ENTREPRENEURSHIP AND SUSTAINABILITY ISSUES
ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

Publisher
http://jssidoi.org/esc/home

INFORMATION TECHNOLOGY FOR INTELLECTUAL ANALYSIS OF ITEM DESCRIPTIONS IN E-COMMERCE*

Olga Cherednichenko [1], Oksana Ivashchenko [2], Marcel Lincenyi [3], Marián Kováč [4]

[1,2,3,4] Bratislava University of Economics and Management, Furdekova 16, 851 04 Bratislava, Slovakia

E-mails: [1] olga.cherednichenko@vsemba.sk; [2] oksana.ivashchenko@vsemba.sk; [3] marcel.lincenyi@vsemba.sk; [4] marian.kovac@vsemba.sk

Received 14 May 2023; accepted 11 September 2023; published 30 September 2023

Abstract. E-commerce is experiencing a robust surge, propelled by the worldwide digital transformation and the mutual advantages accrued by both consumers and merchants. The integration of information technologies has markedly augmented the efficacy of digital enterprise, ushering in novel prospects and shaping innovative business paradigms. Nonetheless, adopting information technology is concomitant with risks, notably concerning safeguarding personal data. This substantiates the significance of research within the domain of artificial intelligence for e-commerce, with particular emphasis on the realm of recommender systems. This paper is dedicated to the discourse surrounding the construction of information technology tailored for processing textual descriptions pertaining to commodities within the e-commerce landscape. Through a qualitative analysis, we elucidate factors that mitigate the risks inherent in unauthorized data access. The cardinal insight discerned is that the apt utilization of product matching technologies empowers the formulation of recommendations devoid of entailing customers' personal data or vendors' proprietary information. A meticulously devised structural model of this information technology is proffered, delineating the principal functional components essential for processing textual data found within electronic trading platforms. Central to our exposition is the exploration of the product comparison predicated on textual depictions. The resolution of this challenge stands to enhance the efficiency of product searches and facilitate product juxtaposition and categorization. The prospective implementation of the propounded information technology, either in its entirety or through its constituent elements, augurs well for sellers, enabling them to improve a pricing strategy and heightened responsiveness to market sales trends. Concurrently, it streamlines the procurement journey for buyers by expediting the identification of requisite goods within the intricate milieu of e-commerce platforms.

Keywords: Information Technology; e-Commerce; Product Matching; Text Processing; Model; Artificial Intelligence

Reference to this paper should be made as follows: Cherednichenko, O., Ivashchenko, O., Lincenyi, M. Kováč, M. 2023. Information technology for intellectual analysis of item descriptions in e-commerce, *Entrepreneurship and Sustainability Issues*, 11(1), 178-190. http://doi.org/10.9770/jesi.2023.11.1(10)

JEL Classifications: M15, O30, C89

Additional disciplines: information and communication; informatics

* *This research is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V01-00078 and the project No. 09I03- 03-V01-00080*

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

1. Introduction

The concept of business digital transformation signifies a pivotal shift wherein digital technologies assume a central role in shaping business operations and broader societal changes (Vaska et al., 2021). Among these transformative technologies, the Internet of Things (IoT) and Artificial Intelligence (AI) stand out as game-changers, wielding a profound impact on the trajectory of business development (Caputo, 2021). Our research is particularly attuned to enhancing product matching processes within the e-commerce sector, recognizing its far-reaching implications for customers and sellers. As underscored by calculations provided by Statista Market Insights in 2023, the compound annual growth rate (CAGR) of retail e-commerce sales is anticipated to maintain an impressive trajectory, projected at a robust 11.16 percent from 2023 to 2027. This surge in e-commerce growth is emblematic of the broader digital transformation sweeping across industries, further emphasizing the significance of optimizing processes like product matching to meet the evolving demands of the digital age.

The rapid growth of the e-commerce sector has ushered in a steep surge in competition, compelling sellers to embrace information technologies and innovations within their business processes. According to recent statistics (Digital Transformation - Statistics & Facts, 2023), three pivotal technologies are steering the course of digital transformation across various economic industries: big data (64%), the cloud (50%), and artificial intelligence (AI) (44%). In recent works (Kumar & Rn, 2019; Bharadiya, 2023), noteworthy trends have emerged regarding utilizing information technologies in e-commerce and the broader business landscape. Chatbots and virtual assistants have emerged as instrumental tools, streamlining customer communication, answering common queries, disseminating seller information, and facilitating smoother transactions. Furthermore, integrating applications offering augmented reality and virtual reality (AR and VR) during the product purchasing process has proven to be a strategic approach to maintaining competitiveness in e-commerce (Magnolia Market, 2023). Additionally, web scraping bots and web scraping APIs have empowered businesses to extract valuable data from websites (Bright Data, 2023; Oxylabs, 2023; Smartproxy, 2023).

Machine learning and AI techniques have emerged as invaluable assets, automating tasks related to data analysis, transformation, and feature extraction, thereby enhancing the decision-making processes and overall business operations. These technologies have been harnessed to collect and interpret customer preferences, transaction data, reviews, and estimations, paving the way for personalized customer support and targeted marketing and advertising campaigns. Notably, Amazon's recommender systems have a rich history (Smith & Linden, 2017; Amazon Science, 2019), with a substantial 35% of customer purchases attributed to their influence, as reported by McKinsey and Company in 2013. Furthermore, contemporary advancements in this field have led to intriguing developments, such as Amazon's Personalization team opting to base learning on a product-level buying history, yielding superior results compared to consumer-level histories. In this intensely competitive e-commerce landscape, integrating these technologies and strategies is pivotal for sellers aiming to thrive and excel in a rapidly evolving digital marketplace.

The challenge of processing product descriptions has been a longstanding concern, with effective solutions contingent upon the quality and variety of available text data. This paper seeks to develop an adaptable and versatile information technology framework for product matching by processing item descriptions. Our approach entails the construction of this information technology as a series of flexible pipelines designed to facilitate product searching, categorization, and matching.

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

2. Theoretical background

Researchers, as noted by (Chatterjee, 2015), define e-commerce as a mode of business where parties engage in transactions over the Internet utilizing various means and technologies. However, the digitalization of business, including e-commerce, brings with it a set of advantages and disadvantages. One notable disadvantage of e-commerce, as highlighted by (Taher, 2021), revolves around security and privacy concerns. Recently, even reputable companies have fallen prey to scam attacks, losing valuable customer data. This issue is further underscored in the work of (Jamra et al., 2020), where the most prevalent security issues in e-commerce encompass credit card fraud, cyberattacks, and threats to sensitive information. The e-commerce transaction process, as elucidated by (Chatterjee, 2015), comprises several intricate steps, each necessitating security and privacy measures.

Additionally, these steps often require the collection of personal customer information. Without robust safety measures, the risks of data breaches and unauthorized use of personal information loom large. Therefore, detecting and managing risks associated with security and privacy are integral components of effective online platform management, yielding benefits for both customers and sellers.

Beyond the sphere of safety and privacy, (Guru et al., 2020), drawing from a review of research papers, identifies three primary categories of perceived risks in e-commerce: performance risk, financial risk, and time-loss risk. Perceived risk characterizes the uncertainty customers experience when making purchasing decisions. Performance risk pertains to whether products meet customers' expectations in terms of functionality. Financial risk relates to whether the quality of online services justifies the monetary investment made by customers. On the other hand, a time-loss risk emerges when customers are dissatisfied with the vast array of products returned in response to their queries, often due to poor product matching, delivery issues, or complicated return procedures. Consequently, time-loss risk can be defined as the time required to purchase, return, or exchange items when customers find them unsatisfactory.

Our suite of information technology solutions assumes a consequential role in mitigating these risks and preserving the integrity of customer data. We propose an information technology for the intelligent analysis of item descriptions, with the primary objective of addressing the inherent temporal inefficiencies in e-commerce operations. Our system components are intentionally designed to refrain from storing or processing personal customer data, thus mitigating the susceptibility to data breaches. In the context of mitigating time-loss risk, our technology excels at matching product features to search queries and recommending products that align with customer preferences. These matching results undergo meticulous calibration to optimize precision and relevance, expediting the shopping process and generating substantial time and cost savings for users. This functionality fosters elevated trust among customers in their engagements with online platforms.

The subject of item matching has been a subject of previous investigation by researchers (Appel et al., 2022; Peeters et al., 2020; Zuo et al., 2020; Strauß et al., 2019; Akritidis et al., 2019; Akritidis & Bozanis, 2018; Mudgal et al., 2018; Kannan et al., 2011; Köpcke et al., 2012; Gopalakrishnan et al., 2012; Zheng, & Sun, 2022). Typically, this involves categorizing attributes through utilizing specific similarity functions designed to process these attributes (Strauß et al., 2019; Łukasik, 2021). An alternative approach for matching unstructured product offerings involves semantic processing of item descriptions (Shah et al., 2018; Ristoski et al., 2018, Nigam et al., 2019; Singh & Shashi, 2019). The central task in this context is addressed as a binary classification problem, where pairs of product offers are assessed to determine whether they describe the same or similar products (Ristoski, 2018, Strauß et al., 2019; Akritidis et al., 2019; Łukasik et al., 2021). It is worth noting that product offers are commonly presented on trading platforms as textual descriptions and specification tables.

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

A diverse range of techniques is employed to tackle the challenge of product matching, spanning natural language processing and machine learning methodologies. There has been a discernible shift in the development of tools and algorithms in recent years. There is a notable transition from traditional statistical and machine learning methods toward adopting deep learning techniques, particularly those rooted in transformer architectures and language models (Peeters et al., 2020; Li, 2020; Ye, 2022). This paradigm shift has resulted in a substantial body of research focused on the implementation of end-to-end entity resolution tasks (Konda, 2018; Konda et al., 2016; Mudgal et al., 2018; Dou et al., 2023; Wang et al., 2021; Peeters et al., 2020, Primpeli & Bizer, 2021). The entire process is consolidated into a single integrated model in this context. Additionally, considerable research efforts have been directed toward the creation and refinement of benchmark datasets tailored for evaluating the efficacy of various product matching approaches (Bizer et al., 2019, Crescenzi et al., 2021; Foxcroft et al., 2021, Peeters et al., 2023, Primpeli et al., 2019; Primpeli & Bizer 2020; Wang et al., 2022). These benchmarked datasets are fundamental for model training and evaluation, facilitating robust comparisons and methodological analyses.

Our research addresses the intricate challenge of item matching within the context of e-commerce platforms. We observe the proliferation of identical physical items offered across multiple e-commerce platforms in the contemporary landscape. Moreover, the same physical item may be presented with disparate descriptions and associated conditions even within a single platform. Consequently, locating a specific item has become a time-intensive undertaking for both customers and sellers. In response, we propose an innovative information technology solution for product matching. This solution amalgamates state-of-the-art approaches in entity resolution, natural language processing, and constructing a flexible data analysis pipeline.

3. Research objective and methodology

Text mining is concerned with extracting structured information from unstructured text collections. The methodologies employed in text mining necessitate task-specific text analysis processes, often comprising multiple interdependent steps. Typically, these processes are implemented through the creation of text analysis pipelines. A significant challenge arises from the fact that these pipelines are predominantly constructed manually, demanding expertise in their design. In addressing the problem of product matching, methods for processing item descriptions can be categorized into distinct stages, encompassing preliminary data processing, tokenization, and the application of specialized processing steps such as clustering and classification. These processing steps collectively constitute the pipeline.

Conceptually, a text analysis pipeline can be represented as a tuple comprising a set of text analysis algorithms and a schedule that dictates the sequence in which these algorithms are applied (Wachsmuth et al., 2013). Each algorithm within the set is responsible for a specific text analysis task, generating information of predefined types as output. To operate effectively, each algorithm relies on the input text and information of specific types, which preceding algorithms must produce within the pipeline. Consequently, the schedule must ensure the fulfilment of input requirements for all algorithms.

The primary objective of text analysis pipelines is to process input texts, transforming them into structured information relevant to specific needs. Consequently, a pipeline's core text analysis task can be summarized as follows: Given a collection or stream of input texts, process them to deduce structured information. The nature of the input texts can range from a confined set of texts in a specific domain to an ongoing stream of open-domain texts from the web. Notably, the composition of text analysis algorithms is inherently task-specific. Thus, when confronted with a text analysis task, a text pipeline is predefined by selecting and scheduling a suitable subset of available text analysis algorithms.

Our approach presents a systematic sequence of steps for processing product text descriptions to identify similar or identical products through machine learning algorithms (see Fig. 1). The initial input data for this process

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

encompasses product titles and product text descriptions. In the first step, we focus on the preprocessing, vectorizing, and clustering of the product text descriptions. At this stage, the user has the flexibility to select the most suitable vectorization and clustering models, considering the research objectives and the inherent characteristics of the data. This step serves the dual purpose of reducing the dataset size and generating groups of text descriptions that correspond to similar products. These text description groups serve as the input for the subsequent step.

Moving to the next stage, the second step, we create core tags for each of these groups. These core tags comprise sets of words that encapsulate essential information about a given group of similar products, providing a succinct yet informative description of the group. This step encompasses several substeps, including preprocessing to generate core tags for each similar product group, cleansing to eliminate duplicates and creating core tags through utilizing a word2vec model in terms of a similarity metric. In the final phase, we generate a reference description for similar products by implementing a model based on a reinforcement learning approach. This reference description is a consolidated representation of the identified similar products, streamlining the understanding and communication of their collective attributes and characteristics.

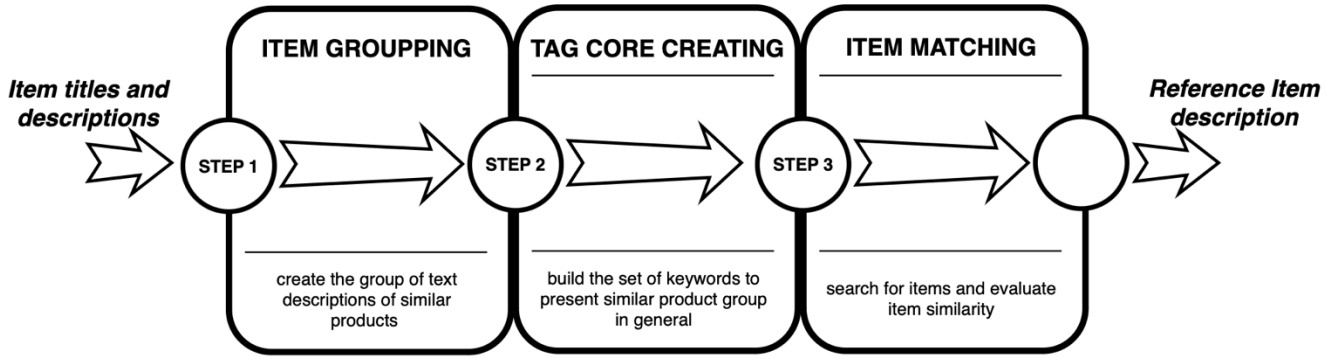

Figure 1. General pipeline for processing product text descriptions

Source: Own work

To establish a functional pipeline, it is imperative to understand its definition, components, and the procedures required for effective deployment. Within the realm of computing systems, a pipeline signifies a structured sequence of operations through which instructions for sequential data processing are systematically passed. It constitutes a methodical framework for the orderly storage, placement, and consecutive data transmission (Wachsmuth, 2015). A distinguishing characteristic of a pipeline lies in the fact that the output generated by the processing of one function serves as the input for the subsequent function in a seamless and interconnected manner.

We implement a data processing pipeline, encompassing critical stages such as data preprocessing, filtering, tokenization, vectorization, and clustering. In this endeavor, we are considering various clustering methods, including well-established techniques such as K-means and SVM (Ahmed et al., 2020; Winters-Hilt & Merat, 2007; Yao et al., 2013; Korovkinas et al., 2019), owing to their widespread applicability in tasks of this nature. Vectorization, a pivotal component of our pipeline, involves converting textual data into numeric representations that encode their underlying meaning. While several vectorization approaches are available, we have opted for the Word Embeddings method, leveraging the Word2Vec model due to its demonstrable superiority in terms of accuracy (Mikolov, Chen et al., 2013; Mikolov et al., 2013).

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

We have chosen the WDC Product Data Corpus for experimentation and validation, accessible at http://webdatacommons.org, as our dataset. We are harnessing publicly available Python libraries, including spaCy, gensim, and scikit-learn to facilitate the implementation process. Our software implementation adheres to a structured Python library format comprising modules that can be seamlessly integrated into other applications. This software solution is designed to serve as an adaptable component, suitable for internal and external deployment within various systems. At its core, it features a pipeline controller responsible for orchestrating interactions with preprocessing processors, a token processor that interfaces with the vector model, and a main processor dedicated to data clustering and classification. This design separates processing into distinct modules, ensuring a flexible and versatile framework for data processing.

The main goal of information technology development is to continually enhance and innovate the capabilities, efficiency, and utility of digital tools and systems to meet evolving technological, business, and societal needs. The main objective of our research is to optimize the analysis and understanding of textual product descriptions, enabling accurate identification, categorization, and matching of products while enhancing user experiences in e-commerce and related applications. To achieve this goal, we establish the information technology tailored for processing textual descriptions pertaining to commodities within the e-commerce landscape.

In the context of product description processing within information technology development, there are the following research questions:

RQ1. How can machine learning algorithms, such as Word Embeddings or Word2Vec, be utilized to generate informative vector representations of product descriptions?

RQ2. What techniques and models are most effective in improving product categorization and matching?

RQ3. How can natural language processing techniques be leveraged to improve the accuracy and efficiency of product matching based on textual descriptions?

4. Results and discussion

The primary objective is to enhance the generality and flexibility of product description processing. Given that we primarily utilize textual representations for product descriptions, applying natural language processing techniques is a promising avenue. It's essential to acknowledge that product descriptions exhibit distinct characteristics. For instance, product names often incorporate terms that may not directly pertain to the product itself (e.g., "new," "incredible," "innovative," etc.). Additionally, product descriptions may undergo automatic translation into English, enabling the detection of non-English words within titles and descriptions.

Moreover, descriptions of product offerings on e-commerce platforms frequently contain abbreviations, size indicators, brand names, model references, product codes, and more. These peculiarities lead to two significant observations: firstly, text description processing necessitates a specialized preprocessing phase, and secondly, traditional natural language processing methods may not yield the desired outcomes. This type of text resembles more of a raw collection of keywords rather than a semantic representation describing a specific product.

Furthermore, it's worth noting that in many instances, product attributes could be more present, complete, or riddled with errors. Extracting and organizing these attributes can significantly enhance the effectiveness of addressing the product-matching challenge. Therefore, natural language processing techniques can be effectively employed to preprocess and cleanse text descriptions, extract values associated with product attributes, and generate a comprehensive set of keywords that encapsulate product descriptions.

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

Given that we process data in textual format, the utilization of embedding methods emerges as a logical choice. We have identified two key tasks for vectorizing text data based on our prior expertise (Cherednichenko et al., 2023) and a wealth of conducted experiments. Firstly, we aim to represent all product descriptions as a collection of keywords that collectively encapsulate the consumer attributes of the product. We assert that keywords associated with a single product should exhibit semantic proximity. To accomplish this, we endorse the approach proposed by (Cherednichenko et al., 2023), which involves tokenizing and cleansing each text description, followed by representing each token (keyword) as a vector. For this purpose, we employ a pre-trained Word2Vec model for the English language from spaCy library. However, it's essential to address the issue of encountering words in the descriptions the model does not recognize. To resolve this, one potential approach is to independently train the model, albeit this entails significant additional resources. Alternatively, if the number of unrecognized words is minimal, they can be eliminated during the preprocessing stage.

The second vectorization task revolves around representing a comprehensive product description in vector format. In this scenario, we posit that two identical products offered by different sellers should yield vector representations that are proximate within the defined vector space. To achieve this, we leverage the available Doc2Vec model from the gensim library. The vector representations of descriptions enable the application of clustering techniques for categorising and grouping similar products or those sharing analogous descriptions.

Given our approach of transforming each textual product representation from an e-commerce platform into a collection of keywords that closely correspond to the product's description, we posit that a cluster of similar products can similarly be represented by a shared set of keywords that collectively characterize the group. We refer to this set as a "tag core," defined as a collection of keywords that semantically depict a group of identical or highly similar products. The threshold for inclusion in the tag core is determined empirically for each cluster of similar products, employing the cosine similarity measure. Thus, we employ a flexible and customizable pipeline (Cherednichenko et al., 2023) for processing descriptions of such product clusters, culminating in the construction of a tag core. The input for this task comprises sets of keywords derived from each product description within the group of similar products, while the output is the resultant tag core.

In summary, building upon the findings in (Cherednichenko, 2023), we propose an information technology framework for processing textual descriptions of products (see Figure 2). This framework streamlines the automation of the entire process, commencing with data collection and preliminary preprocessing of product information sourced from e-commerce platforms and culminating in resolving challenges related to product grouping, categorization, and comparison. The reusability of models and software solutions will enable the construction of customized processing workflows for product descriptions, accounting for the distinct characteristics of product categories and processing objectives. This approach minimizes the labor-intensive research, modeling, design, and approach validation aspects for product comparison tasks. In an extension of our work as presented in (Cherednichenko et al., 2023), we suggest incorporating a crowdsourcing feature to gather data from potential e-commerce customers. Additionally, we emphasize the functions of product matching and the creation of reference descriptions, as illustrated in Figure 2.

The proposed information technology encompasses a suite of functions designed to process textual descriptions of e-commerce products. This technology relies on a family of models encompassing tasks such as data cleansing, tokenization, vectorization, clustering, and classification of texts. These functions are realized through software implementation in the form of flexible custom pipelines, effectively furnishing a mechanism to address the challenges associated with processing textual descriptions. The chosen functions are pivotal in streamlining the preparatory phase of work, which typically demands substantial effort and labor. This phase encompasses activities such as collecting textual descriptions, cleansing them, grouping them, and generating tag cores, ultimately resulting in standardized descriptions for product categories that interest users. This standardization

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

process structures vast datasets and significantly reduces the time required for subsequent product search and matching.

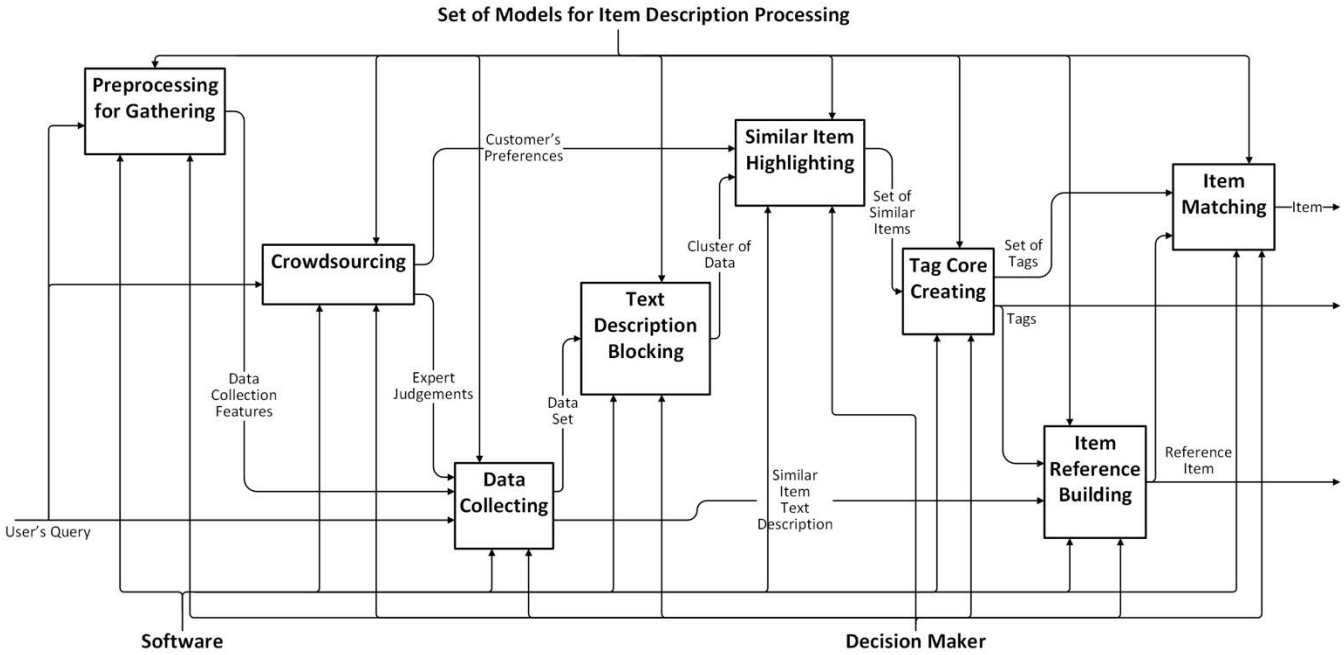

Figure 2. Functional Model of Information technology

Source: Own work

Answering the research questions, we can draw the following conclusions. Addressing the first research question, it is evident that machine learning algorithms, such as Word Embeddings or Word2Vec, prove highly effective in representing keywords within a semantic vector space. These algorithms enable the construction of tag cores for groups of similar products and facilitate clustering of textual descriptions, leading to the formation of product groups. In response to the second research question, we have concluded that to achieve effective categorization and comparison of goods, it is essential to construct a reference description. This reference description serves as the basis for comparison and relies on the cosine similarity measure within the semantic space of product description tags. Regarding the third research question, our findings suggest that while natural language processing technologies may not be ideal for addressing the direct problem of product comparison, techniques such as preprocessing, stop word removal, tokenization, and vectorization have proven to be valuable and can be seamlessly integrated into the overarching information technology.

5. Conclusions

E-commerce, as a mode of business, operates in a realm untethered from physical presence, eschewing the traditional brick-and-mortar establishments and direct face-to-face interactions between sellers and customers (Andonov et al., 2021). In the digital epoch, online platforms are steadily evolving into fiercely competitive arenas, harnessing a diverse array of technologies and innovations. The strategic integration of cutting-edge technologies rooted in artificial intelligence (AI) and machine learning has empowered online retailers to revolutionize their service offerings. The deployment of chatbots, virtual assistants, and recommender systems caters to a more personalized customer experience, delivering both cost-efficient and time-saving services.

ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

Sellers, too, reap the rewards of information technologies. These technologies are invaluable assets for analyzing vast troves of big data, facilitating precise predictions, streamlining decision-making processes, and automating routine business operations and customer interactions. In essence, information technologies have become indispensable allies in the realm of e-commerce, propelling it into the future of business.

The proposed information technology is crucial in safeguarding customer data integrity and streamlining e-commerce efficiency. We offer an intelligent item description analysis solution designed to combat time-related inefficiencies. Our system prioritizes data security by abstaining from personal customer data storage, reducing the risk of breaches. In terms of time efficiency, our technology excels at matching products to search queries based on customer preferences, leading to precise and expedited shopping experiences. This functionality not only saves users time and money but also enhances trust in online platforms

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Amazon Science. The history of Amazon's recommendation algorithm. (2019). https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

Andonov, A., Dimitrov, G. P., & Totev, V. (2021). Impact of E-commerce on Business Performance. *TEM Journal*, 10(4), 1558. https://doi.org/10.18421/TEM104-09

Akritidis, L., & Bozanis, P. (2018). Effective unsupervised matching of product titles with k-combinations and permutations. In 2018 Innovations in Intelligent Systems and Applications (INISTA) (pp. 1-10). https://doi.org/10.1109/INISTA.2018.8466294

Akritidis, L., Fevgas, A., Bozanis, P., & Makris, C. (2019). A Clustering-Based Combinatorial Approach to Unsupervised Matching of Product Titles. arXiv preprint arXiv:1903.04276. https://doi.org/10.1007/s10462-020-09807-8

Appel, A. P., Silva, A. L. D. P., Silva, A. R., Santos, C. D., da Silva, T. L., de Araujo, R. P., & de Aquino, L. C. F. (2022). Item Matching using Text De-scription and Similarity Search. arXiv preprint arXiv:2206.14097.

Bharadiya, J. P. (2023) Machine Learning and AI in Business Intelligence: Trends and Opportunities. *International Journal of Computer (IJC)*, 48(1), 123-134.

Bizer, C., Primpeli, A., & Peeters, R. (2019). Using the semantic web as a source of training data. *Datenbank-Spektrum*, 19, 127-135. https://doi.org/10.1007/s13222-019-00313-y

Bright Data. (2023). https://get.brightdata.com/data-collector7388?sid=web-scraping-tools

Caputo, A., Pizzi, S., Pellegrini, M. M., & Dabić, M. (2021). Digitalization and business models: Where are we going? A science map of the field. *Journal of Business Research*, https://doi.org/10.1016/j.jbusres.2020.09.053

Chatterjee, S. (2015). Security and privacy issues in E-Commerce: A proposed guidelines to mitigate the risk. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 393-396). IEEE. https://doi.org/10.1109/IADCC.2015.7154737

Cherednichenko, O., Ivashchenko, O., Cibák, Ľ., & Lincenyi, M. (2023). Item Matching Model in E-Commerce: How Users Benefit. *Economics and Culture*, 20(1), 77-90. https://doi.org/10.2478/jec-2023-0007

Crescenzi, V., De Angelis, A., Firmani, D., Mazzei, M., Merialdo, P., Piai, F., & Srivastava, D. (2021). Alaska: A flexible benchmark for data integration tasks. arXiv preprint arXiv:2101.11259. https://doi.org/10.48550/arXiv.2101.11259

Digital Transformation - Statistics & Facts. (2023). https://www.statista.com/topics/6778/digital-transformation/#topicOverview

Dou, W., Shen, D., Zhou, X., Nie, T., Kou, Y., Cui, H., & Yu, G. (2023). Soft Target-Enhanced Matching Framework for Deep Entity Matching. https://doi.org/10.1609/aaai.v37i4.25544

Foxcroft, J., Chen, T., Padmanabhan, K., Keng, B., & Antonie, L. (2021). Product Matching Lessons and Recommendations from a Real World Application. In Canadian Conference on AI. https://doi.org/10.21428/594757db.08c5079e

Gopalakrishnan, V., Iyengar, S. P., Madaan, A., Rastogi, R., & Sengamedu, S. (2012). Matching product titles using web-based enrichment. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 605-614). https://doi.org/10.1145/2396761.2396839

Guru, S., Nenavani, J., Patel, V., & Bhatt, N. (2020). Ranking of perceived risks in online shopping. *Decision*, 47, 137-152. https://doi.org/10.1007/s40622-020-00241-x

Jamra, R. K., Anggorojati, B., Sensuse, D. I., & Suryono, R. R. (2020). Systematic Review of Issues and Solutions for Security in E-commerce. In 2020 International Conference on Electrical Engineering and Informatics (ICELTICs) (pp. 1-5). IEEE. https://doi.org/10.1109/ICELTICS50595.2020.9315437

Kannan, A., Givoni, I. E., Agrawal, R., & Fuxman, A. (2011). Matching unstructured product offers to structured product specifications. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 404-412). https://doi.org/10.1145/2020408.2020474

Konda, P. V. (2018). Magellan: Toward building entity matching management systems. The University of Wisconsin-Madison.

Konda, P., Das, S., Doan, A., Ardalan, A., Ballard, J. R., Li, H., ... & Raghavendra, V. (2016). Magellan: toward building entity matching management systems over data science stacks. Proceedings of the VLDB Endowment, 9(13), 1581-1584. https://doi.org/10.14778/3007263.3007314

Korovkinas, K., Danenas, P., & Garšva, G. (2019). SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis. *Baltic Journal of Modern Computing*, 7(1). https://doi.org/10.22364/bjmc.2019.7.1.04

Kumar, D. R., & Rn, R. (2019). A study on E–commerce trends and its advantages in digital era. *International Journal of Research and Analytical Reviews (IJRAR)*, 6(2), 276-281.

Köpcke, H., Thor, A., Thomas, S., & Rahm, E. (2012). Tailoring entity resolution for matching product offers. In Proceedings of the 15th international conference on extending database technology (pp. 545-550). https://doi.org/10.1145/2247596.2247662

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W. C. (2020). Deep entity matching with pre-trained language models. arXiv preprint arXiv:2004.00584. https://doi.org/10.48550/arXiv.2004.00584

Łukasik, S., Michałowski, A., Kowalski, P. A., & Gandomi, A. H. (2021, June). Text-Based Product Matching with Incomplete and Inconsistent Items Descriptions. In International Conference on Computational Science (pp. 92-103). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-77964-1_8

Magnolia Market. (2023). https://apps.apple.com/us/app/magnolia-market/id1263517500

McKinsey and company. (2013). How retailers can keep up with consumers. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., ... & Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In Proceedings of the 2018 International Conference on Management of Data (pp. 19-34). https://doi.org/10.1145/3183713.3196926

Nigam, P., Song, Y., Mohan, V., Lakshman, V., Ding, W., Shingavi, A., ... & Yin, B. (2019). Semantic product search. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2876-2885). https://doi.org/10.1145/3292500.3330759

Oxylabs. (2023). https://oxylabs.go2cloud.org/aff_c?offer_id=7&aff_id=845&url_id=86

Peeters, R., Bizer, C., & Glavaš, G. (2020). Intermediate training of BERT for product matching. *Small*, 745(722), 2-112.

Peeters, R., Der, R. C., & Bizer, C. (2023). WDC Products: A Multi-Dimensional Entity Matching Benchmark. arXiv preprint arXiv:2301.09521. https://doi.org/10.48550/arXiv.2301.09521

Peeters, R., Primpeli, A., Wichtlhuber, B., & Bizer, C. (2020). Using schema. org annotations for training and maintaining product matchers. In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (pp. 195-204). https://doi.org/10.1145/3405962.3405964

Primpeli, A., & Bizer, C. (2020). Profiling entity matching benchmark tasks. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 3101-3108). https://doi.org/10.1145/3340531.3412781

Primpeli, A., & Bizer, C. (2021). Graph-boosted active learning for multi-source entity resolution. In The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20 (pp. 182-199). Springer International Publishing. https://doi.org/10.1007/978-3-030-88361-4_11

Primpeli, A., Peeters, R., & Bizer, C. (2019). The WDC training dataset and gold standard for large-scale product matching. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 381-386). https://doi.org/10.1145/3308560.3316609

Ristoski, P., Petrovski, P., Mika, P., & Paulheim, H. (2018). A machine learning approach for product matching and categorization. *Semantic Web*, 9(5), 707-728. https://doi.org/10.3233/SW-180300

Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10(7), 305-310. https://doi.org/10.14569/ijacsa.2019.0100742

Shah, K., Kopru, S., & Ruvini, J. D. (2018). Neural network based extreme classification and similarity models for product matching. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers) (pp. 8-15). https://doi.org/10.18653/v1/N18-3002

Smartproxy. (2023). http://smartproxy.pxf.io/DVY1yy

Smith, B., & Linden G. (2017). Two decades of recommender systems at Amazon.com. IEEE Internet Computing. https://doi.org/10.1109/MIC.2017.72

Statista Market Insights. (2023). E-commerce retail sales CAGR 2023-2027, by country. https://www.statista.com/forecasts/220177/b2c-e-commerce-sales-cagr-forecast-for-selected-countries

Strauß, O., Almheidat, A., & Kett, H. (2019). Applying Heuristic and Machine Learning Strategies to Product Resolution. In WEBIST (pp. 242-249). https://doi.org/10.5220/0008069402420249

Taher, G. (2021). E-commerce: advantages and limitations. *International Journal of Academic Research in Accounting Finance and Management Sciences*, 11(1), 153-165. https://doi.org/10.6007/IJARAFMS /v11-i1/8987

Vaska, S., Massaro, M., Bagarotto, E. M., & Dal Mas, F. (2021). The digital transformation of business model innovation: A structured literature review. *Frontiers in Psychology*, 11, https://doi.org/10.3389/fpsyg.2020.539363

Wachsmuth, H. (2015) Text Analysis Pipelines. Towards Ad-hoc Large-Scale Text Mining. https://doi.org/10.1007/978-3-319-25741-9

Wachsmuth, H., Rose, M., & Engels, G. (2013). Automatic pipeline construction for real-time annotation. In Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I 14 (pp. 38-49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37247-6_4

Wang, P., Zheng, W., Wang, J., & Pei, J. (2021). Automating entity matching model development. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 1296-1307). IEEE. https://doi.org/10.1109/ICDE51399.2021.00116

Wang, T., Lin, H., Fu, C., Han, X., Sun, L., Xiong, F., ... & Zhu, X. (2022). Bridging the gap between reality and ideality of entity matching: A revisiting and benchmark re-construction. arXiv preprint arXiv:2205.05889. https://doi.org/10.48550/arXiv.2205.05889

Winters-Hilt, S., & Merat, S. (2007). SVM clustering. *BMC Bioinformatics*, 8(7), 1-12. BioMed Central. https://doi.org/10.1186/1471-2105-8-S7-S18

Yao, Y., Liu, Y., Yu, Y., Xu, H., Lv, W., Li, Z., & Chen, X. (2013). K-SVM: An Effective SVM Algorithm Based on K-means Clustering. *J. Comput*., 8(10), 2632-2639. https://doi.org/10.4304/jcp.8.10.2632-2639

Ye, C., Jiang, S., Zhang, H., Wu, Y., Shi, J., Wang, H., & Dai, G. (2022). JointMatcher: Numerically-aware entity matching using pre-trained language models with attention concentration. *Knowledge-Based Systems*, 251, https://doi.org/10.1016/j.knosys.2022.109033

Zheng, X., & Sun, A. (2022). Digitalization and Internationalization: A Study of the Manufacturing Industry in China. *Transformations in Business & Economics*, Vol. 21, No 2B (56B), pp.772-791.

Zuo, Z., Wang, L., Momma, M., Wang, W., Ni, Y., Lin, J., & Sun, Y. (2020). A flexible large-scale similar product identification system in e-commerce. In KDD Workshop on Industrial Recommendation Systems.

Funding: This research is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V01-00078 and the project No. 09I03- 03-V01-00080

Author Contributions: Conceptualization: *Cherednichenko O., Kováč M.*; methodology: *Cherednichenko O., Ivashchenko O.*; data analysis: *Ivashchenko O., Lincenyi M.*, writing—original draft preparation: *Cherednichenko O., Ivashchenko O.;* writing; review and editing: *Cherednichenko O., Lincenyi M.*; visualization: *Cherednichenko O., Kováč M.*. All authors have read and agreed to the published version of the manuscript

Dr. Olga CHEREDNICHENKO, DSc is an Postdoctorant at the Department of Management and Marketing of Bratislava University of Economics and Management (Slovakia), Professor at the Department of Software Engineering and Management Intelligent Technologies of National Technical University “Kharkiv Polytechnic Institute" (Ukraine). Research interests: the mathematical models of intelligent systems with focus on the multi-agents systems. E-mail: olga.cherednichenko@khpi.edu.ua
Bratislava University of Economics and Management, Furdekova 3240/16, 851 04 Bratislava, Slovakia. National Technical University “Kharkiv Polytechnic Institute”, 2, Kyrpychova str., 61002, Kharkiv, Ukraine
ORCID ID: https://orcid.org/0000-0002-9391-5220

Oksana IVASHCHENKO is an PhD student at the Department of Management and Marketing of Bratislava University of Economics and Management (Slovakia). Senoir lector is at the Department of Software Engineering and Management Intelligent Technologies of National Technical University “Kharkiv Polytechnic Institute” (Ukraine). Research interests: data mining, artificial intelligence and machine learning. E-mail: Oksana.Ivashchenko@vsemba.sk
Bratislava University of Economics and Management, Furdekova 3240/16, 851 04 Bratislava, Slovakia.
ORCID ID: https://orcid.org 0000-0003-3636-3914

ENTREPRENEURSHIP AND SUSTAINABILITY ISSUES
ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2023 Volume 11 September 1 (September)
http://doi.org/10.9770/jesi.2023.11.1(10)

doc. PhDr. PaedDr. Marcel LINCENYI, PhD is the vice-rector for doctoral and postgraduate studies of Bratislava University of Economics and management (BUEM), an associate professor and Head of the Department of Management and Marketing of BUEM (Slovakia). Research interests: marketing management, marketing of services, marketing communications E-mail: marcel.lincenyi@vsemba.sk
Bratislava University of Economics and Management, Furdekova 3240/16, 851 04 Bratislava, Slovakia.
ORCID ID: https://orcid.org/ 0000-0002-9076-026X

Ing. Marián KOVÁČ, PhD, is an associate professor of the Department of Management and Marketing of BUEM (Slovakia). Research interests: crisis management, risk management E-mail: marcel.lincenyi@vsemba.sk
Bratislava University of Economics and Management, Furdekova 3240/16, 851 04 Bratislava, Slovakia.
ORCID ID: https://orcid.org/0000-0003-4701-7830