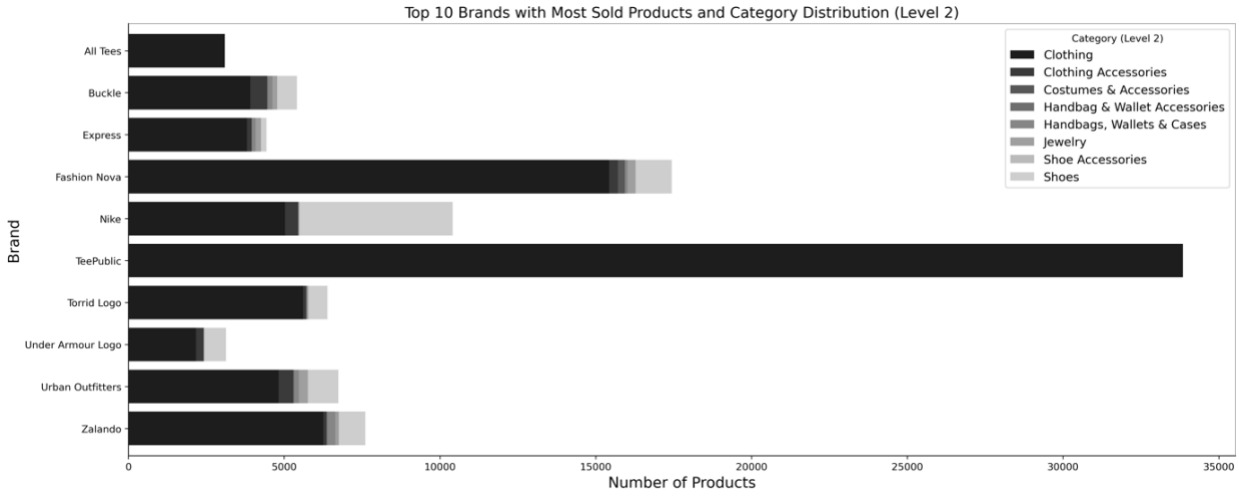


Figure 7: Product Category distribution across the top 10 most frequent Brands

Predominantly, these leading brands are heavily oriented towards clothing. TeePublic emerges as the foremost brand with an exclusive focus on clothing products. Fashion Nova is another significant name, primarily known for its extensive apparel range, although it diversifies with various other product types. Notably, Nike presents a distinct profile in this context, demonstrating an almost balanced product distribution between apparel and shoes. This balanced approach by Nike is an exception in a landscape where brands like TeePublic, Fashion Nova, and All Tees focus predominantly on specific product categories. Such specialization among top brands contributes to the class imbalance observed in the Level 2 category, particularly in the dominance of the „Clothing“ category.

4.5 Image Analysis

In fashion e-commerce, products are presented in various ways, depending on each platform's structure, product assortment, and style. Due to the direct impact of these dissimilarities on embedding generation, our image analysis commences with a focused evaluation of the most prominent visual element: background consistency.

To address this challenge, we deployed a methodical approach where each image underwent a background analysis algorithm designed to discern images with clean, consistent backgrounds from those with varying backdrops. The algorithm applied a Gaussian blur to grayscale versions of images, followed by the application of a binary mask predicated on a specified threshold (Meysenburg 2023). This binary mask is crucial in isolating the product from the background, with the mask threshold set to delineate the product silhouette effectively. Subsequently, the proportion of the image that meets the background criteria, determined by a background threshold parameter, was calculated to ascertain the purity of the background. *Figure 8* shows a subset of clean and unclean background product images.

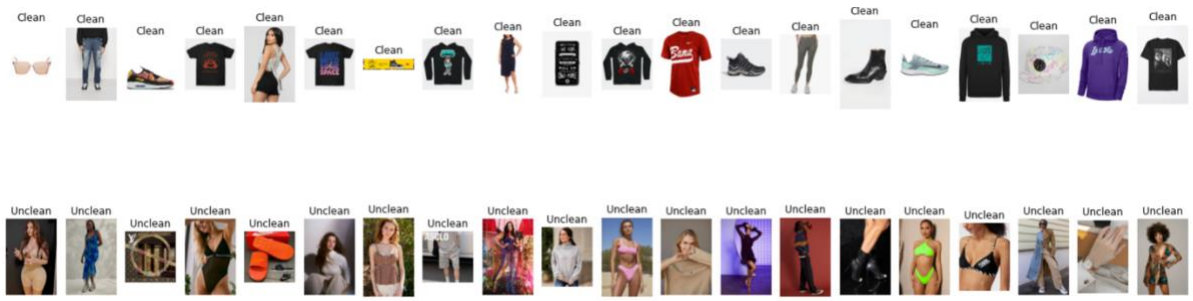


Figure 8: Comparison between pictures with clean and unclean background

In total, the dataset reveals that 97% of images possess clean backgrounds, indicating a high standard of image consistency. This result is primarily due to the fact that the dataset does not encompass marketplaces like Amazon and eBay, where product image variability is typically more pronounced. When broken down by platform, a total of 12 e-commerce sites, including Peek-und-Cloppenburg and Bonprix, exhibited a perfect score, implying an immaculate consistency in their product image presentation. On the lowest end, Fashion Nova is the platform with the highest extent of unclean product images, with a clean background ratio of 84%. Extending the analysis across categories and hierarchy levels reveals further disparities. While the clean background ratio stays consistently above 95% across all hierarchical levels of the taxonomy, categories such as „Lingerie" and „Wristbands" revealed the lowest ratios with 80% and 78%, respectively, indicating that style of product showcasing has a relevant impact on specific fashion segments.

The second challenge related to e-commerce product images emerges from duplicate images, especially when dealing with multinational platforms offering the same product in different markets. Our dataset exhibited an exceedingly low occurrence of duplicate images, as only a singular duplicate instance was identified between two product listings (*Appendix II.II*).

4.6 Platform Analysis

As mentioned in the previous section, the way a product is presented varies significantly from platform to platform. Our dataset, however, deliberately excludes marketplaces with a

high degree of variability in product titles and images, like eBay and Amazon, to maintain a high standard of image and title consistency, thereby ensuring the robustness of subsequent classification tasks. Since Zalando plays a substantial role in our work, this section focuses on Zalando's specific platform characteristics and how they compare to the other 40 platforms in our dataset.

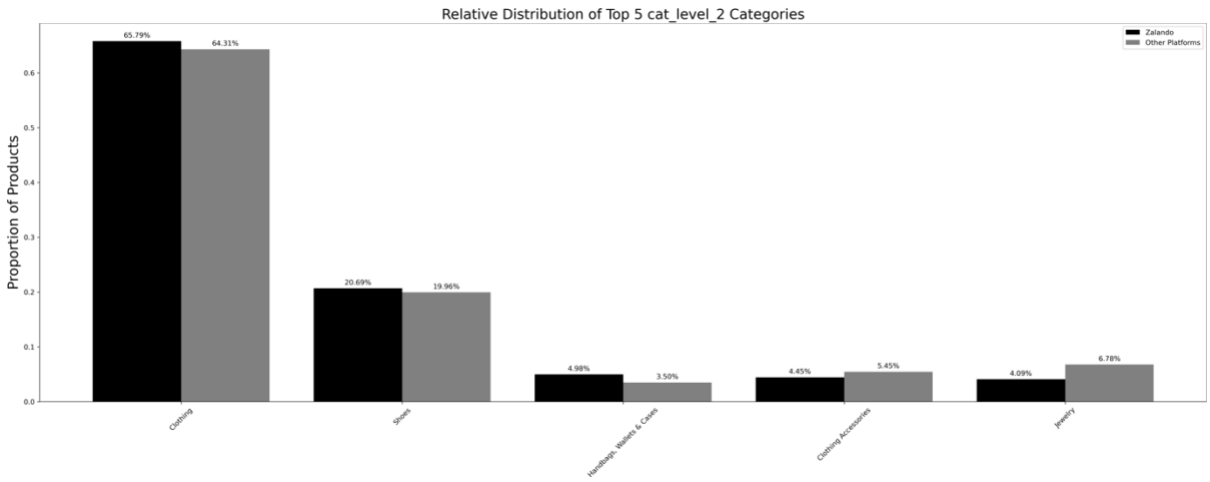


Figure 9: Comparison of Zalando and other platforms on distribution of the top 5 categories on Level 2

Zalando is the most dominant platform in our dataset, contributing to 17.28% of the total records. This share encompasses nearly 47,000 products, a strong differentiator compared to the remaining platforms' average product count of only 5,200. Nevertheless, clear commonalities appear when comparing category distributions across the Google product taxonomy. “Clothing” and “Shoes” are the most prominent categories across all platforms, reflecting a collective emphasis on these essential fashion segments. Both groups are affected by a significant class imbalance since these two categories account for more than 80% of the total records. However, it is noticeable that the other 40 platforms are, in total, slightly more impacted by the imbalance problem, as they together have no other category surpassing a 10% proportion on category Level 3 and 4 (*Appendix II.III*). Furthermore, the depth of product categorization, while slightly more detailed on Zalando, remains substantial across other platforms, with an average hierarchy depth of three for all fashion products for both groups. An

additional similarity emerges in the consistency of image backgrounds across platforms, with an impressive average of over 97% of images showcasing clean backgrounds. The branding landscape presents a convergence of global fashion and sportswear brands, with Nike and Adidas prominently featured across Zalando and other platforms. Nevertheless, platform-specific labels play a role in the case of Zalando and Teebluic, as both brands account for a substantial share of close to 30% of all brands across the dataset (*Appendix II.III*). In terms of product titles, the analysis reveals Zalando's preference for concise descriptions, averaging 30 characters, against a broader average of 39 characters on other platforms. The currency distribution aligns with the geographical focus of each platform. While there are 41 distinct currencies in the dataset, the Euro's prevalence on Zalando (62%) and the US Dollar's dominance across other platforms (80%) reflect the global nature of e-commerce, with major currencies facilitating cross-border transactions and broadening consumer reach (*Appendix II.III*).

4.7 Summary and Insights

The key findings from this exploratory data analysis have important implications for data preprocessing and feature selection in the product classification task. For instance, the analysis shows that most product titles are relatively short in length. This observation carries significant implications, particularly for text-based algorithms used in classification tasks. Some product titles not only exhibit brevity but also contain unusual or minimal text, potentially limiting the available information for classification. These characteristics may pose challenges to the learning process of such models. For a thorough and detailed performance analysis and attempts to enhance each product's interpretability, please refer to *Section A*.

Regarding images, clean and consistent backgrounds help in focusing on the product, enhancing feature extraction, and thereby improving the accuracy of classification and the

effectiveness of clustering. The rarity of duplicate images in the dataset prevents overfitting and ensures a diverse training set, crucial for the generalizability of models. However, the variability in presentation across different platforms and product categories necessitates adaptable algorithms capable of handling these differences, ensuring consistent performance across varied e-commerce environments. For this reason, training the hierarchical models on a cross-platform dataset will be crucial to enable their generalization.

Another critical aspect is how class imbalance in the first hierarchy level propagates to lower-level categories. The imbalance observed at the second level extends to subcategories in subsequent layers. As a result, the class imbalance issue becomes more pronounced in the finer-grained product groups. This phenomenon highlights the need for careful handling of class imbalance at the top level and lower hierarchy levels to ensure fair and accurate analyses across all categories.

Finally, Zalando stands out as the most prominent platform in our dataset, and interestingly, it shares the majority of the main characteristics with our dataset. These characteristics include category distribution, clean background ratio, and hierarchy depth, making it an ideal foundation for an initial proof of concept of our work. We will delve into this approach in greater detail in the following section.

5 Methodology

Our research framework is structured into two main objectives. Firstly, our individual parts aim to establish a proof of concept on a subset of our dataset by focusing on the platform Zalando. As derived from the EDA analysis, the selection of Zalando for the individual contributions of this work is due to its representative characteristics across our dataset, accompanied by the fact that the fashion retailer has a standardized and clean process to define product titles, which allowed us to have higher confidence in label-robustness while training

the algorithms. This first analysis encompasses four approaches: unimodal text-based and image-based classification, multimodal classification using both text and images, and a novel self-supervised learning approach based on images. After training our classification models on the Zalando dataset, we will analyze their capabilities by testing them on Zalando and on the remaining platforms in our dataset. Secondly, we aspire to extend the applicability of our findings across multiple platforms by training and testing the best-performing embeddings and models identified in the initial phase on the complete dataset. *Figure 10* provides an overview of our global structure, described in detail in the subsequent sections of this chapter.

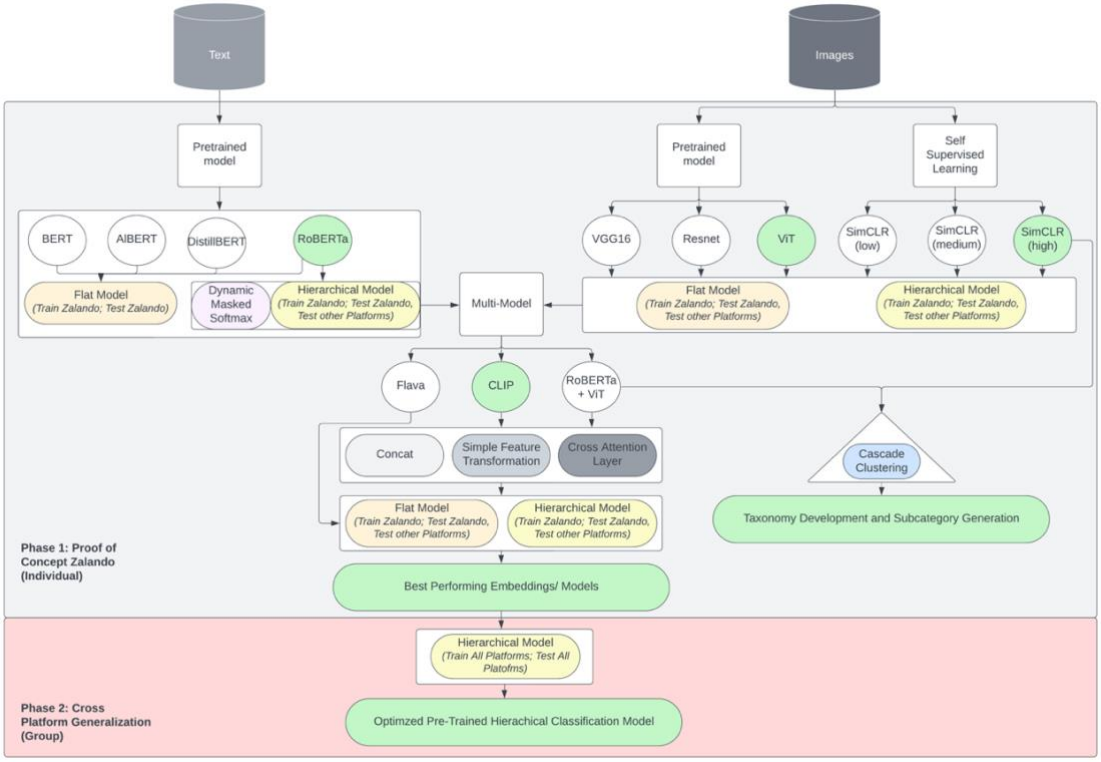


Figure 10: Graphical representation of the Paper's structure and Methodology

5.1 Data Preprocessing

At first, we tackled the issue of inconsistent hierarchy depths within the category column. As highlighted by the EDA, not all products have a classification label for each taxonomy level, resulting in missing values (NaNs) for the deeper hierarchy levels. Such missing data could

introduce significant bias into our classification algorithm. To mitigate this issue, we replaced NaN values with the most specific classification available - the deepest hierarchy level assigned to that product. This approach ensured continuity in the category data, providing a more robust foundation for the subsequent classification tasks.

The second major operation was filtering out infrequent categories. For each categorical column, we removed categories that appear less frequently than a defined threshold that we set to ten. This filtering process ensures that each category within the dataset has a sufficient number of instances, thereby enhancing the robustness of any subsequent statistical or machine-learning analysis.

5.2 Modality Exploration

The four different modalities were analyzed, applied, and tested in the individual papers presented in *Sections A, B, C*, and *D*. The approach of each individual part and key implications for the final analysis of this work are summarized as follows:

In *Paper A*, our exclusive focus is on text data, where we carefully examine and compare several pre-trained models, including BERT, AlBERT, DistillBERT, and RoBERTa. We evaluate these models using a flat and a sophisticated dynamic masked softmax hierarchy classification strategy, ultimately applying the latter classification model to the top-performing BERT variant, RoBERTa. For the shared component, instead of using the tokenized space, we extract RoBERTa CLS embeddings from the pooled output, thereby ensuring comparability with the results from different modalities.

Progressing to *Paper B*, the focus shifts to using the visual input for our product classification. Here, we apply a variety of well-known pre-trained models such as VGG16, ResNet50, and vision transformer models like ViT to generate embeddings and compare their performance in a flat and hierarchical model setup. Our evaluations indicate that ViT

embeddings surpass others in both settings, making it the preferred choice for image embeddings in our final analysis.

Paper C is devoted to exploring a multimodal approach, comparing various embedding generation techniques (RoBERTa and ViT, CLIP, FLAVA) paired with different fusion strategies. The goal is to identify the best-performing combination, comparing its performance with the one of the unimodal approaches. Ultimately, the study shows that CLIP embeddings fused through simple transformation techniques stand out for downstream classification tasks. Consequently, this approach will also be applied to the full dataset.

In a departure from the supervised learning methods of the previous sections, *Paper D* adopts a self-supervised learning framework to generate image embeddings. In assessing three separate SimCLR models, differentiated by their data augmentation intensities, we found that the model with the highest level of augmentation demonstrated superior classification results. Furthermore, this section leverages the assorted text and image embeddings from the other parts to innovate a cascade clustering method that enhances the Google Product Taxonomy by generating novel subcategories within the „Shoes“ division.

5.3 Embeddings Dataframe

After identifying the most successful strategies for each individual approach, we applied them to generate embeddings for the entire cross-platform dataset. To facilitate our future analysis, we have created a singular, cohesive data frame that contains the embeddings generated by ViT, RoBERTa, SimCLR with High Augmentation, and CLIP with Simple Feature Transformation fusion. We chose to separate the creation of embeddings from the classification task to significantly speed up the computation time of the classification process. Thus, once the embeddings were constructed, they could be conveniently fed into any downstream task and tested.

5.4 Model Architecture

To ensure an unbiased comparison between the four different embeddings' performance, we fed all of them individually into the same multilabeled, hierarchical classification neural network. For this purpose, we developed a hierarchical neural network utilizing various fully connected layers, functioning as a classification head. The model contains an initial shared part consisting of a Dense layer, followed by batch normalization and dropout for regularization. It then predicts the most coarse level classes before making subsequent level-by-level class predictions, combining the shared layer's output with previous predictions through additional Dense layers and softmax output layers for each hierarchical level. This architecture allows the model to leverage both general and level-specific features for precise classification across different granularities.

The model is versatile, capable of handling flattened embeddings from diverse data types, including images and text, and supports up to four levels of hierarchical classification. A key feature of our hierarchical model is its flexibility to predict the number of hierarchy levels. While it is designed as a hierarchical classifier, it can also function as a flat classification model depending on the chosen number of levels to predict. For a more detailed presentation of the model, please refer to *Chapter B.3.3*. Based on the cross-platform dataset described above, the model was configured and compiled as outlined in *Chapter B.3.1*.

5.5 Metrics

For evaluating the performance of the various embeddings and models across each hierarchical level, we initially utilize a set of flat layer-wise metrics, and specifically accuracy, precision, recall, and F1 score. When dealing with an imbalanced dataset, where the distribution of classes is uneven, the F1 score becomes particularly valuable. This is because F1, being the harmonic mean of precision and recall, offers a more balanced view of the model's performance,

especially in scenarios where one class dominates over others.

F1 score can be calculated in multiple ways, among others macro and weighted average. F1 macro calculates the metric independently for each class and then takes the average, treating all classes equally regardless of frequency. On the other hand, F1 weighted accounts for class imbalances by weighting the F1 score of each class by its presence in the dataset. To gain a comprehensive understanding of the overall performance of the models in hierarchical classification, we decided to consider both F1 weighted and F1 macro. In this way, we can see how the model performs for both common and uncommon classes. F1 weighted gives us insights into the model's effectiveness in handling the dominant classes, which is crucial for practical applications. At the same time, F1 macro tells us about the model's ability to handle all classes equally, highlighting its effectiveness in dealing with less frequent categories.

According to Silla and Freitas (2010), relying solely on flat classification metrics may not provide sufficient insights into the effectiveness of models on classifying hierarchical data. In response, *HiClass* introduced hierarchical precision (hP), hierarchical recall (hR), and hierarchical F-score (hF) metrics (Miranda, Köhnecke, and Renard 2023). Hierarchical precision measures the accuracy of predicted categories and their parent categories, hierarchical recall assesses the coverage of true categories and their parent categories, and hierarchical F1-score balances precision and recall in the hierarchical context. These metrics are tailored specifically for hierarchical classification scenarios, and we, therefore, include them as an additional approach for a comparative analysis. The definitions for these metrics are as follows:

$$hP = \frac{\sum_i |\alpha_i \cap \beta_i|}{\sum_i |\alpha_i|}$$

Equation 1: Formula for hierarchical precision

$$hR = \frac{\sum_i |\alpha_i \cap \beta_i|}{\sum_i |\beta_i|}$$

Equation 2: Formula for hierarchical recall

$$hF = \frac{2 \times hP \times hR}{hP + hR}$$

Equation 3: Formula for hierarchical F1 Score

where a_i is the set consisting of the most specific classes predicted for example i and all their ancestor classes, while β_i, is the set containing the true most specific classes of test example i and all their ancestors, with summations computed over all test examples.

6 Results

In this chapter, we present the results obtained by comparing the four approaches tested. To grant easier readability, the CLIP model with Feature Tranformation fusion will be addressed as „CLIP".

This comparative analysis not only highlights the relative strengths and weaknesses of the embedding models in a cross-platform context but also illustrates the effectiveness of the different embeddings in improving model performance for different classification tasks.

6.1 Performance Metrics

The analysis of F1 macro scores allows us to understand how the different approaches deal with the classification of unrepresented classes. CLIP consistently outperforms the other models across all hierarchical levels, reaching 93.73% at Level 2 and still performing at Level 4 with 81.89%. RoBERTa follows as the second-best performer, notably at Level 2 with 82.92%, but its performance drops at lower levels. ViT and SimCLR lag behind, with SimCLR showing the lowest scores, barely reaching 24% at Level 4.

		Compared results for Hierarchical Cross-Platform Approach			
		RoBERTa	ViT	CLIP	SimCLR
Accuracy	Level 2	95,77%	96.77%	**99.31%**	93.59%
	Level 3	90.97%	88.24%	**98.25%**	79.22%
	Level 4	90.78%	87.95%	**98.19%**	78.90%
Precision	Level 2	95.77%	96.72%	**99.31%**	93.30%
	Level 3	90.80%	87.76%	**98.24%**	77.66%
	Level 4	90.52%	87.25%	**98.15%**	76.70%
Recall	Level 2	95.75%	96.77%	**99.31%**	93.59%
	Level 3	90.74%	88.24%	**98.25%**	79.22%
	Level 4	90.78%	87.95%	**98.19%**	78.90%
Macro F1	Level 2	82.92%	75.23%	**93.73%**	62.92%
	Level 3	61.03%	50.43%	**88.35%**	30.79%
	Level 4	54.27%	42.65%	**81.89%**	23.87%
Weighted F1	Level 2	95.75%	96.71%	**99.31%**	93.35%
	Level 3	90.74%	87.77%	**98.23%**	77.74%
	Level 4	90.45%	87.39%	**98.15%**	77.11%
Hierarchical Precision		92.39%	91.23%	**98.55%**	83.82%
Hierachical Recall		92.44%	90.62%	**98.48%**	83.71%
Hierarchical weighted F1		92.41%	90.92%	**98.52%**	83.77%

Table 2: Hierarchical Classification results across different Approaches

In terms of weighted F1 scores, which consider the class distribution, all models show better performances compared to F1 macro, with notable improvements on the lower levels. However, CLIP once again outperforms the other approaches, reaching a weighted F1 score of 99.31% on Level 2, 98.23% on Level 3, and 98.15% on Level 4.

When examining the “flat” level-wise metrics, it underscores the superiority of the multi-modal approach, with consistently high values for accuracy, recall, and precision for the CLIP approach. RoBERTa and ViT also exhibit comparable accuracy values at least on the first level, but they fail to transfer this high performance to the other levels. SimCLR, while performing well on the first level, encounters challenges in keeping up with the performance of the other approaches starting from Level 3. The hierarchical precision, recall, and weighted F1 provide a more comprehensive view of the models' ability to predict labels in a multi-level setting. Here, CLIP maintains its dominance with scores exceeding 98% in all three metrics, affirming its robustness in handling varied classes. RoBERTa and ViT show comparable performances, with their combined metrics hovering around the low 90% range. SimCLR, however, trails with scores in the mid-80% range, suggesting limitations in its precision and recall capabilities.

These results indicate that CLIP, with its high scores across all metrics and levels, is particularly effective for hierarchical classification in a cross-platform setting with imbalanced classes. RoBERTa and ViT present viable alternatives but with reduced efficiency, especially at lower hierarchical levels. These models exhibit noteworthy accuracy values, with a particularly striking similarity in performance observed in Layer 2. However, as subsequent layers are considered, ViT experiences a more rapid decline in performance compared to RoBERTa. This decline in ViT's performance at lower hierarchical levels is likely due to the fact that images of products could be similar to each other across different low-level categories, making visual differentiation challenging. RoBERTa, utilizing text data, has an advantage here as textual descriptions often provide clearer distinctions between products.

SimCLR, despite reasonable weighted F1 scores, struggles in terms of macro F1, precision, and recall, indicating a potential shortfall in its ability to balance false positives and false negatives, especially in a class-imbalanced scenario.

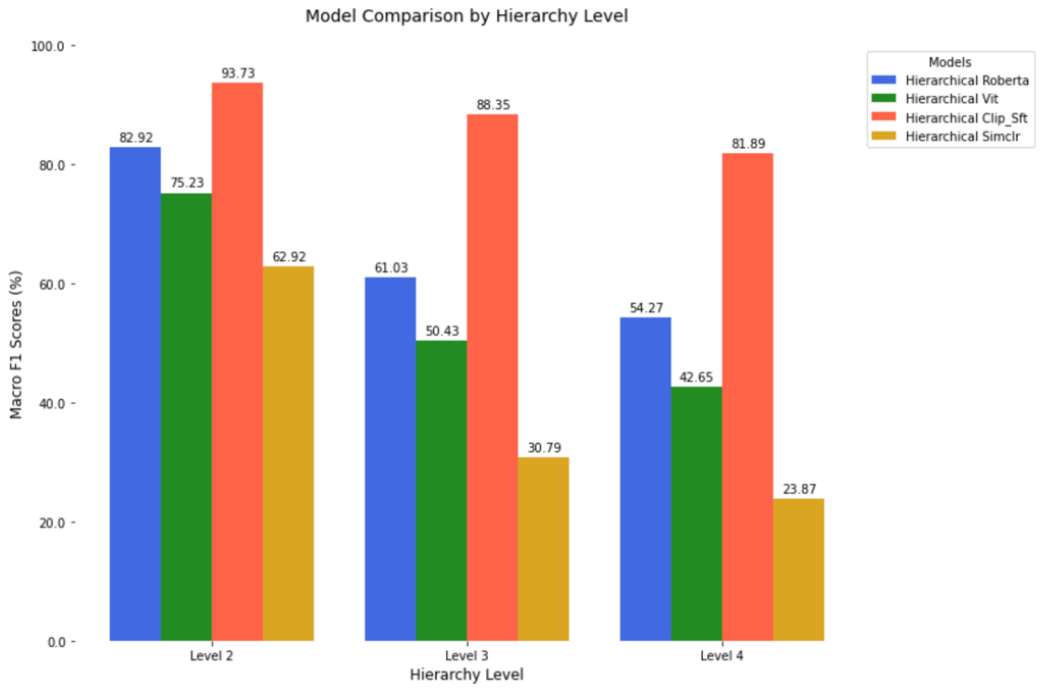


Figure 11: Comparison of Level-specific Macro F1 Scores across different approaches

6.2 Training time and efficiency

Following the comprehensive analysis of performance metrics, we now shift our focus to training specifications, as these are pivotal for real-world applications. In particular, we investigate the distinctive training times and computational efficiency of each model to provide an assessment of their practical feasibility and resource utilization. *Figure 12* provides the total training time and the average time per epoch for each model. It is important to note that we implemented early stopping, which halts the training of each model once the validation loss fails to decrease for five consecutive epochs. This approach is instrumental in optimizing the training process and preventing unnecessary computational expenditures as well as model overfitting.

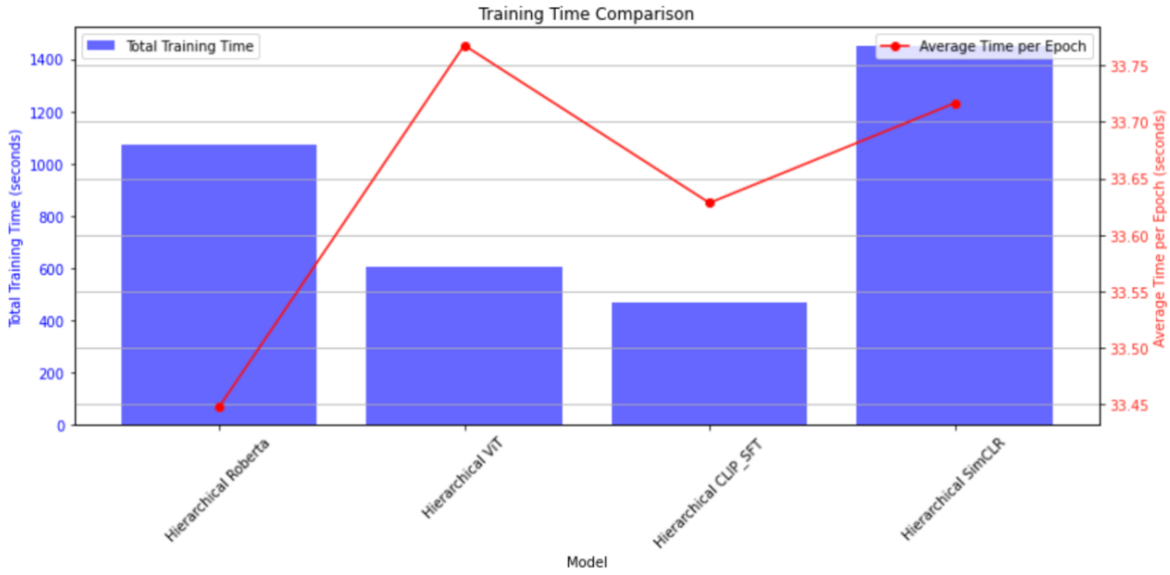


Figure 12: Comparison of Total Training Time and Average Training Time per Epoch across different approaches

In comparing the training times of the four hierarchical models, CLIP emerges as the most time-efficient, converging in just 14 epochs. Conversely, SimCLR requires the most time, necessitating 43 iterations for convergence. All models exhibit an average time per epoch of around 33 seconds, indicating consistent per-cycle efficiency. Overall, the runtime for each model ranges from 12 to 23 minutes.

On the other hand, assessing the time efficiency of embedding generation reveals notable differences. ViT proves highly inefficient, demanding over 52 hours. SimCLR and CLIP both require approximately 5 hours, encompassing embedding generation, fusion, and augmentation. Remarkably, RoBERTa's embeddings are generated in just 25 minutes. Despite not being the fastest, CLIP stands out for striking the optimal balance between results and training times.

6.3 Business Implications

Drawing upon the results where we consistently witnessed robust outcomes across various models, it is evident that the multimodal approach emerges as an optimal strategy in each dimension, rendering it particularly well-suited for real-world scenarios. While each individual model demonstrates commendable performance, the multimodal strategy stands out as the most efficient model, capitalizing on the strengths inherent in both text and image modalities. This integrated methodology holds the promise of a more comprehensive capture of product information, thereby resulting in a noticeable improvement in product classification performance.

7 Conclusion & Limitations

Our investigation, which focused on the hierarchical classification of e-commerce products, explored different modalities that included text, image, and multimodal frameworks. In addition, we explored self-supervised learning methods to determine their effectiveness in generating new taxonomy levels. Optimal models for each modality were identified in individual contributions (*Sections A-D*), which were eventually integrated into a unified framework. This process formed the basis for a comprehensive comparative analysis of the different modalities studied.

Our findings highlight the efficacy of multi-modal approaches, showcasing their superiority

over methods relying solely on a single modality for product classification. The integration of information from both text and image modalities yields enhanced performance, capturing a more comprehensive understanding of the interconnections within the dataset. In particular, fusing the modalities through a neural network proved to be a successful practice to achieve competitive results on our task.

In addition, our exploration of self-supervised learning techniques reveals their advanced capabilities, particularly in the development of new product taxonomies. This approach offers a promising way to refine classification structures without the need for large labeled datasets. The potential for further use and exploration of self-supervised learning methods points to a path towards more sophisticated and adaptive taxonomies that are oriented towards the dynamic nature of product landscapes.

Moving forward, a logical next step involves applying the dynamic masking introduced in *Paper A* to the Multimodal Hierarchical Model introduced in *Paper B*. The masking technique has demonstrated the capability to improve performance across each level, making it an ideal candidate for further exploration and utilization in our study. Furthermore, a future goal would be applying the model not only to the „Apparel and Fashion" category, but to a more inclusive and extensive product dataset. With the needed fine tuning and context adaptation, the Multimodal Approach presented could yield to high performances, supporting the whole e-commerce industry. Finally, a possible next step would be to further dive into the creation of a more detailed taxonomy, not only for the „Shoes" subcategory, but for any needed use case.

In the course of our study, certain limitations impacted the scope and outcomes of our research. Firstly, the fine-tuning phase for all presented models encountered impediments due to limitations in server capacity. The constrained computational resources affected the efficiency of fine-tuning, introducing delays and potential inefficiencies. Secondly, our data preprocessing approach did not address missing values in brand, product title, and image fields;

we opted to remove any records with such deficiencies. This decision may have led to the exclusion of valuable data that could have impacted the models' performance and generalization capabilities. Thirdly, our study did not extend the generalization of models to platforms that deal with products beyond the fashion and apparel domain. This restricts the applicability of our findings to the specific market segment we focused on. Lastly, our classification efforts were limited to the fourth level of the product hierarchy due to a lack of sufficient data records for the fifth level, potentially overlooking finer-grained categorization that could be critical for certain applications. These limitations delineate the boundaries of our study's applicability and suggest areas for future research to enhance the robustness and reach of our models.

PART A – Leveraging dynamic masked softmax and shared hidden layers for hierarchical text-based product classification with BERT

A.1 Introduction

While the term "Natural Language Processing" (NLP) has only recently experienced broad-based adoption in the public domain through the proliferation of ChatGPT, the concept of analyzing human language through computer systems has been of interest to scientists and companies for decades (Chernyavskiy, Ilvovsky, and Nakov 2021). In recent years, significant advances have reshaped the landscape of NLP with the introduction of BERT (Bidirectional Encoder Representations from Transformers), proving to be a turning point in the development of language models (Ozyegen et al. 2022). Unlike previous models that process text unidirectionally, BERT's bidirectional approach enables the simultaneous consideration of preceding and succeeding words (Devlin et al. 2019), enhancing its ability to capture language patterns and relationships. As a result, BERT is particularly potent across a variety of NLP tasks, including text classification, semantic understanding, and contextual language modeling (C. Zhou et al. 2023). This study leverages these unique capabilities of BERT for the underlying hierarchical multi-class classification task. Our approach relies solely on the textual representation of products in the form of their titles and brands to analyze how successfully products are categorized.

The work is structured in two parts: The first part assesses the performance of various BERT-based models using a flat model architecture, considering factors such as performance and computational efficiency. The top-performing model is subsequently used in the second part to construct a hierarchical model, incorporating dynamic masked softmax to optimize performance across different hierarchical levels.

The paper is structured as follows: *Chapter 2* covers the BERT architecture, including fundamentals, variants, and benchmark performance, along with the concept of dynamic

masking. *Chapter 3* outlines data preparation, *Chapter 4* describes and visualizes the model architectures, and *Chapter 5* presents the variant analysis results and dynamic masking effects. The paper concludes in *Chapter 6* with a summary and future research suggestions.

A.2 Background & Related Work

This chapter details the BERT methodology and related variants. Additionally, it introduces the concept of dynamic masked softmax, a key element of our hierarchical model, including a brief review of associated success cases.

A.2.1 Fundamentals

BERT is classified as a pre-trained language model (PTM) and undergoes initial training on extensive unlabeled corpora to establish a strong foundation in natural language understanding (Z. Liu et al. 2021). The pre-training phase exposes BERT to vocabulary from English Wikipedia passages (2,500 million words) and Google's BooksCorpus (800 million words), contributing to BERT's deep language knowledge (Ramprasath et al. 2022).

During pre-training, the model undergoes tasks like masked language modeling (MLM) and next sentence prediction (NSP). MLM enables bidirectional learning from text by masking (hiding) a word in a sentence and forcing BERT to use the words on either side of the covered word bidirectionally to predict the masked word. This unique feature enhances the model's capability to grasp dependencies and relationships across the entire context, fostering a more robust contextual understanding (Wettig et al. 2022). NSP, on the other hand, is used to support BERT learning about relationships between sentences by predicting if a given sentence follows the previous sentence or not (Devlin et al. 2019). BERT's underlying transformer architecture makes it possible to parallelize the extensive pre-training efficiently. A type of neural network architecture, transformers were first introduced in the paper "Attention is All You Need" (Vaswani et al. 2017) and have since become the foundation for various natural language

processing models, including BERT. The critical component of the transformer is the attention mechanism, which signals the model to focus on specific words in a sentence, enabling BERT to consider both preceding and succeeding words simultaneously for each position in a sequence, contributing to its bidirectional processing capability. Generally, a transformer consists of layers, called encoders, which process input data in parallel, thus enhancing the model's ability to capture intricate patterns and relationships within the entire context. Besides providing general-purpose NLP capabilities, pre-trained models like BERT can be tailored to domain-specific tasks through fine-tuning, a process in which additional output layers are added to the pre-trained base model (Church, Chen, and Ma 2021). This dual training process equips BERT with profound language proficiency and contextual understanding, making it the state-of-the-art model for diverse natural language processing tasks (Y. Zhou and Srikumar 2022).

A.2.2 Performance

Pre-trained language models like BERT have been broadly leveraged for recent research on product category classification. For instance, in their comparative analysis, Garrido-Merchán, Gozalo-Brizuela, and González-Carvajal (2023) investigated the performance of BERT in comparison to traditional TF-IDF vocabulary employed in machine learning algorithms. Across four classification scenarios, the study consistently revealed BERT's superior performance over conventional NLP methods. Strikingly, the implementation of BERT was found to be less complex than traditional techniques, underscoring its practicality. The authors emphasized the role of transfer learning in attaining remarkable results, achieving an impressive accuracy of 93.81%, notably outperforming an auto ML-based predictor with an accuracy of 73.99%.

A sizeable share of recent academic literature on product category classification emerged from the highly competitive “Semantic Web Challenge”, in the context of which research teams present key findings and case studies related to NLP. Across the literature published in the context of the Semantic Web Challenge, BERT is referenced as a best-practice model for

product classification tasks. For instance, in a multi-level product classification challenge discussed by Zhang et al. (2020), the predominant trend among top submissions was the adoption of BERT architecture variants, with Zahera and Sherif (2020) from Team DICE introducing ProBERT, a multi-label BERT architecture featuring fully connected neural layers with Sigmoid activations for each classification task. Similarly, the winning team, Rhinobird (Yang et al. 2020), introduced a more sophisticated approach leveraging BERT as the base model. Besides integrating distinct BERT models resulting from harnessing hidden states from the last BERT layers, team Rhinobird proposed a dynamic masked softmax logic that explicitly addressed the dependencies among different category levels. Based on the predicted category, this technique reduced the optimization problem's complexity by filtering out subcategories unrelated to the predicted parent category. Inspired by this work, we introduce dynamic binary masks generated from the predictions of the previous layer.

A.2.3 Variants

Despite its remarkable capabilities, the original BERT model has inherent limitations, which has led to the development of newer variants like ALBERT, DistillBERT, and RoBERTa. These variants address challenges such as computational efficiency, model size, and training speed and offer solutions to improve overall efficiency and effectiveness. BERT variants differ along several model dimensions: parameters represent the weights and biases for predictions and determine a model's complexity. Transformer layers process input data, with more layers enhancing the contextual understanding. Hidden size reflects the dimensionality of internal representations, with higher values capturing more complex relationships but requiring more resources. Attention heads let the model focus on different input parts, improving dependency capture (Acheampong, Nunoo-Mensah, and Chen 2021).

As Devlin et al. (2019) note, originally, two variants of the BERT model were released: the base model with 12 layers, a hidden dimension of 768, and 12 attention heads, and the large

model with 24 layers, a hidden size of 1024, and 16 attention heads.

ALBERT, a lite BERT (Lan et al. 2020), enhances computational efficiency by sharing settings across encoder layers. It separates input and hidden layers, reducing training and inference times while maintaining performance.

Rooted in the assumption that BERT was severely under-trained during pre-training, Y. Liu et al. (2019) introduced RoBERTa (Robustly Optimized BERT Approach), trained with larger batches over more data for a longer time on longer sequences. Further, arguing that next sentence prediction during BERT's pre-training does not significantly improve performance, the authors replaced this step through dynamic masking.

Focused on maximization of computational efficiency rather than model performance, DistilBERT (Distilled BERT), was introduced by Sanh et al. (2019) as a light version of BERT. While DistilBERT displays lower performance on specific performance metrics, it allows for more efficienct training and inference. *Appendix A.I* summarizes the most important attributes of each variant.

A.3 Experimental Set-up

This chapter outlines the preprocessing steps performed to transform the input data into a structured form that aligns with the requirements of the models and gives a comprehensive overview of the fine-tuning approach.

A.3.1 Data Preparation

In this study, we employed a curated dataset, specifically subsampled from the Zalando fashion e-commerce platform by filtering product URLs with the keyword "Zalando". The data comprises 51,837 pre-labeled products organized hierarchically. Our objective is accurate predictions down to the detailed, fourth-level category. Prediction begins from Level 2, encompassing six distinct classes, followed by 38 categories in Level 3 and 60 categories in

Level 4. The data preparation involves merging the *product title* and *brand* name of each item into a single string, to give the models a richer context. The resulting input texts were tokenized, a transformation process whereby the raw input text is transformed into a format compatible with the BERT models. Furthermore, the data preparation phase includes padding and truncation, which addresses the issue of variable input text length. Padding ensures that all input sequences have the same length by adding unique tokens to shorter texts, whereas truncation cuts off the tokens beyond the specified maximum sequence length. By setting the maximum sequence length to *40*, we aim to capture the essential information in the concatenated product titles and brand names while keeping the computational requirements manageable. This choice was informed by the extensive exploratory data analysis conducted in *Chapter 4*. After tokenization, the data was split into training and testing at a ratio of 70:30, where 70% of the data was allocated for training and 30% for testing.

A.3.2 Fine-Tuning

Fine-tuning is the key cornerstone of our model setup, building on the language proficiency attained through the initial pre-training of BERT-based models and infusing further domain-specific language understanding. Due to the thorough pre-training, exhaustive hyperparameter tuning is deemed optional (Sun et al. 2019). To address the class imbalance, we employed focal loss with *alpha* set to 1 and *gamma* to 2, following Lin et al.'s (2017) approach. This enhances sensitivity to minority classes. Hyperparameter tuning, guided by the BERT authors' recommendations, involved fine-tuning learning rate, batch size, and number of epochs:

- Batch Size: Options of 16 or 32
- Learning Rate (AdamW): Choose from 5e-5, 3e-5, or 2e-5
- Number of Epochs: Typically 2, 3, or 4 epochs

While other hyperparameters have been explored in our analysis, this paper concentrates on providing a detailed presentation and explanation of the most significant parameters outlined

above. Detailed information on the chosen parameters for the base models can be found in *Chapter A.4.1.1*, while the ones for the hierarchical model are described in *Chapter A.4.2.1*.

A.4 Models

This chapter explores the two distinct BERT architectures designed for the classification task at hand. The adopted models differ in their complexity but adhere to a consistent overarching logic. The first model serves as the base model and simplifies the classification process by focusing on the last assigned category for each product. It is selected by comparing the performance of the previously described BERT variants. The base model is subsequently used for development of the second model, which introduces a more sophisticated hierarchical structure by adding depth to the classification system. Both model types are compared to highlight key differences in the underlying logic with separate flowcharts for each model architecture facilitating a thorough understanding.

A.4.1 Baseline Model Configuration

Our base configuration simplifies the classification process through a flat model design with a single classifier tailored to the ultimate classification level, i.e., the last assigned category regardless of the classification depth for different products. Mathematically, each label is predicted with a certain probability which can be expressed as

$$p^j = \frac{e^{zj}}{\sum_S k=1\ e^{zk}}$$

Equation A.4: Category possibility (Baseline architecture)

where *p* is the probability of label *j, zj* is the logit corresponding to label *j,* and *S* represents the total number of labels in the last category layer. The variable k serves as a summation index, representing individual labels within the specified range from 1 to *S*, with the resulting summation in the denominator encompassing all logits at this specific level. This expression

highlights that the probability of predicting a particular label is inversely related to the total number of categories in the last category layer.

A detailed model architecture is depicted in *Figure A.13*. The input text, comprising of the product title and brand, undergoes tokenization specific to the chosen model variant. Those tokens are initially mapped to embeddings representing each token's position and semantic information. Next, the transformer layers process the tokens through several layers. The number of layers, hidden dimensions, and attention heads in the layers are variant-dependent and are further detailed in *Appendix A.I*. Thereafter, a single classifier layer tailored to the ultimate classification level that operates on the pooled output is added. This output represents a condensed representation of the input and, therefore, captures the contextual relationship within the text. Following the pooled output, a dropout layer with a rate of 0.1 is applied to prevent overfitting.

For prediction of the final category, two common functions are used: first, the softmax function is utilized to convert the raw output scores into probability distributions, ensuring that the predicted label probabilities sum up to 1, and therefore facilitating the identification of the most probable label. Secondly, the model utilizes the argmax function to select the most probable label based on the logits. Thus, the logits represent the raw outputs indicating the likelihood of each possible label, while the argmax function translates the logits into a concrete prediction.

When switching between the BERT variants, we only focus on adapting the tokenizer and the transformer layer and keep the classification layer and the task-specific components consistent across models.

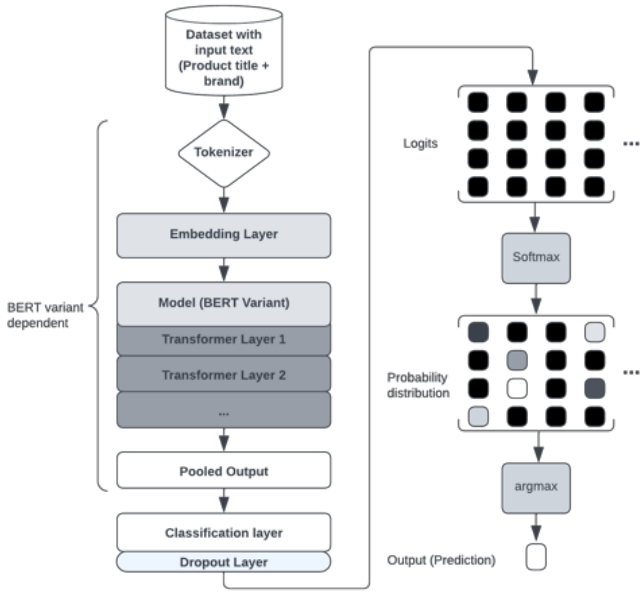


Figure A.13: Baseline model architecture

A.4.1.1 Hyperparameter

To ensure that any observed differences in performance can be attributed to the model architecture or other specific factors rather than variations in hyperparameter choices, we maintained consistent hyperparameters, including a batch size of 32, a training duration of 2 epochs, and a uniform learning rate of 5e-5 with the AdamW optimizer, across all model variants. These parameter choices were determined based on careful observations of training and validation losses as well as overall model performance.

A.4.2 Hierarchical Model Configuration

The second model is designed to capture the hierarchical dependencies within the data. Unlike the flat model, it incorporates multiple layers, each contributing to the prediction process at different hierarchical levels (Level 2, Level 3, and Level 4). This enables the model to learn complex relationships between parent and child categories, potentially enhancing the overall predictive accuracy. We achieve this by leveraging the pre-trained RoBERTa model (identified for its superior performance in the initial phase) and modifying it for hierarchical multi-label

classification by adding three neural output layers – one for each level. These layers correspond to three hierarchical levels in our classification task and function as three independent local classifiers. They are dynamically driven by binary masks generated from the predictions of the previous layer. The following *Equation A.5* represents the probability *p* of a subcategory given its predicted parent category:

$$p^j = \frac{e^{zj * M^{i,j}}}{\sum_S k=1\ e^{zk * M^{i,k}}}$$

Equation A.5: Category possibility (Hierarchical architecture with dynamic masking)

Here, *zj* is the logit for subcategory *j*, and $M^{i,j}$ is the binary mask indicating the relationship between parent category *i* and subcategory *j*. Applying the softmax function to a reduced set of subcategories in layers 3 and 4 elevates the probabilities of relevant subcategories by redistributing the probability mass among them, as the softmax sum remains constant at 1. This adjustment minimizes the impact of irrelevant subcategories and, therefore, potentially enhances the model's predictive accuracy across the deeper layers.

The underlying model architecture can be described as follows: After preprocessing the input text with the RoBERTa tokenizer, the model generates a pooled output, serving as input for the initial layer 2 classifier which in turn produces the initial predictions. As the model progresses through the subsequent layers, it integrates information from the pooled output with the logits from the preceding layer. This integration is enabed by dedicated hidden layers for each level. For instance, in Level 3, the first hidden layer integrates Level 2 information by concatenating the pooled output and Level 2 logits. This concatenated result undergoes a non-linear transformation using the ReLU (Rectified Linear Unit) activation function. Including two additional hidden layers between Level 2 and Level 3, and between Levels 3 and 4, establishes a more expressive feature hierarchy. The shared output size for the layers is set to 512, balancing computational efficiency with the need for complex, hierarchical feature representation.

The decision-making process spans across the levels, incorporating dynamic binary masks

in Level 3 and 4. These masks are applied based on a customized mapping that determines allowable categories, guided by the predictions of the previous level. This strategy identifies logically admissible subcategories, ensuring consistent application of logical constraints within the category taxonomy at all levels of the model. Batch normalization is implemented after the hidden layers for Levels 3 and 4, as well as after obtaining the logits for each level. This enhances model stability during training. Additionally, a dropout rate of 0.1 is between and after the hidden layers for each hierarchical level to mitigate the risk of overfitting. *Figure A.14* illustrates the relationships between the fully connected linear layers, emphasizing the central role of different output layers in the decision-making process. For a detailed overview of the tokenization process and the transformer layers, please refer to *Figure A.13.*

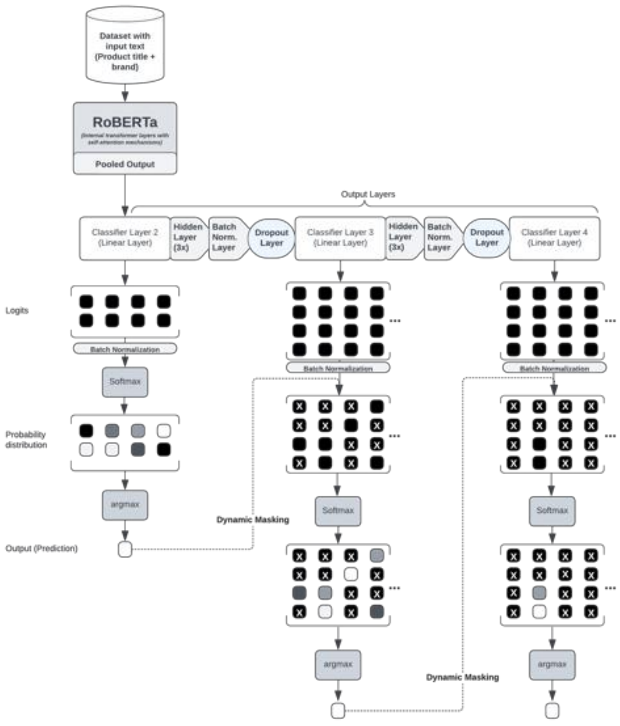


Figure A.14: Hierarchical model architecture with dynamic masking

A.4.2.1 Hyperparameter

While the first layer demonstrates high training performance early on, the knowledge transfer to subsequent layers only becomes evident after 3 to 4 epochs. Therefore, we extend the hierarchical model's training duration to 4 epochs in contrast to the flat model, which

achieves satisfactory results with a 2-epoch training regimen to allow for a more thorough dissemination of knowledge across the hierarchical layers. Our training strategy involves utilizing the AdamW optimizer and a batch size of 32. The BERT parameters, fundamental to the model's architecture, are trained at a higher learning rate of 3e-5, facilitating faster convergence and initial adaptation of the broader contextual embeddings. Conversely, the local classifiers and the hidden layers, responsible for learning hierarchical patterns, are trained at a lower learning rate of 2e-5, allowing for a more detailed and fine-grained adaptation to the hierarchical structures.

A.5 Numeric Results

In this chapter, we present the results of our study in two sections: The first section presents the performance of different BERT variants using the flat model architecture. The second section provides an overview of hierarchical model performance, focusing first on the Zalando test set and examining cross-platform performance thereafter.

A.5.1 Variant Analysis

The performance differences across BERT variants are shown in *Table A.3* and can be attributed to the distinct architectural designs of each model.

Overview results variant analysis - Baseline model configuration

	BERT	ALBERT	DistilBERT	RoBERTa
Accuracy	92,89%	94,28%	93,60%	95,10%
Precision	92,29%	92,42%	93,09%	94,78%
Recall	92,89%	94,28%	93,60%	95,10%
Weighted-F1	92,37%	93,27%	93,22%	94,66%
Training Time (in minutes; for 4 epochs)	~47 mins	~39 mins	~26 mins	~32 mins

Table A.3: Results of BERT variant analysis

As observed in *Table A.3,* BERT has the longest training times and comparatively lower performance, highlighting the need for more efficient variants in resource-intensive

applications. In contrast, ALBERT balances performance and efficiency through shared layer settings, showcasing the importance of optimization in model design. While DistilBERT's reduced parameter size results in slight performance trade-offs on key evaluation metrics compared to ALBERT and RoBERTa, it also drives notably greater training efficiency, making it highly useful in scenarios with limited training and inference budgets and for applications requiring quick decision-making. RoBERTa displays superior accuracy, resulting from its emphasis on prolonged training with augmented data. Incorporating a wide array of text types, such as news articles and open web text, RoBERTa's training regimen exposes the model to a richer variety of linguistic patterns and contexts (Y. Liu et al. 2019). As product titles and brand names often exhibit diverse language styles, abbreviations, and variations not fully captured by models trained on narrower datasets, this ability to understand and generalize from a broader range of language nuances makes RoBERTa uniquely adept to product classification. As a result, RoBERTa shows the strongest performance across variants, while achieving a remarkable reduction of training time by roughly 30% compared to the original BERT, making it the preferred variant for integration in our hierarchical model.

A.5.2 Impact of Dynamic Masked Softmax

The second part of the study evaluates to which extent the hierarchical structure enhances the model's ability to understand and process complex relationships within the given data. The hierarchical model's results, detailed in *Table A.4*, encompass flat and hierarchical performance metrics. For an in-depth description of metrics, please refer to *Chapter 5.5*.

Results of hierarchical model with dynamic masked softmax (Zalando test set)

	Layer 2	Layer 3	Layer 4		
Accuracy	95,68%	93,42%	90,07%	Hierarchical Recall	90,07%
Precision	96,45%	94,03%	86,41%	Hierarchical Precision	90,07%
Recall	95,68%	93,42%	90,07%	Hierarchical F1	90,07%
Weighted-F1	95,85%	92,81%	87,51%		
Average Accuracy		93,06%			

Table A.4: Zalando test results hierarchical model

With an impressive accuracy of nearly 96% in layer 2, the applied RoBERTA model demonstrates remarkable performance. In addition, the modest reduction in accuracy between layer 2 and 3 (~3 percentage points) and layer 2 and 4 (~5 percentage points) provides evidence for the efficiacy of the employed dynamic masking mechanism, which is further corroborated by the coherence between the hierarchical metrics and flat performance metrics in Level 4, emphasizing the ability of the masking mechanism to adeptly handle invalid categories.

The category-level model performance overview comparing predicted with actual categories (*Appendix A.II*) highlights the impact of error propagation, with inaccurate layer 2 predictions cascading down into subsequent layers. The application of the Focal Loss function, designed to address the class imbalance, coupled with RoBERTA's pre-trained capabilities and the innovative model architecture, appears to be highly effective in significantly mitigating this issue.

However, the introduction of increased model architecture complexity, driven by dynamic masking and the addition of extra hidden layers, results in training times longer than those observed for flat models. Spanning 4 epochs and all three levels, the training process took approximately 2.5 hours, underscoring the inherent trade-off between computational cost and improved interpretability, enabled by heightened model sophistication (*Appendix A.III*).

Following the evaluation of the hierarchical model on the Zalando test set, our analysis extends to examine model cross-platform performance, as summarized in *Table A.5*.

Results of hierarchical model with dynamic masked softmax (Cross-Platform test set)

	Layer 2	Layer 3	Layer 4		
Accuracy	69,00%	66,28%	65,81%	Hierarchical Recall	65,81%
Precision	82,95%	78,35%	76,18%	Hierarchical Precision	65,81%
Recall	69,00%	66,28%	65,81%	Hierarchical F1	65,81%
Weighted-F1	71,09%	67,28%	66,31%		
Average Accuracy		67,03%			

Table A.5: Cross-Platform test results hierarchical model

In cross-platform inference, a performance drop is evident in layer 2, with an accuracy of 69%, affecting all subsequent layers. While the misclassifications originating in the first classification layer propagate to the lower layers, the dynamic nature of the applied masking logic proves efficient, reflected in minimal losses after the first classification level. *Appendix A.IV* provides a detailed overview of predicted and actual labels, revealing numerous misclassifications in Level 2.

To fully comprehend these misclassifications, we examined specific cases, such as the product named “Hot Stuff Hot Stuff”, incorrectly labeled as “Clothing Accessories” instead of “Clothing”. This misclassification can be attributed to the fact that Zalando probably maintains similar descriptions for all of its products. However, when the model is applied to other fashion platforms with significantly different product description styles, classification issues arise. Additionally, it becomes evident that the available information for this product does not provide clear indications of the expected category, making it challenging for the model to leverage its pre-trained capabilities effectively. This emphasizes the importance of comprehensive and informative textual content for accurate predictions, especially in cross-platform scenarios.

In summary, the incorporation of dynamic masking presents both advantages and challenges. On the one hand, dynamic masking optimizes the classification task by selectively masking sub-levels unrelated to the predicted parent category, potentially enhancing performance, especially in scenarios characterized by a multitude of possible categories. Further, additional layers contribute to the model's capacity to capture and transfer intricate hierarchical relationships within the data, enabling improved knowledge transfer capabilities and driving a heightened ability to navigate and comprehend complex hierarchical structures within the data. However, this heightened performance also comes with trade-offs such as longer training times. In addition, the approach introduces the risk of compounding misclassifications, especially when parent category predictions are inaccurate, highlighting the

importance of model design to ensure optimal performance across all hierarchical levels as model complexity increases.

A.6 Conclusion & Limitations

Our research demonstrates the consistently strong performance of BERT-based models across various configurations. In navigating the trade-offs between the advantages and challenges introduced by dynamic masking and additional hidden layers, careful consideration of the specific characteristics of underlying data, the complexity of hierarchical relationships, and the computational resources available are paramount. Our research, therefore contributes to the growing body of evidence supporting the efficacy of leveraging pre-trained language models in combination with innovative design modifications. Potential future research questions include several directions to further improve the models' effectiveness and applicability. For instance, future research could explore incorporating more product-related details, like descriptions, to deepen the models' understanding. This is particularly relevant, as observed during cross-platform inference, where the model's performance is contingent on the availability and quality of textual content. Furthermore, conducting a detailed analysis of the misclassified products at each level would help identify the challenges for the hierarchical model, providing valuable information for refining and improving its predictive capabilities. While the present study has focused on models that produced satisfactory results, using multilingual BERT variants remains an exciting potential avenue for future research, opening up possibilities for broader applications and linguistic diversity in hierarchical classification. Finally, due to server dependencies during model training, a comprehensive tuning of the hyperparameters was only possible to a limited extent. Given increased server capacity, there could have been a more extensive exploration of optimal model configurations through comprehensive hyperparameter tuning and thorough re-evaluion.

References

Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen. 2021. "Transformer Models for Text-Based Emotion Detection: A Review of BERT-Based Approaches." *Artificial Intelligence Review* 54 (8): 5789–5829. https://doi.org/10.1007/s10462-021-09958-2.

Cevahir, Ali, and Koji Murakami. 2016. "Large-scale multi-class and hierarchical product categorization for an e-commerce giant." *International Conference on Computational Linguistics*, December, 525–35. https://aclanthology.org/C16-1051.

Chen, Lei, Houwei Chou, Yandi Xia, and Hirokazu Miyake. 2021. "Multimodal Item Categorization Fully Based on Transformer." Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), January, 111–15. https://doi.org/10.18653/v1/2021.ecnlp-1.13.

Chernyavskiy, Anton, Dmitry Ilvovsky, and Preslav Nakov. 2021. "Transformers: 'The End of History' for NLP?" *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2105.00813

Church, Kenneth, Zeyu Chen, and Yanjun Ma. 2021. "Emerging Trends: A Gentle Introduction to Fine-Tuning." *Natural Language Engineering* 27 (6). https://doi.org/10.1017/S1351324921000322

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.1810.04805

Gao, Dehong, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Wei Yi, Yi Hu, and Hao Wang. 2020. "FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval." *Proceedings of the 43rd International ACM SIGIR Conference on Research*

and Development in Information Retrieval, July, 2251–60. https://doi.org/10.1145/3397271.3401430.

Gao, Dehong, Wenjing Yang, Huiling Zhou, Yi Wei, Hu, and Hao Wang. 2020. "Deep Hierarchical Classification for Category Prediction in E-commerce System." *arXiv Preprint*, January. https://doi.org/10.18653/v1/2020.ecnlp-1.10.

Garrido-Merchán, Eduardo C., Roberto Gozalo-Brizuela, and Santiago González-Carvajal. 2023. "Comparing BERT against Traditional Machine Learning Models in Text Classification." *Journal of Computational and Cognitive Engineering*, 2(4). https://doi.org/10.47852/bonviewJCCE3202838

Gupta, Vivek, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. 2016. "Product Classification in E-Commerce using Distributional Semantics." *International Conference on Computational Linguistics*, June, 536–46. https://doi.org/10.48550/arXiv.1606.06083

Kalva, P.R., Fabrició Enembreck, and Alessandro L. Koerich. 2007. "WEB image classification based on the fusion of image and text classifiers." *Proceedings of the International Conference on Document Analysis and Recognition*, September. https://doi.org/10.1109/icdar.2007.4378772.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations." *arXiv (Cornell University),* September. https://doi.org/10.48550/arxiv.1909.11942.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv (Cornell University)*, July. https://doi.org/10.48550/arxiv.1907.11692.

Liu, Zhuang, Wayne Lin, Ya Shi, and Jun Zhao. 2021. “A Robustly Optimized BERT Pre-Training Approach with Post-Training.” In *Lecture Notes in Computer Science*, 471–84. https://doi.org/10.1007/978-3-030-84186-7_31.

Miranda, Fábio, Niklas Köhnecke, and Bernhard Y. Renard. 2023. “HiClass: A Python Library for Local Hierarchical Classification Compatible with Scikit-Learn.” *arXiv (Cornell University),* December. https://doi.org/10.48550/arxiv.2112.06560.

Ozyegen, Ozan, Hadi Jahanshahi, Mücahit Çevik, Beste Bulut, Deniz Yigit, Fahrettin F. Gonen, and Ayşe Başar. 2022. “Classifying Multi-Level Product Categories Using Dynamic Masking and Transformer Models.” *Journal of Data, Information and Management* 4 (1): 71–85. https://doi.org/10.1007/s42488-022-00066-6.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” *arXiv (Cornell University)*, October. https://doi.org/10.48550/arXiv.1910.01108.

Shen, Dan, Jean David Ruvini, Manas Somaiya, and Neel Sundaresan. 2011. “Item categorization in the e-commerce domain.” *In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*, October, 1921–24. https://doi.org/10.1145/2063576.2063855.

Silla, Carlos N., and Alex A. Freitas. 2010. “A Survey of Hierarchical Classification across Different Application Domains.” *Data Mining and Knowledge Discovery* 22 (1–2): 31–72. https://doi.org/10.1007/s10618-010-0175-9.

Sun, Chi, Xipeng Qiu, Yige Xu and Xuanjing Huang. 2019a. “How to Fine-Tune BERT for Text Classification?” In *Lecture Notes in Computer Science*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16.

Tagliabue, Jacopo, Ciro Greco, Jean-Francis Roy, Binqing Yu, Patrick John Chia, Federico Bianchi and Giovanni Cassani. 2021. “SIGIR 2021 E-Commerce Workshop Data Challenge.” https://arxiv.org/abs/2104.09423.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is All you Need.” *arXiv (Cornell University)* 30 (June) : 5998–6008. https://arxiv.org/pdf/1706.03762v5.

Wettig, Alexander, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. “Should You Mask 15% in Masked Language Modeling?” arXiv (Cornell University), February. 10.18653/v1/2023.eacl-main.217.

Yang, Li, E. Shijia, Shiyao Xu and Yang Xiang. 2020. “Bert with Dynamic Masked Softmax and Pseudo Labeling for Hierarchical Product Classification.” *MWPD@ISWC* (2020). https://ceur-ws.org/Vol-2720/paper6.pdf.

Yu, Longlong, Edgar Simo-Serra, Francesc Moreno-Noguer, and Antonio Bandera. 2017. “Multi-modal Embedding for Main Product Detection in Fashion.” *Proceedings of the IEEE International Conference on Computer Vision Workshops*, October, 2236–42. https://doi.org/10.1109/iccvw.2017.261.

Zahavy, Tom, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor. 2018. “Is a picture worth a thousand words? A deep Multi-Modal architecture for product classification in E-Commerce.” *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). https://doi.org/10.1609/aaai.v32i1.11419.

Zhou, Chuangbing, Qian Li, Chen Li, Yuanyuan Wang, Yixin Liu, Guangjing Wang, Kai Zhang, et al. 2023. “A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT.” *arXiv (Cornell University)*, February. https://doi.org/10.48550/arxiv.2302.09419

Zhou, Yichu, and Vivek Srikumar. 2022. “A Closer Look at How Fine-Tuning Changes BERT.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), January. https://doi.org/10.18653/v1/2022.acl-long.75

List of Figures

List of Tables

List of Equations

Appendix

I. Data Dictionary

Variable	Description	Data Source
Product URL	The url of the product which works as a unique identifier	Grips
Gtin	A standardized international product identification number. This is unique for a given product across different retailers	Grips
Product Title	The title of the product as denoted in the metadata of the url	Grips
Brand	The brand of the product	Grips
Currency	The currency of the price	Grips
Price	The price of a product on a given website	Grips
SKU	A retailer dependent internal ID	Grips
Offer Id	A process ID from Grips	Grips
Batch Id	A process ID from Grips	Grips
Category	An assigned category following the google product taxonomy, assigned by a third-party provider	Grips
cat_level_1	First level category, based on Category column	Self-engineered
cat_level_2	Second level category, based on Category column	Self-engineered
cat_level_3	Third level category, based on Category column	Self-engineered
cat_level_4	Fourth level category, based on Category column	Self-engineered
cat_level_5	Fifth level category, based on Category column	Self-engineered
last_cat_level	Last level category, based on Category column	Self-engineered
hierarchy_depth	Hierarchical depth to which the procut is categorized to, based on Category column	Self-engineered
platfom	E-commerce platform on which the product is offered, based on Product URL	Self-engineered
title_length	Count of characters in the product title, based on Product Title	Self-engineered
CleanBackground	Based on an algorithm that applies Gaussian blur and a binary mask to distinguish between products with clean and varying backgrounds, calculating background purity based on a set threshold (bool)	Self-engineered

II. EDA

II.I Hierarchical Analysis

Hierarchy Depth Statistics by Cat_level_2 Category

Category	Average Depth	Maximum Depth	Minimum Depth
Shoes	2.00	2	2
Shoe Accessories	2.81	3	2
Handbag & Wallet Accessories	3.00	3	3
Handbags, Wallets & Cases	3.00	3	3
Jewelry	3.01	4	2
Clothing Accessories	3.03	4	3
Costumes & Accessories	3.05	4	3
Clothing	3.19	5	2

Disitinct Hierarchical Combinations after Category

Category	Count
Clothing	53.00
Costumes & Accessories	22.00
Jewelry	13.00
Costumes & Accessories	5.00
Handbags, Wallets & Cases	3.00
Handbag & Wallet Accessories	2.00
Shoe Accessories	2.00
Shoes	1.00

II.II Image Analysis

Average CBR per Category Level

Category Level	CBR
cat_level_1	97.01%
cat_level_2	95.80%
cat_level_3	95.60%
cat_level_4	95.71%

Category with lowest CBR per Level

Category Level	Category Name	CBR
cat_level_2	Costumes & Accessories	83.65%
cat_level_3	Wristbands	80.00%
cat_level_4	Lingerie	77.58%

Top 5 Platform with highest CBR

Platform	CBR
asics	100.00%
bonprix	100.00%
yoox	100.00%
teepublic	100.00%
peek-und-cloppenburg	100.00%

Top 5 Platform with lowest CBR

Platform	CBR
torrid	94.34%
buyma	90.78%
pacsun	86.95%
urbanoutfitters	85.40%
fashionnova	83.98%

Number of Duplicate Image URLs

Platform	Count
aboutyou	1

II.III Platform Analysis

Relative Category Distribution (Level 3)		
	Zalando	Other Platforms
Shirts & Tops	27.97%	41.55%
Shoes	23.43%	22.62%
Outerwear	12.18%	6.05%
Pants	11.66%	7.21%
Dresses	6.67%	6.13%
Underwear & Socks	6.15%	3.16%
Handbags	4.77%	3.04%
Shorts	2.75%	2.18%
Hats	2.70%	2.71%
Watches	1.72%	5.36%

Top Brands Distribution		
	Zalando	Other Platforms
Zalando	15.05%	15.06%
Nike Performance	2.58%	7.76%
adidas Performance	2.31%	4.63%
Puma	2.08%	3.00%
Nike Sportswear	1.92%	2.84%
Tommy Hilfiger	1.47%	2.41%
adidas Originals	1.27%	1.97%
Classic Outfits	0.99%	1.39%
The North Face	0.94%	1.38%
Guess	0.92%	1.36%

Title Lenght		
	Zalando	Other Platforms
Mean	30.29	39.14
Median	29.00	37.00
Min	3	2
Max	132	255

Hierarchy Depth		
	Zalando	Other Platforms
Mean	2.98	2.91
Median	3.00	3.00
Min	2	2
Max	5	5

Product Count and CBR				
	Zalando	Other Platforms Avg	Other Platf	Other Platforms Max
Product Count	46952	5227	1	39010
CBR	0.97	0.98	0.84	1.00

Top Currencies Zalando	
	Share
EUR	62.42%
GBP	37.10%
PLN	0.23%
CHF	0.15%
DKK	0.04%
CZK	0.02%
NOK	0.02%
SEK	0.01%
HRK	0.00%

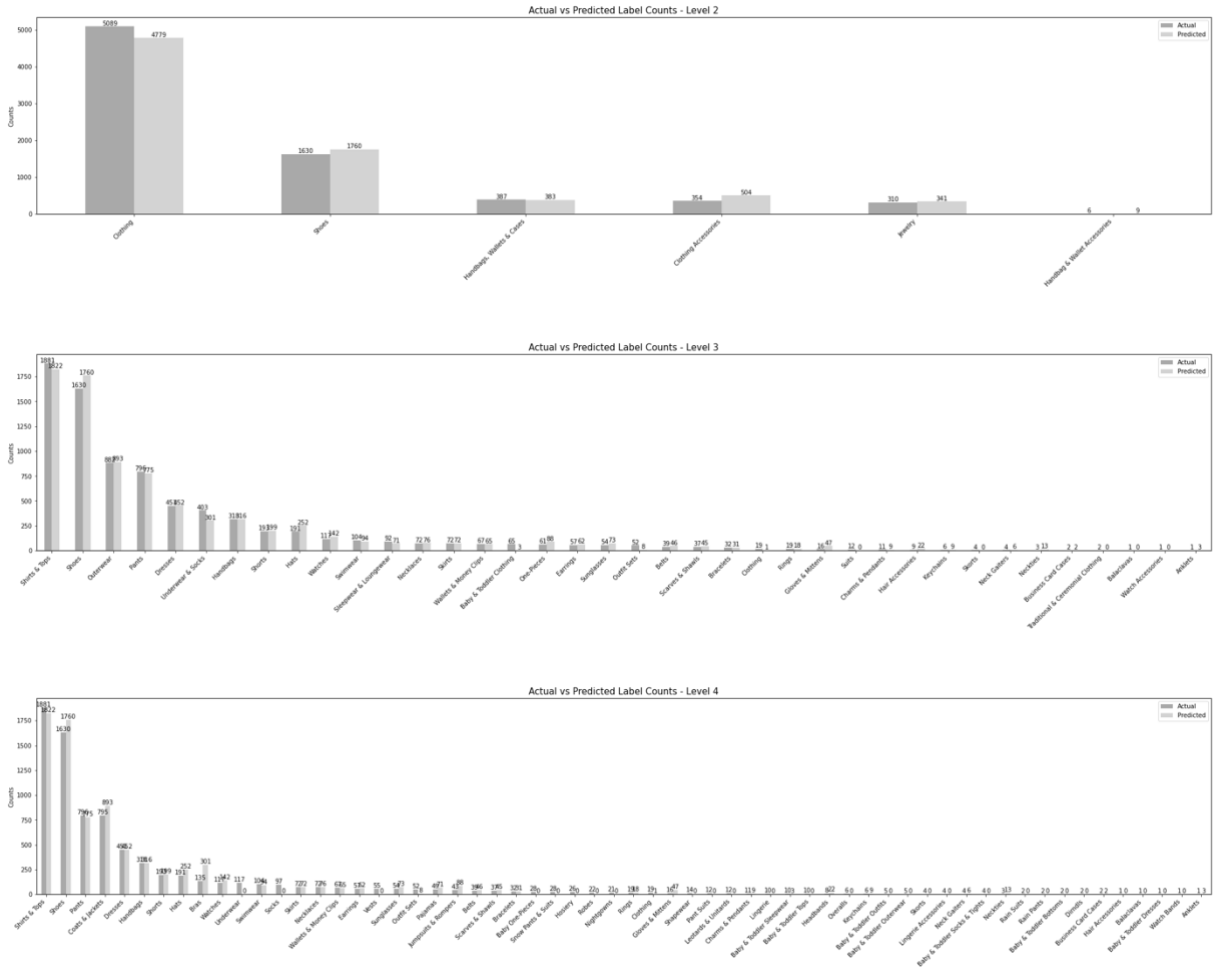
Top Currencies Other Platforms	
	Share
USD	80.26%
EUR	14.73%
GBP	3.82%
SAR	0.14%
PHP	0.12%
AUD	0.12%
MYR	0.12%
COP	0.08%
IDR	0.08%

A.I. Overview of Bert variants and most important attributes

Overview of BERT variants

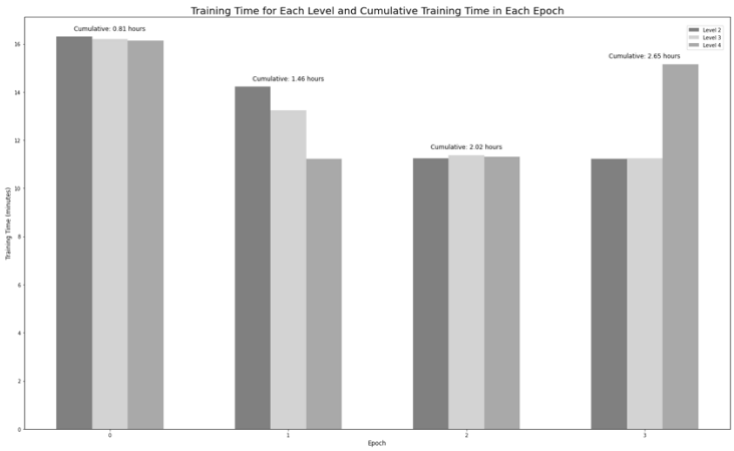
	BERT	ALBERT	DistilBERT	RoBERTa
Layers / Hidden Dimensions / Attention Heads	Base: 12 / 748 / 12 Large: 24 / 1024 / 16	Base: 12 / 748 / 12 Large: 24 / 1024 / 16	Base: 6 / 748 / 12	Base: 12 / 748 / 12 Large: 24 / 1024 / 16
Parameters *(base version)*	110 million	12 million	66 million	125 million
Data	16GB (BooksCorpus + English Wikipedia)	16GB (BooksCorpus + English Wikipedia)	16GB (BooksCorpus + English Wikipedia)	160GB (BERT + CCNews + OpenWebText + Stories)
Method	MLM & NSP	BERT with reduced parameters	BERT Distillation	BERT without NSP, Dynamic Masking

A.II. Actual vs. Predicted Categories Dynamic Masked Softmax Hierarchical Model on Zalando Test Set



A.III. Training Times Hierarchical Dynamic Masked Softmax Model (Zalando

Training)



A.IV. Actual vs. Predicted Categories Dynamic Masked Softmax Hierarchical Model on Cross-Platform Test Set