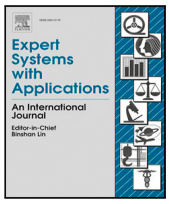
Expert Systems With Applications 287 (2025) 127702

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



What is in a title? Characterizing product titles in e-commerce

Sharon Hirsch [a,*], Ido Guy [a], Slava Novgorodov [b], Gal Lavee [c], Bracha Shapira [a]

[a] Ben-Gurion University of the Negev, Beer Sheva, Israel
[b] Tel Aviv University, Tel Aviv, Israel
[c] eBay Research, Netanya, Israel

ARTICLE INFO

Dataset link: https://drive.google.com/drive/folders/1RsxFzlsvYUUUxuZcoOs3d9MA-UbFNayv?usp=sharing

Keywords:
Electronic commerce
Product titles
Title analysis
Linguistic analysis

ABSTRACT

Product titles play a central role in connecting e-commerce buyers and sellers, concisely conveying the product information required to facilitate a transaction. Product titles are prominently presented in e-commerce search results, recommendations, and browse pages. In addition, many e-commerce tasks, such as matching, categorization, and product recommendations, heavily rely on product titles as a signal. In this paper, we report a comprehensive analysis of the linguistic characteristics of e-commerce titles. Specifically, we consider syntax, content, order of words, and attribute distribution. We compare these characteristics to those of the language used in other e-commerce and web corpora. Our analysis reveals a variety of unique properties of e-commerce title language. We consider the practical implications of our analysis with an empirical evaluation of modeling approaches on two real-world e-commerce tasks directly involving titles. Our findings suggest a number of practical approaches to applying natural language processing techniques over e-commerce titles. Based on these findings, we developed a set of clear, actionable guidelines for creating effective titles that can be adapted by e-commerce platforms.

1. Introduction

Modern electronic commerce products are stored and presented to buyers as multi-faceted objects, including such facets as images, descriptions, and semi-structured attributes in the form of name and value pairs. The *product title* (or *title* for short) is an important unstructured text facet used by sellers to convey the essence of the product to potential buyers. While there is some variance across different online e-commerce platforms on which facets constitute an e-commerce product, every product definition contains the title facet. Furthermore, the title is often displayed prominently in views of the product such as search results, product recommendations, and the product landing page. Finally, product titles are often indexed by search engines and recommendation systems to allow the discovery of products in search and personalization scenarios.

When constructing titles, e-commerce sellers must consider multiple objectives: (1) *Information* — summarizing the important aspects of the product; (2) *Marketing* — differentiating the product from others in search results or recommendation channels; (3) *Discoverability* — allowing algorithmic solutions for searching and browsing to discover the product and surface it to potential buyers. Further, to encourage brevity, most e-commerce platforms limit the title length (usually to 70–150 characters). Thus, the title author must determine how to trade off the above objectives given this constraint. The consequence of these dynamics is a unique and complex "language" of e-commerce titles, with its own vocabulary, syntax, and other linguistic qualities. These are important to understand even in an era where large language models become more widespread, as these may still require appropriate fine tuning or prompting.

The ubiquity of titles, combined with their inherent summarization capability, has driven research on using titles as a signal in a variety of e-commerce tasks. E-Commerce search (Bell et al., 2018), recommendation (Schafer et al., 2001), categorization (Shen et al., 2012), matching of inventory items to catalog products (Shah et al., 2018) , and others have all used titles as a primary signal. Furthermore, much research has been devoted to extracting structured product attributes from titles (More, 2016; Roy et al., 2021; Zheng et al., 2018). A recent demonstration of product title's centrality to the e-commerce domain is the introduction of a "next product title generation" task in the Amazon KDD cup 2023 (Deotte et al., 2023; Lee et al., 2023; Miyamoto et al., 2023). Despite the outsized role title facets occupy in the field of e-commerce, to the best of our knowledge, no previous work has studied their characteristics in a comprehensive manner.

* Corresponding author.
E-mail addresses: hirschsh@post.bgu.ac.il (S. Hirsch), gid@bgu.ac.il (I. Guy), slavanov@post.tau.ac.il (S. Novgorodov), glavee@ebay.com (G. Lavee), bshapira@bgu.ac.il (B. Shapira).

https://doi.org/10.1016/j.eswa.2025.127702
Received 27 March 2024; Received in revised form 23 February 2025; Accepted 10 April 2025
Available online 13 May 2025

In this work, we present a large-scale analysis of the language of e-commerce titles. Our analysis examines syntactic and lexical properties. We compare these properties across various text datasets, from both inside and outside the e-commerce domain. In addition, we use language modeling techniques to quantify the similarity of the title language to the other "languages". The results of our analysis allow us to reason about the common and different between e-commerce titles and other corpora across various dimensions. Our analysis treats e-commerce titles as a language, and considers many types of linguistic characteristics. To assess which of these have a practical impact on real-world e-commerce scenarios, we explore two practical tasks that are purely based on titles: *product title selection* and *product categorization*. We experiment with several modeling approaches and examine the performance based on different title properties. For each task, we conduct a feature importance analysis, which altogether inform the creation of a recommended set of guidelines for writing effective e-commerce titles.

The key contributions of this work are summarized as follows:

- An analysis of the "language" of e-commerce titles, a product facet whose importance is widely recognized, but whose properties have heretofore not been studied comprehensively.
- A comparative study of e-commerce titles, contrasting the properties of this data with other e-commerce and web corpora.
- An empirical evaluation of modern modeling approaches on real-world e-commerce tasks directly applied to titles.
- A clear set of guidelines for creating effective titles for products, which can be adapted by e-commerce platforms.

Our findings suggest a number of practical approaches to constructing e-commerce applications that rely on title information. We conclude the paper by summarizing the key findings and discussing their implications.

2. Related work

It is widely recognized that titles have an important role summarizing and promoting a wide range of complex content across many areas. Titles and their characteristics have been studied in a number of domains, including books (Symes, 1992), newspaper articles (Su et al., 2019; Vasilyev et al., 2019), academic papers (Hartley, 2005; Letchford et al., 2015), Wikipedia entries (Medelyan et al., 2008; Yıldırım et al., 2016), and YouTube videos (Larasati & Moehkardi, 2019). These studies consider properties such as title aesthetics, readability, and marketing effectiveness.

In the domain of news articles, a number of applications rely on the title facet. Su et al. (2019) propose a method for the identification of fake news articles using the article's title. They compare a number of neural embedding models. Vasilyev et al. (2019) propose a method for the automatic generation of titles for news articles. This method is trained on news articles and their corresponding headlines.

In the domain of e-commerce, to our knowledge, product titles have not been studied rigorously, but are widely acknowledged to be an important facet and are used to represent the product in many applied studies that have a particular downstream objective. Nicholson and Paranjpe (2013) build a Naïve Bayes model to predict the end price of an auction on eBay using title keywords as binary features. The study finds that specific words in product titles can influence the final sale price, with certain keywords boosting or lowering the probability of a successful sale in specific category. (Cholakov, 2009) builds software to allow sellers to optimize the title for their own listing by comparing to other listings in the same space. The task of e-commerce search, aiming at improving the relevance of products surfaced to free-text user queries, makes extensive use of titles, and has been considered in many recent studies (Bell et al., 2018; Sondhi et al., 2018; Tsagkias et al., 2021). For instance, Bell et al. (2018) define a method for weighting title terms to improve search relevance. Their work demonstrates how title terms are directly used to score and rank products for search queries, emphasizing their role in determining relevance. Additionally, the category information is treated as a term, since it is strongly correlated with relevance. In our research, we investigate how the title structure impacts the categorization task.

Extracting name-value product attributes from titles is another research area that has received much attention (More, 2016; Putthividhya & Hu, 2011; Xu et al., 2019; Zheng et al., 2018). Putthividhya and Hu (2011) present a named entity recognition (NER) system for extracting attribute values from e-commerce titles, which are in turn normalized based on n-gram substring matching. More (2016) proposes a sequence labeling algorithm with a "structured perceptron" to extract the brand from a product title. Recent studies used pre-trained language models (PLM) and large language models (LLM) (Baumann et al., 2024; Brinkmann et al., 2023; Roy et al., 2021; Wang et al., 2020; Yang, Wang, et al., 2023; Yang et al., 2022) to extract name-value attributes. AVEQA and MAVEQA (Wang et al., 2020; Yang et al., 2022) formulate the extraction task as a question answering problem using different pre-trained language models in their experiments. Particularly, each attribute is treated as a question, and they search for the best answer span in the product context (e.g., product title) that corresponds to the value. Another line of work applies prompt tuning to find the prompt that works best to extract the name-value attributes (Brinkmann et al., 2023; Yang, Wang, et al., 2023).

Generation of product titles from other modalities has also been recently explored (Camargo de Souza et al., 2018; Hancock et al., 2019; Ueffing et al., 2018; Yang, Liu, et al., 2023). Camargo de Souza et al. (2018) describe a method for generating the title of an e-commerce product given the titles of noisy seller-generated listings describing the product. In addition, they attempt to assess whether a title is suitable by predicting its quality. For this task, they use a set of features, including a bag-of-words representation of the item title, the length of the titles in tokens and characters, ratios of title length to aggregated average, maximum or minimum title length, and counts of repeated tokens, among others. Similarly, in our study, we focus on the task of title selection, but the set of features we use is determined by the linguistic analysis we perform, including features such as part-of-speech occurrences, proportions of stop words, punctuation marks, capitalized words, all-uppercase words, PCFG parse score, and more. (Yang, Liu, et al., 2023) build a set of prompts from different modalities, such as image and attributes, to generate titles of new products with limited labeled data. The multimodal prompt learning approach for product title generation aligns with our work by emphasizing the importance of structured title information in categorization and retrieval. While their study focuses on generating titles using modalities, our work provides a detailed linguistic analysis of title structures, which can enhance such models by informing attribute selection and optimizing title composition for better search and classification performance.

The use of titles has attracted attention in recent conference challenges focused on e-commerce applications. In the CIKM AnalytiCup 2017 challenge (Nguyen et al., 2017; Tay, 2017), the task of product title grading was introduced: given a product title, description, and attributes, two types of quality scores are considered: clarity and conciseness. The SIGIR eCom 2018 Data Challenge (eCom 2018 Data Challenge, 2025) focused on a large-scale taxonomy classification task, where the goal was to predict a product's category based on its title. Categorizing products within a structured taxonomy is a fundamental challenge for e-commerce platforms, with applications in personalized search, recommendations, and query understanding. Similarly to these two challenges, our work addresses the two tasks of title selection and product categorization. Through a detailed linguistic analysis of e-commerce titles, we examine structural and lexical properties that impact the performance of these tasks. The KDD Cup 2023 challenge focused on building a multilingual recommendation system that is based on session data (Deotte et al., 2023; Lee et al., 2023; Miyamoto et al., 2023). As part of the challenge, one of the tasks was to predict

the title for the next engaged product, given the user's history of interactions with product titles. The generated titles can improve various downstream tasks, such as cold-start recommendation of new products, unseen during the training phase.

As large e-commerce platforms are often backed by category hierarchies, product categorization is one of their most fundamental tasks. This task is often heavily reliant on product titles, as these are generally available for most products (Cevahir & Murakami, 2016a; Goumy & Mejri, 2018; Kozareva, 2015; Xia et al., 2017). Product matching, the problem of finding a correspondence between listed items to canonical catalog product representations, is also an important business problem for e-commerce platforms. Approaches to address this task naturally rely heavily on the title signal Shah et al., 2018; Stein et al., 2019. Other e-commerce applications making use of titles include title translation (Calixto et al., 2017; Chen et al., 2016), high-cardinality (a.k.a "extreme") classification (Shen et al., 2012), and title compression for mobile devices (Sun et al., 2018) and voice assistants (Mane et al., 2020). Our work provides an in-depth study of product titles, revealing and quantifying different unique qualities, with implications to downstream applications.

3. E-commerce title guidelines

The language of e-commerce titles is shaped by the guidelines communicated to sellers by the e-commerce platform. The large global e-commerce platforms provide instructions that share many similar aspects. In this section, we briefly review the principles and guidelines by a few of the leading e-commerce platforms.

Title guidelines on eBay (eBay Title Policy, 2024) strictly limit the length to 80 characters, and suggest that the title should contain the product's top features, while being as short as possible. In addition, titles should be unique, well written, and easy to read. The platform suggest to organize the features in a readable, logical order. Using acronyms, like NIB, is also not recommended, since the buyers may not understand them. Furthermore, titles should not contain misspelled words, information irrelevant to the product, or any false or misleading information.

On Amazon (Amazon Title Policy, 2024), the recommended title length is also 80 characters, but this limit is not enforced, and in some categories titles exist which contain 150 to 200 characters. Titles should contain the product's type (e.g., 'microwave' or 'umbrella') and should not contain promotional phrases (e.g., "free shipping" or "100% quality guaranteed"). Additionally, the title should include the minimum information needed to identify the item and nothing more. Each word in the title, excluding prepositions, should be capitalized. Numeral representation of numbers are preferred to "spelled out" representation ("2" rather than "two"). Subjective commentary (e.g., "hot item" or "best seller") should be avoided. Abbreviating measurements (e.g., "cm" or "oz") is also discouraged. Including merchant name in the title is not recommended, since it is not part of the original product name and may confuse buyers.

The title guidelines on Alibaba (Alibaba Title Policy, 2025) suggest including key product features, avoiding special characters and repeated keywords, and not using only a brand name or model number as the product title. A recent Alibaba study (Sun et al., 2018) also proposes product title guidelines. Essentially, titles should contain key product details, such as the brand name or commodity name, and should not contain irrelevant information.

The Walmart product title guidelines (Walmart Title Policy) recommend having clean and concise product titles and setting the title length to 50 – 75 characters. Titles should include key attributes that buyers are likely to search for, while avoiding "keyword stuffing" (Zuze & Weideman, 2013), i.e., repeating the same words or phrases so often that it sounds unnatural. Color, brand and model are good keyword examples to include in titles (Walmart Listing Optimization Guide).

Some special characters, such as exclamation marks, asterisks or trademark symbols, are also often forbidden by leading e-commerce platforms, unless they are part of the brand.

In our study, we consider the different aspects of the above guidelines and demonstrate how they influence the language and quality of product titles. In Section 11, we suggest, based on our findings, a revised set of principle guidelines that can be used by e-commerce practitioners to promote higher quality product titles.

4. Datasets and characteristics

The analysis in this work is based on a number of datasets from both e-commerce and general web contexts. We consider datasets from two major online e-commerce platforms, *eBay* and *Amazon*. Our analysis focuses on three of the most popular e-commerce product domains: *Electronics*, *Fashion*, and *Home & Garden*.

We consider two sources of e-commerce product titles:

- **eTitles** — refers to a dataset of product titles listed on *eBay*'s U.S. site during November 2020.
- **ATitles** — refers to a publicly available dataset of product titles from *Amazon* U.S. (Ni et al., 2019). These datasets are composed of the three product domains mentioned above.

In our analysis, we set out to compare and contrast the properties of e-commerce titles with other types of e-commerce textual data. To this end, we make use of the following datasets:

- **Queries** — freetext queries submitted to the *eBay* U.S. search engine (Trotman et al., 2017) during November 2020. We only consider queries categorized into one of the aforementioned domains.
- **Descriptions** — all product descriptions in the three domains from the publicly available *Amazon* U.S. dataset (Ni et al., 2019). Descriptions, like titles, are an unstructured text representation of the product. However, they are not bounded by character limits and their presentation is typically less prominent than the product's title.
- **Reviews** — all reviews in the three product domains from the publicly-available *Amazon* U.S. dataset (Ni et al., 2019). Reviews typically reflect subjective customer perspectives over a variety of product aspects (Popescu & Etzioni, 2007), as well as personal experiences, descriptions, and tips.
- **Questions** — all questions from the three product domains on a publicly available dataset of *Amazon* U.S. products, containing question and answer data (Wan & McAuley, 2016).
- **Answers** — all answers from the three product domains on a publicly available dataset of *Amazon* U.S. products, containing question and answer data (Wan & McAuley, 2016).

In our analysis, we also consider two, more general, web corpora:

- **Wikipedia** — a 1 GB sample of the English Wikipedia February 2021 dump, processed using WikiExtractor (Attardi, 2015).
- **News** — 1.2 GB of news articles in the years 2015-2017 (News Dataset, 2022), compiled from 15 US mainstream digital and print publishers, including *The New York Times*, *Business Insider*, and *The Guardian*.

In the remainder of this paper, we use the sentence as the basic unit of analysis. Thus, for datasets that contain multi-sentence documents, we apply an initial preprocessing step of sentence splitting (Loper & Bird, 2002). Following this step, the datasets are considered as a collection of sentences.

Table 1 presents summary statistics of the datasets described above. It can be observed that title length is similar across all *eBay* product domains, in the range of 11–12 tokens (median). However, *Amazon*

Table 1
Datasets and characteristics, including the number of sentences, median and average sentence length, and portions of stop words, punctuation marks, capitalized, and all-capitalized tokens.

Corpus	Source	#Sentences	Med (avg) len	StopW	Punct	Caps	AllCaps
Titles Electronics	*eBay*	7,587,256	11 (11.5)	4.7%	9.2%	61.9%	12.6%
Titles Fashion	*eBay*	38,818,187	11 (11.2)	2.5%	5.1%	73.5%	11.5%
Titles H&G	*eBay*	10,984,188	12 (11.5)	4.5%	5.9%	72.2%	7.7%
Titles Electronics	*Amazon*	786,272	15 (18.2)	5.6%	12.2%	58.6%	8.9%
Titles Fashion	*Amazon*	2,665,528	10 (10.3)	3.5%	5.3%	80.0%	4.1%
Titles H&G	*Amazon*	571,532	12 (14.1)	4.1%	11.5%	64.7%	5.1%
Queries Electronics	*eBay*	44,065,252	2 (2.8)	1.1%	0.4%	7.6%	2.5%
Queries Fashion	*eBay*	28,755,546	3 (2.8)	1.4%	0.4%	7.1%	1.9%
Queries H&G	*eBay*	18,086,969	3 (2.8)	1.3%	0.4%	7.4%	1.7%
Descriptions Electronics	*Amazon*	4,395,700	17 (20.1)	24.0%	13.5%	23.3%	4.3%
Descriptions Fashion	*Amazon*	10,140,475	15 (17.2)	26.5%	13.6%	20.0%	1.9%
Descriptions H&G	*Amazon*	3,257,256	15 (17.3)	24.2%	14.3%	19.8%	2.6%
Reviews Electronics	*Amazon*	4,676,621	14 (16.0)	36.4%	12.8%	14.2%	4.3%
Reviews Fashion	*Amazon*	2,922,032	11 (12.6)	35.5%	13.8%	15.2%	4.3%
Reviews H&G	*Amazon*	3,709,199	13 (14.9)	36.1%	12.6%	13.5%	3.6%
Questions Electronics	*Amazon*	454,212	10 (11.1)	39.8%	12.5%	15.8%	3.8%
Questions Fashion	*Amazon*	33,132	9 (10.2)	38.7%	13.4%	13.4%	3.3%
Questions H&G	*Amazon*	152,008	10 (11.0)	40.8%	12.5%	13.3%	2.7%
Answers Electronics	*Amazon*	805,751	12 (14.3)	37.7%	13.0%	17.0%	4.0%
Answers Fashion	*Amazon*	52,087	11 (12.9)	36.3%	13.4%	16.0%	4.2%
Answers H&G	*Amazon*	254,427	12 (13.9)	37.6%	12.7%	15.3%	3.2%
Entries	Wikipedia	1,544,187	23 (24.7)	32.3%	13.5%	17.3%	0.9%
Articles	News	5,365,725	21 (23.6)	34.5%	11.2%	15.1%	1.5%

title lengths exhibit more variation across the product domains, ranging from 10 (*Fashion*) to 15 (*Electronics*). Recall from the discussion in Section 3 that *eBay* enforces a cap on title length, while *Amazon* only has a non-binding length recommendation in place. The table reflects this disparity in the guidelines.

Considering the length of non-title datasets, we observe that queries (across all categories) are significantly shorter than titles, while "web" sentences tend to be longer (23 tokens for Wikipedia, 21 tokens for News). Furthermore, e-commerce descriptions also tend to be somewhat longer than other types of e-commerce data such as titles, reviews, questions, and answers.

Examining stop words (Loper & Bird, 2002), we can see that e-commerce titles across the two platforms and three product domains contain few stop words relative to other types of e-commerce and web data. Only the queries dataset contains a lower portion of stop words. On this dimension, titles are closer to queries than they are to other types of e-commerce data, which are more similar to web corpora. E-Commerce descriptions have a relatively lower proportion of stop words, but still considerably higher than titles.

The portion of punctuation marks is also lower for titles than for all datasets, aside from queries. It is somewhat higher for *Electronics* titles, with commas, dashes, and parentheses being the most common marks on this domain (e.g., "*LaCie Rugged 2 TB, External, USB-C, (STFR2000800) Hard Drive*").

One prominently characterizing property of titles is the portion of capitalized tokens, which is substantially higher than in all other datasets, at well over 50% across all title domains and sources. All-uppercase tokens, which are more common on all e-commerce domains compared to web, are also especially common on titles.

The remainder of our analysis is organized as follows. Section 5 examines syntactic properties of product titles, including parts of speech and parsing score. Section 6 looks into the vocabulary of product titles, inspecting the different token types, the distribution of stop words, and the use of attributes. Section 7, studies the ordering of product title tokens by assessing how predictable it is. In Section 8, we explicitly compare, using language modeling analysis, the similarity between title language and other e-commerce textual facets, in light of the differences revealed in the preceding sections. The final two sections of our analysis (Sections 9 and 10) examine the performance characteristics of different models over two fundamental e-commerce tasks that are solely based on titles: product title selection and product categorization. In addition to evaluating performance, these two sections reflect on the analysis in the previous sections by considering feature importance and examining results segmented according to key title properties. In Section 11, we build on the findings in all the preceding sections to suggest a revised set of guidelines for product title creation.

5. Syntactic characteristics

In this section, we delve deeper into syntactic analysis of e-commerce titles in comparison with the other datasets. Our analysis examines the distribution of key part-of-speech (POS) tags, using the Stanza POS tagger (Qi et al., 2020), and the parse score, using the Stanford parser (Klein & Manning, 2003). For the e-commerce datasets, we unified the analysis across all three domains, since the differences among them were minor.

5.1. Parts of speech

Table 2 displays the portion of POS tags (out of all tokens in the corpus) across different corpora. The table shows only the most prevalent tags: nouns (NN), adjectives (JJ), verbs (VB), prepositions (IN), determiners (DT), pronouns (PR), and adverbs (RB). For the e-commerce title datasets, nouns are the dominant part of speech. Both *eBay* and *Amazon* datasets contain a noun proportion greater than 65%. The titles' noun proportion is nearly double that of descriptions and more than three times that of reviews. This phenomenon may be explained by the fact that many e-commerce titles are composed of product attribute values, which are largely nouns. E-Commerce queries contain an even higher proportion of nouns than the titles, as Table 2 shows, at 74.7% of all tokens (for general web search queries, a lower 64.2% has been reported Guy, 2016).

Considering the non-nouns, e-commerce titles have very low verb, preposition, determiner, pronoun, and adverb counts compared to other e-commerce datasets. We also observe that e-commerce reviews, questions, and answers share a similar POS distribution profile. All these have a relatively large (6–10%) proportion of pronouns. This may be because of the first-person context of these datasets. Overall, it can be

Table 2
Part-of-speech distribution across datasets.

	NN	JJ	VB	IN	DT	PR	RB
eBay Titles	67.1%	6.7%	1.9%	2.8%	0.4%	0.1%	0.3%
Amazon Titles	65.4%	5.6%	1.7%	3.3%	0.4%	0.2%	0.2%
eBay Queries	74.7%	9.0%	2.9%	1.2%	0.3%	0.1%	0.2%
Amazon Descriptions	35.2%	8.8%	10.7%	8.2%	6.3%	2.7%	2.3%
Amazon Reviews	18.9%	8.6%	16.5%	8.7%	9.4%	8.8%	7.3%
Amazon Questions	23.9%	5.2%	17.1%	7.9%	10.8%	6.6%	2.9%
Amazon Answers	19.8%	6.4%	16.2%	8.5%	9.2%	9.5%	6.3%
Wikipedia	29.9%	7.4%	11.9%	12.7%	9.4%	2.1%	3.1%
News	27.1%	6.1%	14.9%	11.2%	8.7%	4.8%	4.2%

Table 3
Parsing score across datasets.

	eTitles	ATitles	Queries	Descr	Revs	Qs	As	Wikip	News
Average	−12.5	−12.1	−15.6	−8.8	−7.4	−7.5	−7.2	−7.3	−7.4
Median	−12.4	−12.0	−15.2	−8.3	−6.9	−7.2	−6.8	−7.1	−6.9

observed that despite the length differences, POS distribution on titles is closest to queries. Like queries, product titles are rich with nouns, and substantially sparser on all other parts of speech, aside from adjectives.

Since nouns in titles are so prevalent, we further analyzed the type of nouns that appear in e-commerce titles. Proper nouns are more frequent in titles (*eBay* 45.8% of all nouns, *Amazon* 44.1%) than in queries (33.5%), descriptions (16.0%), and reviews and q&a, where they are especially uncommon (7.6%–11.9%). In fact, titles are the only e-commerce corpus where the portion of proper nouns out of all nouns is higher than Wikipedia (36.0%) and News (35.3%). The high occurrence of proper nouns in titles is related to product attributes, which we further consider in Section 6.2. We also analyzed the proportion of plural forms among common nouns. Plural noun forms are infrequent in e-commerce titles (9.8% and 8.2% of all common nouns for *eBay* and *Amazon*, respectively). They are more common in queries (14.6%), descriptions (18.7%), and reviews (21.0%), and more common yet in web corpora (28.1% and 28.6% on Wikipedia and News, respectively).

5.2. Parsing characteristics

The final part of our syntactic analysis examines the *parse score*, which is the length-normalized log probability of the parse tree, using the Probabilistic Context Free Grammar (PCFG) of the syntactic parser. This score serves as a proxy for grammaticality, as low scores indicate sentences whose parses are atypical. Table 3 presents the average and median parse scores across our sentence datasets. It can be seen that the parsing score of e-commerce titles, in both eTitles and ATitles, are rather low, reflecting poor grammaticality. Among all other datasets, these parsing scores are closest to e-commerce queries, even though the average token count of titles is 4 times that of queries, as shown in Table 1. On the other hand, while the length of e-commerce titles is similar to sentences of e-commerce descriptions, reviews, questions, and answers, the parse score of these datasets are more similar to the web datasets, which contain many natural language sentences.

6. Lexical characteristics

In this section, we examine titles' vocabulary in comparison, where relevant, to other datasets. We first examine the distribution of stop words and then perform an analysis of the main token types that compose titles and specifically examine attribute values.

Table 4
Most frequent stop words across datasets.

eTitles	ATitles	Queries	Descr	Revs	Qs	As	Wikip	News
for	for	for	the	the	the	the	the	the
with	with	and	and	and	this	it	of	to
and	and	of	a	a	a	a	and	of
of	in	the	to	to	is	to	in	a
in	the	with	of	it	it	and	to	and

6.1. Stop words

While no unified definition of stop words exists, they are a fixed list of common words that are ignored by search engines and several natural language processing tasks (Loper & Bird, 2002). Table 4 shows the top 5 stop words in each of our datasets. Recall from Table 1, that stop words are generally less prevalent in e-commerce titles and queries than in other datasets. It can be seen that `for` and `with` are the most common stop words for both the *eBay* and *Amazon* titles. These "connector words" are often used in e-commerce to indicate the compatibility or target audience of a product, as well as a part/component or a complementary product. For example, in the title *Shockproof Armor Case Cover for Huawei Mate 20*, `for` indicates the compatibility of the phone cover to a particular brand and model. In the title *Nikon Pronea Film Camera with travel case*, `with` connects the main product to its complementary accessory.

Considering all e-commerce datasets in the table excluding titles and queries, we observe that the connector words `for` and `with` are not on their top lists. On the other hand, common stop words in these datasets, such as `the`, `a`, and `to`, are not among the top stop words in the title datasets, and are more similar to the appearance of stop words in web sentences.

6.2. Attributes

The vast majority of tokens that appear in titles correspond to product attribute values (More, 2016; Putthividhya & Hu, 2011). In this section, we quantify this phenomenon on the *eBay* titles dataset and inspect how the attribute content varies over different product domains. We also examine which other types of tokens compose product titles, alongside attribute values.

6.2.1. Token types

In order to carry out the analysis, we manually labeled each token in a title with a *token type*. The labeling was carried out in two phases. The exploratory phase served to identify the major token types and improve annotation task definition. These were then used to carry out the collection phase of tagging, which provided the data for our analysis. During the collection phase, 3 annotators, none of whom participated in the exploratory phase, labeled the tokens of 500 titles randomly sampled from each of the three product domains. To facilitate accurate labeling, annotators were given access to other product facets besides the title including product image, description, and structured attributes. The Fleiss' Kappa among annotators on this task, measured over 100 random titles, was 0.83, indicating high agreement (Fleiss, 1971).

Table 5 presents the distribution of title tokens across the 8 major token types identified by our analysis. Most prominently, we observe that the *attribute value* type describes over 80% of the tokens in all three domains. This result is consistent with previous research (More, 2016; Putthividhya & Hu, 2011) and with the part-of-speech analysis in Table 2, which shows high noun content in title data.

Marketing token type includes the following examples, among others: *very good luster*, *100% natural*, or *free shipping*. This type of token accounts for nearly 4% of the tokens across all domains. *Attribute name* tokens are extremely rare, indicating that title authors rely on the

Table 5
Distribution of token types on *eBay* titles.

	Electronics	*Fashion*	*Home & Garden*
Attribute Value	80.2%	82.9%	82.8%
Punctuation	10.1%	6.0%	7.0%
Marketing	3.8%	3.9%	3.7%
Function Word	3.2%	1.5%	3.3%
Attribute Name	0.3%	3.6%	0.3%
Descriptive	1.1%	1.7%	2.3%
External Reference	0.1%	0.1%	0.1%
Other	1.2%	0.3%	0.5%

attribute value to convey its context implicitly. A notable exception to this observation is the *Fashion* domain, where 3.6% of the tokens are attribute names. This disparity is largely due to a single token, 'size', which appears in over 30% of the *Fashion* titles. The *Descriptive* token type refers to product details that are not attributes (e.g., `Warm` in *Baby Girls Warm Fleece Dress*). These account for 1% to 2% of the tokens. *External reference* tokens (e.g., *read description* or *see video*) are uncommon (0.1%) in all domains. The 'other' category includes even more infrequent token types, such as usage (e.g., for `skin problems`), selling service (e.g., `repair`), and internal identifiers used by sellers (e.g., training shorts 984).

6.2.2. Attribute values

We have seen attribute values make up the majority of tokens in e-commerce titles. We therefore set out to explore which particular product attributes appear in the title and in what form. To scale up our analysis, we use a neural attribute extraction tool (Xin et al., 2018). This allows us to automatically extract the attribute name corresponding to each attribute value in the entire *eBay* titles dataset. Table 6 shows the most common attribute values (aggregated by their corresponding attribute name) for each product domain. The second column of each product domain shows the portion of titles in which this attribute name appears. It can be seen that differences in product domain affect which attributes appear in the title. The *type* attribute (e.g., `Shoes`, `Vest`, or `Chair`) is more common in the *Fashion* and *Home & Garden* domains (>85%) than in *Electronics* (<65%). Titles in the *Electronics* domain are more likely to contain a `model` or `brand` token than the other domains. Many attributes appear on the top list of only a single domain. These include `game name` and `connectivity` in *Electronics*, `size` and `gender` in *Fashion*, and `u.o.m` (units of measure, e.g., 'kg' or 'cm') and `number of pieces` in *Home & Garden*.

The rightmost columns of each product domain in Table 6 show the position distribution for each attribute. Prefix and suffix refer to the first and second half of the title, respectively (for titles of odd length, the middle token was neither considered part of the prefix nor the suffix). It can be observed that `brand` attribute values frequently occur in the prefix of the title. `Game name` in *Electronics*, `gender` in *Fashion*, and `#pieces` in *Home & Garden* also tend to appear in the prefix. On the other hand, `condition` and `size` attribute values tend to appear in the suffix of the title. `Color` attribute values tend towards the suffix in *Electronics* and *Home & Garden* titles, but not in *Fashion*. Other attribute values spread rather evenly across the title, between prefix and suffix. Prominent examples include `model`, `type`, and `material`.

7. Order predictability

In this section, we study the ordering of tokens in the product title. Particularly, we consider the question: is there a natural ordering of a given set of title tokens? More practically, can the order of a set of title tokens be predicted by a trained model?

To analyze these questions, we considered subsets of the datasets in Section 4 which contain only sentences with precisely 10 tokens (excluding queries, as they rarely include 10 tokens). Each such subset consists of 100,000 sentences selected randomly across all product domains. In the learning phase, we trained a model to generate each sentence given a noisy version of the sentence, created by randomly permuting the tokens in the sentence. To this end, we employed Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2020). BART is a "denoising encoder" trained to reconstruct text samples from a corrupted version of those samples, and is thus a natural fit for our task. We randomly split each of the sentence sets into training (80%) and test (20%) sets. The training set was further split into training (80%) and validation (20%) sets to enable hyperparameter tuning for BART.

Intuitively, if a model trained on the learning task preforms well on a hold-out test set, we claim token order information is predictable in this dataset. To quantify this effect, we used the *BLEU* metric, often used in machine translation (Papineni et al., 2002). This metric was also used in the KDD Cup 2023 challenge for the title generation task (Deotte et al., 2023). BLEU-m is the geometric mean of n-gram precision for $n = 1 \dots m$. In our analysis, we made use of the BLEU-3 and BLEU-4 metrics over the test set. We also considered the simpler *Match* metric, which is the percentage of samples that were exactly reconstructed from their permuted tokens.

Table 7 presents the results of the order analysis. In the first section, we consider the task of reconstructing the entire sentence from a corrupted version. To obtain further insights, the second and third sections consider the task of reconstructing the prefix (first 5 tokens) and suffix (last 5 tokens) of the sentence, respectively. Examining the first section of the table, we observe that all the metrics we considered, BLEU-3, BLEU-4 and Match, confirm one another and imply correlated rankings of the datasets. The scores for both *eBay* and *Amazon* titles are considerably lower than for all other datasets. This indicates that token order is relatively less predictable in e-commerce titles. E-Commerce descriptions are noticeably more predictable than titles. Other e-commerce datasets (reviews, questions, and answers) are even more predictable, comparable to the web datasets, Wikipedia and News.

Comparing the second and third sections, we observe that the prefix and suffix of the title are equally (un)predictable. Considering non-title datasets, we observe that the prefix is more predictable than the suffix, even as absolute predictability varies. This gives another indication of the lower importance of token order in titles.

8. Language model similarity

Our analysis thus far has revealed unique lexical and syntactic characteristics of e-commerce titles compared to other corpora. In this section, we set out to examine the similarity of titles to other datasets using language modeling. Particularly, if each of our datasets is created by a generative process, how likely is that process to generate an e-commerce title?

To this end, we build language models using our comparison datasets and analyze how well the *eBay* title data is predicted by these models. More specifically, we build n-gram language models (LMs) with Kneser–Ney smoothing (Chen & Goodman, 1999) for $n = 1, 2, 3$ (i.e., unigram, bigram and trigram). For each e-commerce dataset, we fit language models for each of the product domains: *Electronics*, *Fashion*, and *Home & Garden*. For the web datasets, which do not decompose into product domains, we fit language models using the entire data. We use the language model *perplexity* to quantify similarity between titles and other datasets (Rosenfeld, 2000). Given a probabilistic model, perplexity is a measure of how well the model predicts a particular sample. Intuitively, if a language model fit on some sentence data predicts e-commerce titles well, that data is similar to e-commerce titles.

More formally, let $D = \{x_i\}_{i=1}^{N}$ denote a dataset of N independent samples. Further, let $P(x_i \mid \theta)$ denote the probability of a particular sample x_i given the model parameters θ. The *cross entropy* of the dataset

Table 6
Top title attributes, with the percentage of containing titles ('%') and distribution between prefix ('P') and suffix ('S').

Electronics				*Fashion*				*Home & Garden*			
Attribute	%	P	S	Attribute	%	P	S	Attribute	%	P	S
model	83.5%	50.2%	40.6%	type	85.9%	34.5%	57.7%	type	88.5%	42.1%	47.6%
brand	73.3%	81.3%	16.2%	brand	59.0%	88.6%	9.8%	brand	47.0%	83.2%	14.9%
type	64.6%	40.6%	50.5%	color	53.0%	48.8%	42.5%	model	34.7%	52.8%	40.3%
condition	21.5%	30.4%	67.0%	size	43.5%	25.6%	70.1%	material	33.7%	50.3%	44.0%
game name	18.2%	68.0%	21.8%	gender	42.8%	78.8%	18.3%	u.o.m	26.0%	41.7%	50.4%
connectivity	16.8%	44.6%	48.6%	material	35.6%	43.9%	48.4%	color	20.5%	40.2%	55.3%
color	16.4%	33.5%	61.0%	model	22.3%	50.1%	44.3%	#pieces	17.8%	63.4%	33.7%

Table 7
BLEU scores and portion of exact matches of BART model for predicting token order within 10-token sentences.

	Whole Sentence			Prefix			Suffix		
	BLEU-3	BLEU-4	Match	BLEU-3	BLEU-4	Match	BLEU-3	BLEU-4	Match
eBay Titles	0.42	0.29	14.9%	0.29	0.22	18.6%	0.31	0.23	19.5%
Amazon Titles	0.41	0.26	8.2%	0.30	0.21	15.1%	0.27	0.17	13.1%
Descriptions	0.52	0.39	15.7%	0.38	0.29	23.3%	0.30	0.22	18.0%
Reviews	0.73	0.62	35.0%	0.58	0.50	44.2%	0.51	0.42	37.9%
Questions	0.72	0.61	33.4%	0.68	0.61	53.6%	0.49	0.40	34.8%
Answers	0.73	0.62	34.1%	0.59	0.52	45.5%	0.49	0.41	36.4%
Wikipedia	0.70	0.59	29.8%	0.55	0.46	39.8%	0.47	0.38	33.0%
News	0.75	0.64	35.7%	0.57	0.49	43.6%	0.54	0.45	39.9%

relative to the model is given by $H_\theta(D) = \sum_{x_i \in D} -\frac{1}{N} \log_2 P(x_i \mid \theta)$. The perplexity is then derived by $2^{H_\theta(D)}$.

In our analysis, the model is the unigram, bigram, or trigram LM. We compute the perplexity as above, substituting the median of $\log P(x_i \mid \theta)$ for the mean in the cross-entropy expression.[1] Perplexity can be conceived as an aggregate inverse probability. That is, lower perplexity scores correspond to better models, or in our context, stronger similarity with titles.

Table 8 shows the application of the perplexity measure to model dataset similarity using the LMs described above. Specifically, we measure the perplexity of each of the three product domains of the *eBay* titles dataset. Note that for the purpose of this analysis, the comparison is aligned. That is, for each domain, we consider the LM trained over this domain only (e.g. *eBay* titles from the *Electronics* domain will be evaluated by models fit on the *Electronics* domain). The exception to this are the web datasets, which are not stratified into product domains, and are used in their entirety across all domains, as a reference baseline.

Examining the columns corresponding to unigram LMs in the table, we observe that *eBay* titles are most similar to *eBay* queries (lowest perplexity values). The next most similar datasets are the questions and ATitles datasets. Since unigram models reflect the vocabulary of the language, we confirm the similarity of titles and queries, observed in previous analysis. The fact that titles from *eBay* are closer to *eBay* queries than they are to titles from competing platform *Amazon*, could be explained by imperfect domain alignment as well as differences in the selling flow and title authorship guidelines between the two platforms (Section 3).

Inspection of the columns corresponding to bigrams and trigrams reveals a narrower similarity gap between queries and ATitles. In the *Fashion* domain, the similarity to *Amazon* titles is slightly higher than to *eBay* queries. As these LMs capture more local structure, we conclude that the similarity in structure between the title datasets offsets the differences in word choice reflected in the unigram results. Similarly, we observe that descriptions are also closer to titles in language structure. All of these findings are consistent across all three e-commerce domains.

[1] We use median rather than average due to the high variance among title scores.

9. Product title selection

Product titles are often an important input to supervised learning models trained to solve key practical downstream tasks in e-commerce (Kozareva, 2015; Shah et al., 2018; Tsagkias et al., 2021; Tzaban et al., 2020). In the final part of our analysis, we examine two e-commerce downstream tasks that are directly based on product titles, and reflect upon our analysis from the previous sections. The first task, described in this section, is *Product Title Selection*: given two e-commerce titles describing the same catalog product, determine which title better represents the product. The task is motivated by an important application called *product construction*. Oftentimes, different sellers will list the same product with different listing titles. For the canonical product title, there are many desirable properties (see Section 3). In this application, we aim to select the best seller-provided title among many candidates. A similar task was introduced in the CIKM AnalytiCup 2017 challenge, which focused on grading product titles (Nguyen et al., 2017; Tay, 2017). We address this task as a supervised learning problem and recruit human annotators in order to produce ground truth labels.

The decision of whether a title is appropriate for a product is complex and multi-dimensional (see the discussion on title guidelines in Section 3). Thus, in order to simplify the collection of labels, annotators were presented with a pair of titles, and asked to select the better of the two. This facilitates the decision by allowing the annotator to focus on the difference between the two titles. To collect the dataset used in this analysis, we employed 22 in-house annotators with expertise in the product domains *Electronics*, *Fashion*, and *Home & Garden*, who were asked to choose the better of two titles, representing the same product. Annotators were trained for this task in several rounds, and their guidelines included different considerations for selecting a better title, such as the inclusion of useful information, the exclusion of irrelevant or redundant information, and the overall structure and formatting quality.

To create a dataset for annotation, we sampled at random products from each of the three domains in the *eTitles* dataset described in Section 4. For each product, we considered all pairs of seller-provided titles and sampled one pair at random (if exists), out of all pairs that fulfilled the following three conditions: (1) originated from different sellers; (2) did not have one title being a substring of the other; (3) did not have a complete overlap between their tokens. Overall, the annotators labeled 5000 pairs of titles across the three product domains. The Fleiss' Kappa, measured among three of the annotators

Table 8
Perplexity of *eBay* titles w.r.t other language models.

	Electronics			*Fashion*			*Home & Garden*		
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
ATitles	2285	1685	1430	1474	836	830	2424	4231	3791
Queries	325	909	635	463	1031	865	438	2680	2081
Descriptions	4380	2592	2271	4662	1897	1672	3783	6558	5523
Reviews	3825	3701	3586	2323	3886	3786	3272	9,884	9652
Questions	1274	3593	3443	1062	5194	3611	1432	8349	8062
Answers	2618	4157	4036	2168	7826	5166	2584	9994	9748
Wikipedia	15,897	21,586	21,117	37,935	26,354	26,047	19,157	29,197	28,752
News	16,363	35,996	33,809	32,391	25,499	24,356	17,997	34,187	32,350

over 100 random titles, was 0.70 for this task, indicating the selection is often difficult even for humans (Fleiss, 1971).[2] The labeled set was randomly split into training (80%) and test (20%) sets. The training set was further split into training (80%) and validation (20%) to enable tuning the relevant hyperparameters for each model.

In order to model product titles and learn how to select a better title out of a pair of input titles, we experimented with three types of models: first, we considered gradient boosting models, XGBoost (Chen & Guestrin, 2016) and CatBoost (Dorogush et al., 2018), which take explicit features as their input. We considered a set of features inspired by the analysis in previous sections, including the portion of stop words, punctuation marks, capitalized words and all-uppercase words (Section 4). Additional features included part-of-speech occurrence (Section 5.1), the PCFG parse score (Section 5.2), frequent unigrams and bigrams (Section 6.1), and common attributes (Section 6.2).

Secondly, we considered unidirectional and bidirectional variants of a recurrent neural network with long short term memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997), which models sequential dependence, but only indirectly encodes the features considered above. Token representations in these models were initialized using word2vec (Mikolov et al., 2013) embeddings pre-trained on a large corpus of over 10M product titles. These vectors encode global word-sense in the domain of e-commerce titles. Finally, we considered *cased* and *uncased* variants of the transformer-based BERT architecture (Devlin et al., 2018), pretrained on a large corpus of over 10M product titles. Conceptually, this pretraining allows the model to have *contextual* embeddings of each token representing the domain-specific word-sense adjusted for the context of the entire title. Like the recurrent models, the transformer models only encode the features we considered above indirectly (if at all).

Table 9 shows the accuracy of the six different models on the Product Title Selection task. It can be observed that the BERT-based architectures (both cased and uncased) outperform the models trained on explicit features and the LSTM models with pre-trained word2vec embeddings. Furthermore, using case-sensitive tokens outperforms the uncased BERT variant. The rather noticeable gap between the cased and uncased BERT variants is in accordance with our findings in Section 4, observing e-commerce titles having an exceptionally high proportion of capitalized tokens compared to any other dataset. These apparently play an important role when selecting the BERT variant for the product title selection task.

The LSTM models with pre-trained word2vec embeddings underperform other model classes. We conjecture that the lack of inductive bias provided by explicit feature encoding combined with only global embedding (in contrast to contextual embedding available for BERT) of word-sense hinders this model class from learning this task to high accuracy.

The overall accuracy for the product title selection task, as presented in Table 9, is not very high, reflecting the difficulty in the task, as

[2] An anonymized sample of the labeled set can be found in https://anonymous.4open.science/r/Choose-better-title. The full set of labeled title pairs will be released if the paper is accepted.

Table 9
Accuracy of pairwise title selection classifiers.

XGBoost	CatBoost	UniLSTM	BiLSTM	BERT-cased	BERT-uncased
68.3%	69.8%	64.6%	66.3%	**81.2%**	78.3%

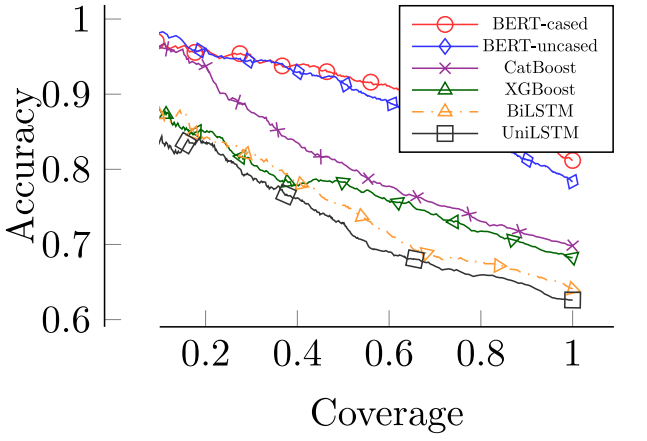


Fig. 1. Accuracy by coverage for the title selection task.

in some cases a pair of titles can be comparable in their suitability to represent the product. If the variance in difficulty observed in the results is reflected in the confidence output of the model, it may be possible to trade-off coverage for accuracy. That is, if we allow the model to "abstain" from a decision on some portion of the test samples (where it has low confidence in its response), the responses we do collect may be more accurate. To investigate this, we set a confidence threshold, and calculate the accuracy for only those examples for which the model has made a prediction with confidence above the threshold. Practically, since our models all yield a score between 0 and 1, we use this score as the model confidence. Varying the threshold allows us to plot the accuracy versus coverage tradeoff, as shown in Fig. 1. Indeed, a clear and smooth trade-off can be observed for all the approaches shown. This confirms that the trained models are, in some capacity, aware when a comparison is more difficult and the difference between titles more subtle. For the best performing model, BERT-cased, if we tolerate coverage of 80%, accuracy reaches 85.7%. For coverage of 60%, accuracy rises to 91.1%.

In order to analyze the contribution of different types of features, we performed a feature importance analysis. This type of analysis is well understood for models that take explicit features (e.g. XGboost) (Pedregosa et al., 2011). Modern neural models, however, do not take explicit features and, thus, standard feature importance techniques do not apply to them. Still, we are interested in the question of whether these architectures indirectly capture these types of features in its dense vector representation of the problem. To answer this question, we adapt the idea of *Concept Activation Vectors* recently proposed in the model interpretability literature (Kim et al., 2018). Essentially, each feature is

Table 10
Top 10 features by their importance.

Feature	Type
num_tokens	concept-based, explicit-based
aspect_brand	concept-based, explicit-based
aspect_color	concept-based, explicit-based
num_tokens_all_caps	concept-based, explicit-based
pcfg_score	concept-based
num_attributes	concept-based, explicit-based
num_verbs	explicit-based
num_adjectives	explicit-based
sentiment_score	explicit-based
word_len_entropy	concept-based

Table 11
Accuracy, macro precision, and macro recall of product categorization classifiers.

	XGBoost	CatBoost	UniLSTM	BiLSTM	BERT-cased	BERT-uncased
Accuracy	81.6%	72.5%	89.1%	89.3%	88.6%	**91.0%**
Macro-P	80.9%	73.5%	86.4%	86.7%	85.0%	**88.1%**
Macro-R	73.6%	59.5%	82.7%	83.3%	80.9%	**86.4%**

Table 12
Accuracy of product categorization using BERT-uncased by title length.

Length (#tokens)	1–7	8–9	10–11	12	13	14	15+
% of all titles	13.1%	12.1%	21.7%	14.6%	13.3%	10.1%	15.2%
Accuracy	88.0%	90.0%	91.0%	91.4%	91.9%	92.3%	92.4%

encoded as a binary class. A linear model is then trained to predict this class label using the final dense layer of the *trained* model as its feature input. If the linear model is a good predictor of this concept on a hold-out set, then we claim that the concept is encoded well by the neural representation. Since concepts can be mapped to features, we use this technique to indirectly assign feature importance in models that do not explicitly encode features.

We applied feature importance analysis to all models except the LSTM variants, which we excluded owing to their poor performance on the task. The results of the analysis is depicted in Table 10. This analysis confirms the importance of capitalization for this task, as all models scored capitalization features (*num_tokens_all_caps*) among the most important ones, together with the number of attributes (*num_attributes*) and PCFG parse score (*pcfg_score*). Additional important features across models were binary valued features indicating the occurrence of verbs (*num_verbs*) and `brand` (*aspect_brand*) and `color` (*aspect_color*) attributes. Explicit feature models assigned more importance to POS-related features (e.g., *num_verbs*, *num_adjectives*), which were not found to be meaningful in the analysis of BERT models. Furthermore, features related to sentiment (*sentiment_score*) were found to be important for explicit feature models but not for BERT models. These findings suggest that BERT does not rely on these features when deciding on this task. Overall, the findings depicted in Table 10 play a central role in forming our recommended set of guidelines in Section 11.

10. Product categorization

One of the most fundamental tasks in e-commerce, is the assignment of a given product into one of a wide variety of categories, which reflects its type (e.g., a guitar, a table, or a t-shirt). Accurately identifying the product's category allows its proper presentation on the platform, along with the relevant attributes (e.g., zoom range and resolution for a camera), and organizing the set of products offered for sale in a way that enables easy navigation. In addition, category identification allows to assist sellers when they upload products into the platform, as well as support search and recommendation on the buyers' side. A number of studies explored the task of product categorization using different facets of the product, including its description (Cevahir & Murakami, 2016b; Chen & Warren, 2013; Li et al., 2018), reviews (Huang et al., 2012), attributes (Krishnan & Amarthaluri, 2019), and images (Ristoski et al., 2018; Wirojwatanakul & Wangperawong, 2019), often combining multiple facets (Cevahir & Murakami, 2016a; Shen et al., 2012). Despite being relatively short, the product's title is also a highly productive source for categorization, typically available for all products (Hasson et al., 2021; Paulucio et al., 2020; Shen et al., 2012). A recent data challenge at SIGIR eCom 2018 also focused on this task, aiming to predict a product's category based on its title (eCom 2018 Data Challenge, 2025).

In this section, we examine the use of product titles, with their unique characteristics depicted in previous section, for the task of product categorization. To this end, we considered all titles assigned to a product category in the *eTitles* dataset (Section 4), and filtered these to include only categories with at least 100 titles. This resulted in a set of 14.6 million titles spanning 586 categories, across the three domains: *Electronics* (3.6M), *Fashion* (3.5M), and *Home & Garden* (7.5M). We experimented with the same set of six models and the same set of features as described for the title selection task (Section 9). Similarly, the set was randomly split into training (80%) and test (20%), while the training set was further split into training (80%) and validation (20%) to enable hyperparameter tuning.

Table 11 shows the accuracy, macro precision, and macro recall over the test set. The macro metrics consider the average precision and recall across all classes, while assigning each class with an equal weight, regardless of the number of its associated instances. Overall, the BERT and LSTM models achieve high performance for this task, with accuracy exceeding 88.5%, while the gradient boosting models substantially underperform. It appears that these explicit feature-based models fail to capture many of the patterns that serve to accurately identify the category based on the product's title. The best preforming model for this task across all metrics is the BERT-uncased variant, outperforming both LSTM variants, and, by contrast to the Product Title Selection task, also the BERT-cased variant, with accuracy reaching 91%. Apparently, as opposed to the title selection task, casing does not play a meaningful role in categorization, making BERT-uncased a more suitable modeling approach. Overall, these results establish that the title is indeed an effective source for product categorization. The performance results for categorization across the three domains were very similar. For instance, the BERT-uncased model achieved accuracy of 90.9%, 91.2%, and 90.9% for *Electronics*, *Fashion*, and *Home & Garden*, respectively. We therefore present the analysis across all three domains combined, without separating.

Feature importance analysis identified the occurrence of `type`, `brand`, and `gender` among the most important features. Additional important features included words that identify the product's type (e.g., doorbell, backpack, toaster) and units of measure tokens, such as '120v' and 'ounce'. However, since the feature-based models poorly performed for this task, we also set out to explore the correlation of different title characteristics examined in the previous sections with the product categorization performance. To this end, we conducted analysis of the performance (accuracy, in particular) of the BERT-uncased model, while segmenting the test set according to different title properties. The large size of our test set enables such analysis, while preserving a substantial set of instances in each analyzed segment. The most distinctive signal could be observed, as might be expected, for the title length. Table 12 shows that the longer the title, the higher the performance, up to 92.4% accuracy for titles of 15 tokens or more.

Other than length, we examined the performance of the BERT-uncased categorization model according to a variety of characteristics, aligned with the analysis presented in Sections 4, 5, and 6. Table 13 presents the accuracy difference for titles that contain key properties (e.g., a certain POS tag or attribute) compared to all other titles. Since many of these characteristics are correlated with the title length (e.g., the probability for a title to contain a punctuation mark increases with its length), we also examine the same performance difference,

Table 13
Delta in accuracy of BERT-uncased for product categorization over titles containing different properties, considering all titles and 12-token titles only. The '%' column marks the portion of titles containing the respective property.

Contains	%	All	12 tokens	Contains	%	All	12 tokens
Stopword	38.9%	−0.2%	−1.2%	Brand	58.8%	+1.0%	+1.2%
Punctuation	42.9%	+0.7%	+0.5%	Color	20.6%	+2.1%	+1.6%
Proper noun	73.6%	+1.1%	+1.2%	Material	22.0%	+0.9%	+0.1%
Adjective	50.6%	+1.3%	+0.5%	Measure	24.0%	+0.9%	+0.7%
Preposition	33.3%	−0.2%	−1.1%	Series	44.2%	+0.3%	+0.3%

considering titles of 12 tokens only, which is the median title length within the *eTitles* dataset (Section 4). The table also presents, for each property, the portion of titles that contain it out of all titles. It can be seen that titles that include a stop-word or a preposition yield lower performance than titles that do not contain these (especially when fixing the length to 12), implying stop-words and prepositions on titles do not substantially contribute to the categorization task. On the other hand, the occurrence of a proper noun coincides with a gain in categorization accuracy, presumably as proper nouns tend to be quite specific and therefore more likely to disclose the category.

In terms of attributes (right section of Table 13), titles that contain a *brand* name, and even more so a *color*, yield higher categorization performance (even for a fixed length). While brands are often specific to a single or very few categories, the association of colors with high categorization performance is more surprising, since the common colors are not category revealing on their own. On the other hand, the occurrence of *material* on a title does not seem to correspond with higher performance.

In addition to the properties shown on Table 13, we found that categorization accuracy was lower for titles with particularly low parsing score (bottom 10%) and low number of nouns (half of the tokens or fewer, accounting for 15% of all titles). We did not find any correlation of the categorization performance with capitalization properties as well as parsing, POS tags, and token type properties beyond the ones reported above.

11. Title guidelines revisited

In Section 3, we reviewed the title creation principles and guidelines used by leading e-commerce platforms. Our analysis and experimentation revealed various insights that can help revise the guidelines for more effective title creation. In this section, we suggest several key principles derived from our findings, which together form a recommended guidelines for e-commerce platforms.

Based on our in-depth analysis across sections, particularly regarding product categorization and title selection tasks, with an emphasis on feature importance analysis, we propose the following guiding principles for developing title creation guidelines in e-commerce platforms. It is important to note that these guidelines were developed specifically for the tasks analyzed, and for other tasks, different or even contradictory guidelines may emerge.

- Consider the key attributes that characterize your product by order of importance to the product's description. These can include general attributes, such as `brand` or category-specific attributes, such as `zoom range` for cameras, and `material` for shirts. The title structure consists mainly of product attributes and there is a limit to the allowed number of tokens the seller can include. Selecting attributes based on their importance is therefore essential for accurately describing the product.
- When relevant, make sure to include the `brand`, the `color`, and the `type` of your product in the title. These attributes are highly important for the title selection and categorization tasks, as they provide category-specific details and highlight key product information.
- Include the top-ranked attributes by their value only (e.g., "red" or "large"), and avoid mentioning the explicit attribute name (e.g., "red color" or "large size"). Mentioning the explicit attribute name is almost always redundant and takes potential room of other valuable information.
- Order the attribute values by their importance. Many applications use product titles for downstream tasks, such as categorization or search ranking. In case of a length limit where the full title cannot be used, ensuring that the most important attributes are listed first ensures that key information will still be included in these applications. In addition, for better readability, more important attributes should appear earlier in the title, as this helps customers quickly understand what the product is.
- When relevant, use unit of measures next to numbers (e.g., 5 kg or 110v). In the task of categorization, units of measurement were identified as important features, as they are often category-specific.
- Make your title as close to natural language as possible. Avoid using pure keyword staffing (long chain of unrelated nouns) and try to connect your attribute values with adjectives and verbs. Well-structured language is important for the title selection task and generally improves the readability of the title.
- Use prepositions, such as `for` when describing compatible models or `with` when describing bundles. Use other functional words to disambiguate or increase readability, as needed.
- Use capitalization, especially for names of brands and models, but avoid altogether using all-capitalized terms. Capitalization improves the readability of the title and helps identify named entities such as brands or models.
- Avoid any redundant or subjective information. Marketing statements are not going to make your product more prominent. These types of information do not improve the product's ranking in search results or categorization and may negatively impact the product's visibility, as search algorithms prioritize accurate and relevant information to match user queries.
- After considering all the above, include as many attributes as the title length limitation allows.[3] The number of attributes is important in the tasks of title selection and categorization, as having more relevant attributes makes the title more informative.

Based on our extensive analysis of data across two leading e-commerce platforms and various key e-commerce domains, and based on our experimentation with real-world title-based tasks, we believe these guidelines can help improve product titles on e-commerce platforms. Each of our principles stem from a key analysis and experimentation results. Future work should further validate this by conducting online experimentation and comparing the effect of the title guidelines on search effectiveness, sales, and other key metrics.

Table 14 compares the title guidelines of four major e-commerce platforms reviewed in Section 3 with our proposed guidelines for title creation. For each guideline a 'v' indicates it is included in the platform's existing guidelines. This comparison helps identify key differences and highlights areas for improvement. Some principles, such as including key product attributes and avoiding redundancy and marketing terms are widely adopted across most platforms. Other guidelines, such as including brand, color, and type, ranking attributes by importance, and avoiding all caps are explicitly followed only by some of the platforms. Noticeably, guidelines like using only attribute values, adding units next to numbers, using natural language, and incorporating prepositions for clarity are absent from all platforms, suggesting a lack of emphasis on readability aspects.

[3] Maximum title length in words or characters is platform-specific.

Table 14
Existing title guidelines in e-commerce platforms compared to the proposed title creation guidelines.

Guideline	eBay	Amazon	Walmart	Alibaba
Prioritize key product attributes.	v	v		v
Mention brand, color, and type.		v	v	
Use values, not attribute names.				
Rank attributes by importance.		v		v
Add units next to numbers.				
Use natural language, not keyword stuffing.			v	
Use prepositions for clarity.				
Capitalize properly.		v		
Avoid all caps.	v	v		
Avoid redundancy and marketing terms.	v	v	v	
Maximize attributes within length limit.				

12. Discussion and implications

12.1. Linguistic characteristics of the titles

Our analysis reveals a variety of unique properties exhibited by product titles, which can assist researchers and practitioners when they explore downstream tasks that use titles as their primary source of information. We establish that product titles have only loose language structure and low grammaticality, with supporting evidence spanning the various sections of our analysis. First, titles contain many capitalized tokens, but relatively few punctuation marks and functional words. Furthermore, the majority of title tokens are nouns, especially proper nouns, rarely appearing in the plural form. Verbs, pronouns, determiners, and prepositions are rare in titles. Additionally, our analysis found title sentences exhibit low parsing confidence. Finally, the most common stop words in titles are the connectors '`for`' and '`with`', typically used to feature compatible or complementary products, while other common stop words are not as frequent as in more grammatical text.

In our analysis, we observed that over 80% of the title tokens correspond to product attribute values. Other token types, such as marketing, descriptive statements, and attribute names account for only a small portion of the title. Some attributes, such as `brand`, more commonly appear at the beginning of the title. These observations have implications for title summarization (Mane et al., 2020; Sun et al., 2018), which aims at displaying a shortened version of the title, possibly for presentation on small-screen mobile devices. Since titles consist largely of attribute values, a key part of the task is understanding the relative importance and dependency relations between these attribute values.

Our analysis of token ordering in titles concluded that it is difficult, given a set of title tokens, to accurately predict their relative order in the title. That is, titles are closer, in nature, to an unordered set of attribute tokens than to a grammatical natural language sentence. There is an implication of this observation for practitioners of e-commerce titles. Many common natural language processing (NLP) methods (e.g. word2vec Mikolov et al., 2013) are based on the distributional hypothesis (Firth, 1957), which states that "a word is characterized by the company it keeps". This assumption breaks down somewhat in the regime of unordered sentences such as titles, and thus these methods may be less effective when applied to e-commerce titles.

12.2. Comparison of titles with other e-commerce data

A significant portion of our analysis was devoted to characterizing product titles relative to sentences drawn from other types of texts. On many dimensions, titles were found to be most similar to queries submitted on e-commerce search. Shared characteristics include high noun content, low grammaticality, and similar word choice. Language modeling analysis showed that titles are indeed close to queries, despite being much longer. This makes titles an especially productive facet for product search (Sondhi et al., 2018; Su et al., 2018; Tsagkias et al., 2021), since they are natural to match to user-submitted queries.

Product titles are also comparatively related to sentences from e-commerce descriptions. First, these sentences are of similar length to titles. The distribution of parts-of-speech is somewhat similar (though less so than queries). Token order in descriptions is more predictable than in titles, but less predictable than in product reviews and web natural language. Our language model analysis shows description bigram language models predict titles better than unigram models, alluding to some shared language structure between titles and descriptions. On many measures, description sentences are somewhere in-between natural language sentences and e-commerce titles. One reason for this may be that descriptions are in fact a mix of grammatical natural language text and "title-like" sentence constructions that contain many product attributes. This insight has implications for the design of e-commerce systems, which often treat text from titles and descriptions interchangeably (Ghani et al., 2006; Zheng et al., 2018). Understanding the similarities and differences between these text sources has a potential to improve performance on such applications.

12.3. Implications of titles for downstream tasks and practitioners

We also studied modeling practices for supervised tasks based on product title data. We considered three approaches: explicit feature-based models using boosted decision trees, LSTM networks with pre-trained word embedding, and pre-trained transformer networks (following recent trends Luo et al., 2020; Nguyen et al., 2020; Wang & Fu, 2020). We empirically evaluated these approaches on two supervised learning tasks of varying difficulty, data volume, and labeling method, motivated by real-world application: product title selection and product categorization. The transformer-based approaches dominated in performance on both tasks, with the cased variant more effective for title selection, and uncased variant outperforming for categorization. Our analysis showed that capitalized tokens are extremely widespread on product titles, compared to any other textual dataset; different downstream tasks leveraging titles should therefore carefully consider different case modeling approaches in order to optimized their performance.

Feature importance inspection highlighted several other properties, such as the occurrence of certain attributes and part of speech tags, as especially helpful for downstream tasks. Finally, the unique set of product title characteristics demonstrated throughout this work implies that state-of-the-art NLP methods should be further adapted when used for tasks that heavily rely on titles. For example, fine-tuning a pre-trained model using title data or appropriately prompting a large language model with title-based examples may be necessary to optimize performance. We leave this for further exploration in future work.

Overall, the findings of the paper have several outcomes that impact different usages of titles for e-commerce applications and downstream tasks. Table 15 summarizes the potential implications of the different analyses performed in our study and how they can be applied to e-commerce applications.

Based on our experimentation and analysis, we composed a set of recommended guidelines for writing high-quality titles for products on e-commerce platforms. Having such guidelines is crucial since

Table 15
Potential implications of the analyses for e-commerce applications.

Analysis	Usage
Syntactic Characteristics	We can use POS tags and PCFG parse scores to assess title quality. Titles with unusual syntactic structures (e.g., missing nouns, excessive verbs or adjectives) may indicate spam or low-quality listings. Titles that deviate significantly from the expected POS distribution or have low PCFG scores can be flagged as potential anomalies. This approach helps identify low-quality titles that need revision, while also recognizing high-quality titles that can be used to train language models or support downstream tasks. Additionally, longer titles tend to improve product categorization performance, whereas titles with low parsing scores or a low number of nouns correlate with poorer categorization accuracy.
Lexical Characteristics	Identifying and extracting key attributes from product titles can enhance key e-commerce applications. In search ranking, important attributes (e.g., brand, type, size, material) can be given greater weight to improve retrieval accuracy. In title summarization, prioritizing essential attributes ensures that shortened titles retain the most relevant information, making them more effective for display on small-screen mobile devices. Additionally, in recommendation systems, extracted attributes help suggest similar products based on shared features, enhancing personalization and user experience. Moreover, certain attribute tokens significantly improve categorization accuracy by incorporating domain-specific terminology. Titles containing brand names, colors, and product types perform better, regardless of title length, as these attributes help distinguish between closely related categories. Units of measurement (e.g., '120V' or 'ounce') also serve as strong category indicators. These findings highlight the importance of word choice in product categorization, suggesting that focusing on informative attributes can enhance classification accuracy in e-commerce systems.
Order Predictability	Our findings suggest that product titles do not follow a strict or predictable word order, which has important implications for the effectiveness of language models such as Word2Vec, BERT, and Transformer-based architectures, which rely on contextual relationships between words. These NLP models assume that a word's position within a sequence carries meaning; however, the inconsistent structure of product titles may reduce their effectiveness, making it difficult to capture meaningful representations and potentially lowering the quality of learned embeddings. Additionally, autoregressive models such as GPT-4 and T5, which generate text word-by-word based on learned order sequences, might produce low-quality or inconsistent titles due to the lack of a structured order in training data. This issue could impact tasks such as search ranking, classification, and title generation. To mitigate these challenges, e-commerce applications may benefit from emphasizing keyword-based techniques (e.g., BM25, bag-of-words representations) and graph-based embeddings, which model relationships between entities such as brands, product attributes, and category features, ultimately enhancing performance.
Language Model Similarity	This analysis shows that the language used in product titles closely resembles to that of queries. This finding is consistent with other analyses in this work, which demonstrate that while differing in length, product titles and e-commerce search queries share many common language characteristics. This has important implications for title-query discrimination in search scenarios, where distinguishing between the two is crucial. Users often paste an exact product title into the search prompt, which occurs when the user has navigational intent (Fuchs et al., 2020), i.e. they want to find a particular product, rather than browse through a set of products. Better understanding of this scenario can potentially improve search relevance and the overall e-commerce search experience. For instance, a navigational search for a product title could be tuned to link instantly to a product's page when a title is detected. While product titles share linguistic characteristics with queries, they are relatively different from e-commerce descriptions, even though titles and descriptions are considered interchangeably in some studies (Ghani et al., 2006; Zheng et al., 2018). Our analysis shows that titles primarily consist of attribute values, whereas descriptions are more similar to natural language. This distinction is important for LLM-based generation, as titles may be more challenging to generate fluently due to their loosely structured linguistic form. This has implications for application such as automated product listing.

structured titles improve product visibility, which plays a key role in effective categorization and title selection. Each principle in the guidelines is grounded in our analysis and experimentation results. Since each of the platforms has its set of rules for writing titles, we believe that our proposed guidelines can help refine and revise existing practices while enhancing the user experience in general. While this improvement can contribute particularly to optimizing the downstream tasks (e.g., product categorization), it also has the potential to improve the platform's overall performance in key business metrics (e.g., revenue).

12.4. Future directions

An important direction for future work is the conduction of in-vivo experimentation via A/B test on a real e-commerce platform. This will allow to assess the impact of titles not only on downstream tasks such as product categorization, but can also serve to study titles' impact on overall platform performance. This experimentation can reveal the effect on key business metrics of user engagement such as number of clicks, add-to-carts, and purchases (or, in a more holistic view, the total sale volume on the platform), as well as time spent on the platform, users' return rates, and overall satisfaction, reported directly or inferred from user behavior.

CRediT authorship contribution statement

Sharon Hirsch: Conceptualization, Software, Investigation, Writing – original draft. **Ido Guy:** Conceptualization, Writing – original draft, Supervision. **Slava Novgorodov:** Software, Writing – original draft. **Gal Lavee:** Investigation, Writing – original draft. **Bracha Shapira:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset for the task of product title selection is available here https://drive.google.com/drive/folders/1RsxFzlsvYUUUxuZcoOs3d9MA-UbFNayv?usp=sharing.

References

Alibaba Title Policy (2025). Product name guidelines at alibaba. URL https://activity.alibaba.com/ggs/product_name.html.

Amazon Title Policy (2024). Product title requirements at amazon. URL https://sellercentral.amazon.com/gp/help/external/YTR6SYGFA5E3EQC.

Attardi, G. (2015). Wikiextractor. https://github.com/attardi/wikiextractor,

Baumann, N., Brinkmann, A., & Bizer, C. (2024). Using LLMs for the extraction and normalization of product attribute values. arXiv preprint arXiv:2403.02130.

Bell, A., Senthil Kumar, P., & Miranda, D. (2018). The title says it all: A title term weighting strategy for ecommerce ranking. In *Proc. of CIKM* (pp. 2233–2241).

Brinkmann, A., Shraga, R., & Bizer, C. (2023). Product attribute value extraction using large language models. arXiv preprint arXiv:2310.12537.

Calixto, I., Stein, D., Matusov, E., Castilho, S., & Way, A. (2017). Human evaluation of multi-modal neural machine translation: A case-study on E-commerce listing titles. In *Proc. of vL@eACL* (pp. 31–37).

Camargo de Souza, J. G., Kozielski, M., Mathur, P., Chang, E., Guerini, M., Negri, M., Turchi, M., & Matusov, E. (2018). Generating E-commerce product titles and predicting their quality. In *Proc. of INLG*.

Cevahir, A., & Murakami, K. (2016a). Large-scale multi-class and hierarchical product categorization for an E-commerce giant. In *Proc. of COLING* (pp. 525–535).

Cevahir, A., & Murakami, K. (2016b). Large-scale multi-class and hierarchical product categorization for an E-commerce giant. In *Proc. of COLING* (pp. 525–535).

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proc. of KDD* (pp. 785–794).

Chen, W., Matusov, E., Khadivi, S., & Peter, J.-T. (2016). Guided alignment training for topic-aware neural machine translation. In *Proc. of AMTA* (pp. 121–134).

Chen, J., & Warren, D. (2013). Cost-sensitive learning for large-scale hierarchical classification. In *Proc. of CIKM* (pp. 1351–1360).

Cholakov, N. (2009). Researching product tiles in EBay using the EBay API. In *Proc. of compSysTech*.

Deotte, C., Puget, J.-F., Schifferer, B., & Titericz, G. (2023). Winning amazon KDD cup'23. In *Amazon KDD cup 2023 workshop*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint, arXiv:1810.04805.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. ArXiv Preprint, arXiv:1810.11363.

eBay Title Policy (2024). Title guidelines at eBay. URL https://developer.ebay.com/api-docs/user-guides/static/trading-user-guide/listing-title.html#.

eCom 2018 Data Challenge (2025). SIGIR ecom 2018 data challenge. URL https://sigir-ecom.github.io/ecom2018/data-task.html.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In *Studies in linguistic analysis (special volume of the philological society) 1952–59*, 1–32.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fuchs, G., Acriche, Y., Hasson, I., & Petrov, P. (2020). Intent-driven similarity in E-commerce listings. In *Proc. of CIKM* (pp. 2437–2444).

Ghani, R., Probst, K., Liu, Y., Krema, M., & Fano, A. (2006). Text mining for product attribute extraction. *SIGKDD Explor. Newsl.*, *8*(1), 41–48.

Goumy, S., & Mejri, M.-A. (2018). Ecommerce product title classification. In *ECOM@SIGIR*.

Guy, I. (2016). Searching by talking: Analysis of voice queries on mobile web search. In *Proc. of SIGIR* (pp. 35–44).

Hancock, B., Lee, H., & Yu, C. (2019). Generating titles for web tables. In *The world wide web conference* (pp. 638–647).

Hartley, J. (2005). To attract or to inform: What are titles for? *Journal of Technical Writing and Communication*, *35*, 203–213.

Hasson, I., Novgorodov, S., Fuchs, G., & Acriche, Y. (2021). Category recognition in e-commerce using sequence-to-sequence hierarchical classification. In *Proc. of WSDM* (pp. 902–905).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, S., Liu, X., Peng, X., & Niu, Z. (2012). Fine-grained product features extraction and categorization in reviews opinion mining. In *2012 IEEE 12th international conference on data mining workshops* (pp. 680–686). IEEE.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proc. of ICML* (pp. 2668–2677).

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proc. of ACL* (pp. 423–430).

Kozareva, Z. (2015). Everyone likes shopping! multi-class product categorization for e-commerce. In *Proc. of NAACL-HLT* (pp. 1329–1333).

Krishnan, A., & Amarthaluri, A. (2019). Large scale product categorization using structured and unstructured attributes. ArXiv Preprint, arXiv:1903.04254.

Larasati, L., & Moehkardi, R. R. D. (2019). Unique keywords found in the titles of YouTube beauty and fashion videos. *Lexicon*, *6*(2).

Lee, H., Chang, Y., Choi, K., Lee, J., & Ko, Y. (2023). Statistical and generative models with subtitle extraction for next product title generation. In *Amazon KDD cup 2023 workshop*.

Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. *Royal Society Open Science*, *2*(8), Article 150266.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL* (pp. 7871–7880).

Li, M. Y., Kok, S., & Tan, L. (2018). Don't classify, translate: Multi-level E-commerce product categorization via machine translation. ArXiv Preprint, arXiv:1812.05774.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *cs/0205028*, ArXiv Preprint.

Luo, X., Liu, L., Yang, Y., Bo, L., Cao, Y., Wu, J., Li, Q., Yang, K., & Zhu, K. Q. (2020). AliCoCo: Alibaba e-commerce cognitive concept net. In *Proc. of the SIGMOD* (pp. 313–327).

Mane, M. R., Kedia, S., Mantha, A., Guo, S., & Achan, K. (2020). Product title generation for conversational systems using BERT. ArXiv Preprint, arXiv:2007.11768.

Medelyan, O., Witten, I. H., & Milne, D. (2008). Topic indexing with wikipedia. *1*, In *Proc. of the AAAI wikiAI workshop* (pp. 19–24).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miyamoto, J., Hirano, S., Makino, S., Uekado, K., & Lao, X. (2023). Next product title generation in E-commerce: Rule-based methods and autoencoder model. In *Amazon KDD cup 2023 workshop*.

More, A. (2016). Attribute extraction from product titles in ecommerce. ArXiv Preprint, arXiv:1608.04670.

News Dataset (2022). All the news dataset. URL https://www.kaggle.com/snapcrack/all-the-news.

Nguyen, T. T., Fani, H., Bagheri, E., & Titericz, G. (2017). Bagging model for product title quality with noise.

Nguyen, T. V., Rao, N., & Subbian, K. (2020). Learning robust models for e-commerce product search. ArXiv Preprint, arXiv:2005.03624.

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proc. of EMNLP-IJCNLP* (pp. 188–197).

Nicholson, D., & Paranjpe, R. (2013). A novel method for predicting the end-price of eBay auctions.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL* (pp. 311–318).

Paulucio, L. S., Paixão, T. M., Berriel, R. F., De Souza, A. F., Badue, C., & Oliveira-Santos, T. (2020). Product categorization by title using deep neural networks as feature extractor. In *Proc. of IJCNN* (pp. 1–7).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9–28).

Putthividhya, D., & Hu, J. (2011). Bootstrapped named entity recognition for product attribute extraction. In *Proc. of EMNLP* (pp. 1557–1567).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proc. of ACL demos* (pp. 101–108).

Ristoski, P., Petrovski, P., Mika, P., & Paulheim, H. (2018). A machine learning approach for product matching and categorization. *Semantic Web*, *9*(5), 707–728.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, *88*(8), 1270–1278.

Roy, K., Goyal, P., & Pandey, M. (2021). Attribute value generation from product title using language models. In *Proceedings of the 4th workshop on e-commerce and NLP* (pp. 13–17).

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1), 115–153.

Shah, K., Kopru, S., & Ruvini, J. D. (2018). Neural network based extreme classification and similarity models for product matching. In *Proc. of ACL* (pp. 8–15).

Shen, D., Ruvini, J.-D., & Sarwar, B. (2012). Large-scale item categorization for e-commerce. In *Proc. of CIKM* (pp. 595–604).

Sondhi, P., Sharma, M., Kolari, P., & Zhai, C. (2018). A taxonomy of queries for E-commerce search. In *Proc. of SIGIR* (pp. 1245–1248).

Stein, D., Shterionov, D., & Way, A. (2019). Towards language-agnostic alignment of product titles and descriptions: A neural approach. In *Proc. of WWW (companion)* (pp. 387–392).

Su, N., He, J., Liu, Y., Zhang, M., & Ma, S. (2018). User intent, behaviour, and perceived satisfaction in product search. In *Proc. of WSDM* (pp. 547–555).

Su, T., Macdonald, C., & Ounis, I. (2019). Ensembles of recurrent networks for classifying the relationship of fake news titles. In *Proc. of SIGIR* (pp. 893–896).

Sun, F., Jiang, P., Sun, H., Pei, C., Ou, W., & Wang, X. (2018). Multi-source pointer network for product title summarization. In *Proc. of CIKM* (pp. 7–16).

Symes, C. (1992). You can't judge a book by its cover: The aesthetics of titles and other epitextual devices. *Journal of Aesthetic Education*, *26*(3), 17–26.

Tay, M. (2017). *CIKM analyticup 2017-lazada product title ality challenge: A bag of features for short text classification: Technical Report*, Tech report in School of Computer Science and Engineering, Nanyang ....

Trotman, A., Degenhardt, J., & Kallumadi, S. (2017). The architecture of ebay search. In *ECOM@ SIGIR workshop*.

Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V., & de Rijke, M. (2021). Challenges and research opportunities in ecommerce search and recommendations. *SIGIR Forum*, *54*.

Tzaban, H., Guy, I., Greenstein-Messica, A., Dagan, A., Rokach, L., & Shapira, B. (2020). Product bundle identification using semi-supervised learning. In *Proc. of SIGIR* (pp. 791–800).

Ueffing, N., de Souza, J. G., & Leusch, G. (2018). Quality estimation for automatically generated titles of ecommerce browse pages. In *Proc. of NAACL* (pp. 52–59).

Vasilyev, O., Grek, T., & Bohannon, J. (2019). Headline generation: Learning from decomposable document titles. ArXiv Preprint, arXiv:1904.08455.

Walmart Listing Optimization Guide Listing Quality Optimization guide, xURL https://marketplace.walmart.com/wp-content/uploads/2020/09/WMP_listing_quality_optimization_guide.pdf.

Walmart Title Policy Guidelines & Requirements for Product Title, Description, & Features at Walmart, URL https://sellerhelp.walmart.com/s/guide?article=000006404&language=en_US#producttitle.

Wan, M., & McAuley, J. (2016). Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *Proc. of ICDM* (pp. 489–498).

Wang, T., & Fu, Y. (2020). Item-based collaborative filtering with BERT. In *Proc. of the ECNLP* (pp. 54–58).

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., Yu, Z., & Elsas, J. (2020). Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 47–55).

Wirojwatanakul, P., & Wangperawong, A. (2019). Multi-label product categorization using multi-modal fusion models. ArXiv Preprint, arXiv:1907.00420.

Xia, Y., Levine, A., Das, P., Di Fabbrizio, G., Shinzato, K., & Datta, A. (2017). Large-scale categorization of japanese product titles using neural attention models. In *Proc. of EACL* (pp. 663–668).

Xin, Y., Hart, E., Mahajan, V., & Ruvini, J. D. (2018). Learning better internal structure of words for sequence labeling. In *Proc. of EMNLP* (pp. 2584–2593).

Xu, H., Wang, W., Mao, X., Jiang, X., & Lan, M. (2019). Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proc. of ACL* (pp. 5214–5223).

Yang, B., Liu, F., Li, Z., Yin, Q., You, C., Yin, B., & Zou, Y. (2023). Multimodal prompt learning for product title generation with extremely limited labels. arXiv preprint arXiv:2307.01969.

Yang, L., Wang, Q., Wang, J., Quan, X., Feng, F., Chen, Y., Khabsa, M., Wang, S., Xu, Z., & Liu, D. (2023). Mixpave: Mix-prompt tuning for few-shot product attribute value extraction. In *Findings of the association for computational linguistics: ACL 2023* (pp. 9978–9991).

Yang, L., Wang, Q., Yu, Z., Kulkarni, A., Sanghai, S., Shu, B., Elsas, J., & Kanagal, B. (2022). Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining* (pp. 1256–1265).

Yıldırım, A., Üsküdarlı, S., & Özgür, A. (2016). Identifying topics in microblogs using wikipedia. *PloS One*, *11*(3), Article e0151885.

Zheng, G., Mukherjee, S., Dong, X. L., & Li, F. (2018). Opentag: Open attribute value extraction from product profiles. In *Proc. of KDD* (pp. 1049–1058).

Zuze, H., & Weideman, M. (2013). Keyword stuffing and the big three search engines. *Online Information Review*.