Digitalization and Management Innovation III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA250005

ML-Based AI Conceptual Framework for Assessing SMEs Digitalization Through the Lens of Sentiment Analysis

Boryana PELOVA[a,1] and Gloria HRISTOVA[a]
[a] *Faculty of Economics and Business Administration, Sofia University "St. Kliment Ohridski", Bulgaria*
ORCiD ID: Boryana Pelova https://orcid.org/0000-0002-1393-1002
ORCiD ID: Gloria Hristova https://orcid.org/0000-0002-3950-4201

Abstract. Tools from the machine learning and data mining domain become even more popular in the fields of economics, entrepreneurship, and policy-making. At the same time, the research on small and medium-sized enterprises (SMEs) is getting amplifying importance for governments and policy-makers especially when it comes to support of SME's digitalization. A good understanding of the current level of digitalization of SMEs by industries is a prerequisite for design and integration of effective national policies. The goal of this paper is to design the architecture of an ML-based AI conceptual framework for assessing SMEs digitalization. We do this from the perspective of customers assuming that their preferences are absorbed in the publicly available (online) data that they generate in social media and community forums. This approach forms a significant contribution of this paper. Furthermore, we define an algorithm for data preparation, and we develop an algorithm based on sentiment analysis, which generates a set of industry-specific digitalization indices, which is another important contribution of this paper.

Keywords. Digitalization, small and medium-sized enterprises (SMEs), machine learning, natural language processing, artificial intelligence, sentiment analysis.

1. Introduction

Small and medium enterprises (SMEs) are nowadays an important driving force towards economic development, particularly in developing and emerging economies (see for example [1] and [2] as well as the references therein). The latest figures as published in [3] emphasizing SMEs importance for the EU economy are as follows. They constitute 99.8% of the number of enterprises in the EU indicating a growth of 2.7% for the period 2021-2022; the share of employed by SMEs is 64.4% with SMEs value added share amounting to 51.8%. Furthermore, [2] highlights the crucial role of SMEs for poverty alleviation and sustainable economic growth.

[1] Corresponding Author: Boryana Pelova, Faculty of Economics and Business Administration, Sofia University "St. Kliment Ohridski", Bulgaria; E-mail: bpelova@feb.uni-sofia.bg

The study and results are financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008. The dissemination costs and travel expenses are financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008 on activity 3.3.

This motivates the increased research interest in the determinants behind success and survival of SMEs [4]. As outlined in [2] information communication technology and changes of consumer preferences are among the major challenges within the context of economic globalization. In particular, the rapid development of information communication technology speeds up globalization and market liberalization, which amplifies the importance for SMEs to integrate technological innovations in the business.

However, while technological innovation fosters the growth of firms operating in high-tech resource rich environments of growing demand, it might be an additional challenge to small firms' survival in a disrupted declining industry [5]. Combined with fast changing customer preferences shaping an increasing demand for high-quality, diverse goods and services, the pressure towards digitalization of SMEs is determined mainly by the necessity to stay competitive at the globalized markets even for firms operating at the local market only.

Therefore, well-designed government policies in support of SMEs digitalization are imperative in order to mitigate this pressure and the starting point for such policies is the good understanding of the current level of digitalization of SMEs. This task requires sound assessment tools. Our paper contributes to this problem through embracing the power of data analytics. Application of tools from the machine learning and data mining domain has already proven their ability to provide new specific knowledge that might complement and support the intuition behind design and implementation of policies. Examples such as the SHARE project evidence this.

Taking into account the importance of SMEs and their survival for national economies, the goal of this paper is to design the architecture of an ML-based AI conceptual framework for assessing SMEs digitalization. We do this from the perspective of customers assuming that their preferences are absorbed in the publicly available (online) data that they generate in social media and community forums. This approach forms a significant contribution of this paper. Furthermore, we define an algorithm for data preparation, and we develop an algorithm based on sentiment analysis, which generates a set of industry-specific digitalization indices, which is another important contribution of this paper.

The rest of the paper is organized as follows. In Section 2 is provided a review of related studies. In Section 3 is introduced a case study that illustrates distinctly the research gap and the necessity to develop the proposed conceptual framework. Section 4 introduces the developed integrated framework for data engineering, knowledge extraction, and analytics of SME's data, followed by a discussion in Section 5.

2. Literature Review

The current study explores the possibilities of assessing SMEs digitalization level by employing a data-driven approach relying on the application of AI techniques. Since the beginning of the XXIst century, sentiment analysis represents a very active research area that integrates various natural language processing and mathematical approaches to understand and interpret the sentiment or subjective information expressed in a piece of text [6].

Sentiment analysis is an AI approach that has numerous applications in studies devoted to understanding public opinion and might be considered as an important complementary tool in large-scale surveys utilized in the social and management science [7]. When it comes to the assessment of digitalization level of a

service/company/industry, sentiment analysis might be helpful in at least several directions. First, it might be effectively used to provide insights into how the digitalization efforts are being perceived by the customer or general public. Second, it might be used in establishing benchmarks and key performance indicators (KPIs) for digitalization initiatives based on public sentiments. Furthermore, the analysis of internal communication can provide insights into the effectiveness of company's digital tools and processes as perceived by employees (internal subjective perspective). Following is a brief literature review of recent studies focused on digitalization assessment through the application of sentiment analysis and other NLP techniques.

Several up-to-date research papers examine the potential of AI, and natural language processing in particular, in the assessment of digitalization efforts in the government domain. In [8] is proposed an architecture of an ML system for mining public opinion on e-government services. The system combines topic modeling, sentiment analysis and visualization techniques applied on news and citizens' discussions in forums in an attempt to capture the progress in e-government development and the expressed opinions towards public e-services in Bulgaria. The main objective in [9] is to leverage state-of-the-art natural language processing technologies, particularly transformer-based language models, for public opinion analysis. The study focuses on the analysis of citizens' sentiments and emotions regarding the digitalization of services in the educational, administration, and health public sectors. Main source of data are discussions in community forums. The published results reveal interesting insights into public sentiments and emotions towards learning in an electronic environment, usage of electronic signature, e-voting, e-health systems etc.

In [10] sentiment analysis is applied on the aspect level in order to extract explicit aspects related to government software applications. The proposed approach captures reactions and emotions to e-government services development with the help of deep learning methods applied on citizen reviews. Muliawaty et al. [11] focus on the potential of big data and sentiment analysis technologies for enhancing bureaucratic public services and adoption of digital technologies. The authors propose a design of a sentiment analysis application that utilizes Twitter posts to understand public opinion on existing bureaucracy services. Yue et al. [12] present an insightful study on public perceptions towards smart city construction in China. Similar to the approach adopted in related studies, the authors apply a combination of sentiment analysis and topic modeling techniques on social media public comments. The utilization of the established topic modeling algorithm Latent Dirichlet Allocation (LDA) leads to the extraction of various topics related to smart cities among which are captured public concerns about technology applications, digital transformation of enterprises and digital industry economy. After topic extraction, sentiments in each topic are detected through deep learning techniques in order to gain deeper insights into public opinion.

Another study utilizing sentiment analysis to capture citizen opinion on government digitalization efforts is proposed in [13]. The authors develop an e-government gamification model aimed at motivating citizens to actively use digital public services for paying bills, renewing licenses etc. In [14] is presented novel research on the identification of emerging technologies in public services delivery with the aim of guiding policymakers in their implementation. The authors apply natural language processing techniques on a large database of academic papers focused on the topics of digital governance and digital democracy. Cluster analysis and network visualization techniques are applied in order to develop a better understanding of the role of emerging technologies in e-government and current trends in the field. Other studies aimed at

assessing digitalization efforts in the government domain through the application of AI techniques can be found in [15, 16, 17, 18].

Our review of related work reveals that in the last two years there has been an increasing volume of studies focused on the application of modern AI technologies in attempt to evaluate e-government development and various digitalization efforts in the public sector. We would adapt the recent developments from this domain to the domain of SMEs digitalization for the following reasons. AI provides extremely powerful tools that have the ability to reveal insights that could not be captured otherwise. Sentiment analysis and topic modeling are among the most frequently applied natural language processing techniques used in discovering public opinion towards digitalization initiatives and measures. The results from the reviewed studies have important implications for policy-makers and experts engaged in the development and implementation of policies related to digitalization within the government domain.

Our literature review reveals the almost complete lack of studies focused on the potential of AI applications in support of SMEs digitalization efforts. To the best of the authors' knowledge currently there are no research papers exploring the possible uses of natural language processing in the analysis of public opinion on SMEs current digitalization level and efforts. We bridge this gap in the literature by proposing a conceptual framework for assessing SMEs digitalization based on machine learning and natural language processing techniques.

3. Fostering Understanding of SMEs Effective Digitalization: A Case Study from Bulgaria

To illustrate how the proposed conceptual framework contributes to existing approaches for digitalization assessment, we provide as a context a case study based on survey data analysis. We should note that this survey is conducted under the same research project as the current paper. Details are provided in the first footnote of this text. Data for this study was collected through a national representative sociological survey among 1000 SMEs for the country of Bulgaria. The survey used a quota sampling technique based on type of SMEs (50% microenterprises and 50% small and medium-sized enterprises), Nomenclature of Territorial Units for Statistics NUTS 2 developed by the EU, and economic sector based on NACE Rev. 2 (Section A - Agriculture, Forestry and Fishing; high-technology and medium-high technology; medium-low technology and low technology; knowledge-intensive services; less knowledge-intensive services).

As explained in the introduction section, the rapid development of information communication technology speeds up globalization and market liberalization, which amplifies the importance for SMEs to integrate effectively technological innovations in the business. However, while technological innovation is known to foster the growth of high-tech firms, it might be an additional challenge to low-tech firms. Therefore, in this example we focus our attention on a subsample of firms that operate in the high-technology and medium-high technology sector, and in the medium-low technology and low technology sector. For simplicity of notation, we refer to the former as high-tech sector and to the latter as low-tech sector. Out of 1000 firms constituting our sample, 104 are classified as high-tech firms, and 127 firms are classified as low-tech firms. We consider the answers provided in a conducted survey on the self-assessed degree of digitalization and the integration of digital technologies in their businesses.

Figure 1 represents visually the self-assessed degree of digitalization at Likert scale ranging from 1 to 5, where 1 indicates that the firm considers itself as not digitalized at all and 5 stands for fully digitalized firm self-assessment. The share of companies at each level of digitalization for the high-tech and low-tech sector is represented as a percentage of all the SMEs in the respective sector. We could easily see that the highest saturation of companies is at levels 3 and 4. Even though high-tech firms are more digitalized as compared to low-tech SMEs, the differences are not as severe as we could intuitively assume.

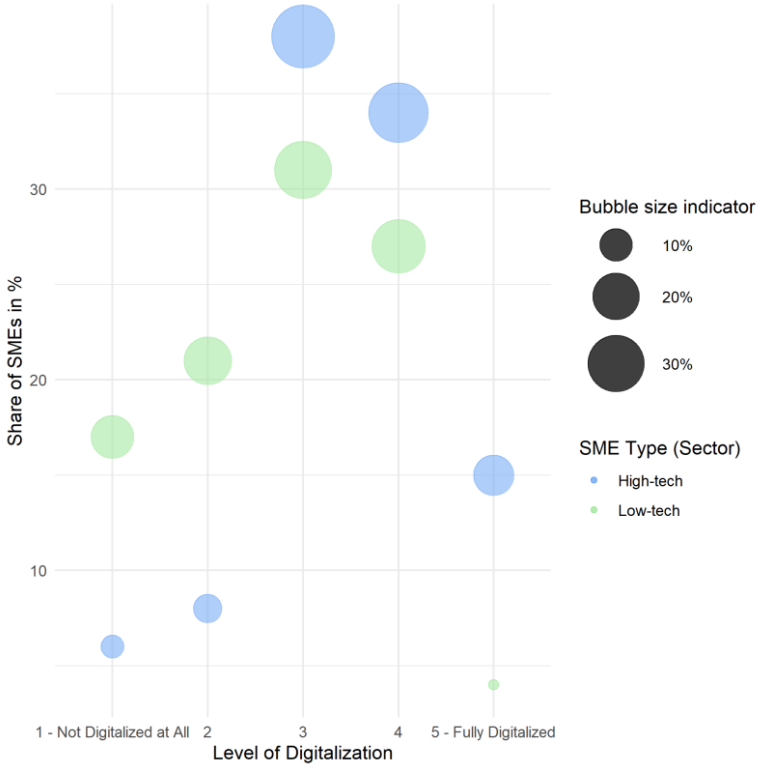

Figure 1. Self-assessed level of digitalization for high-tech and low-tech Bulgarian SMEs

Consequently, we look closely at the integration of two particular technologies that are intimately related to economic globalization. These are e-commerce, which includes online sales via website, as well as social media. Results are summarized in Figure 2. We observe that the share of SMEs integrating e-commerce in their business is approximately the same for high-tech and low-tech firms and even though the percentage of high-tech companies integrating social media is greater than that for low-tech companies, the latter show catch-up numbers. However, these numbers are not indicative on plausible differences in the effectiveness of integration of digital technologies between high-tech and low-tech companies. Customer opinion might provide valuable insights fostering comprehensive understanding of such kind of differences.

Therefore, we propose a conceptual framework embracing the power of publicly available text data that aims to complement the understanding of SMEs digitalization

traditionally assessed through survey data analysis. Customer opinion that might be found in community forums, social media platforms etc. contains important information on the degree of effective adoption of some common digital technologies by SMEs such as usage of e-trade instruments and social media presence. In order to illustrate this concept, we provide an example based on simulated data depicted at Figure 3.

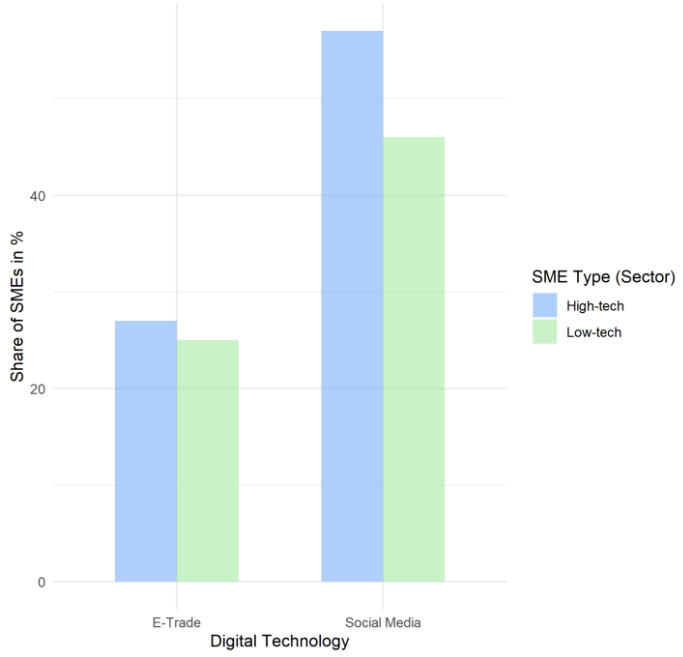

Figure 2. Digital technology adoption in high-tech and low-tech Bulgarian SMEs

We have generated three examples of customer feedback for a company that has a website for online sales and uses social media. While we observe integration of these technologies in its business, the client sentiment might be viewed as an additional source of information that could foster understanding on how effectively they actually work. In the next section we describe the proposed conceptual framework based on machine learning and natural language processing techniques that enable understanding SMEs digitalization from the client's perspective.

Customer feedback 1 (Text 1): *"Wow, their **website** is incredibly disorganized! It's frustrating trying to find the information I need. Plus, with the frequent occurrence of recent data breaches, I'm really worried about the **security** of my personal information when using this **website**."*

Customer feedback 2 (Text 2): *"I really like their digital **delivery** management process. It is really convenient to track your order and know when to expect it!"*

Customer feedback 3 (Text 3): *"OMG! Love their **Instagram** page! The visuals are great, and I like it that there is information about prices."*

Figure 3. Simulated customer feedback data

4. ML-based AI Conceptual Framework for Assessing Digitalization in SMEs

The main objective of the current paper is to develop a conceptual framework for assessing SMEs digitalization at industry level via application of ML and AI methods on SME's data. Figure 4 illustrates the main stages and key elements in the proposed framework.

In a nutshell, the framework relies on publicly available online data on SMEs operating in different industries. We motivate our choice of data as follows. We assume that opinions expressed in social media and community forums absorb customer preferences. This would allow us to assess SMEs digitalization through the lens of customers, which is a novel way to approach the issue.

Data is then processed via NLP and ML techniques so as to extract synthetic features subject to further analysis. These features are integrated to create an empirical study database that will be later used to derive insights on SMEs digitalization at industry level. Nevertheless, it is important to note that the proposed conceptual framework could be applied to evaluate the digitalization initiatives of SMEs not just at the industry level but also at the level of individual companies.

An important result, coming as an output of the conceptual framework for data analytics of SME's data, is the delivery of an empirical database providing insights into public opinion on SMEs digitalization efforts on industry level. The database encompasses insights from external sources, providing a comprehensive collection of objective digitalization measures for SMEs. Thus, such a database might be considered as an important complement to surveys focused on measuring digitalization development and efforts based on the internal "subjective" perspective of company's management board and staff.

Five important stages could be outlined in the proposed conceptual framework for assessing SMEs digitalization. The next subsections will describe in detail each of the main analytical stages of the framework displayed in Figure 4.

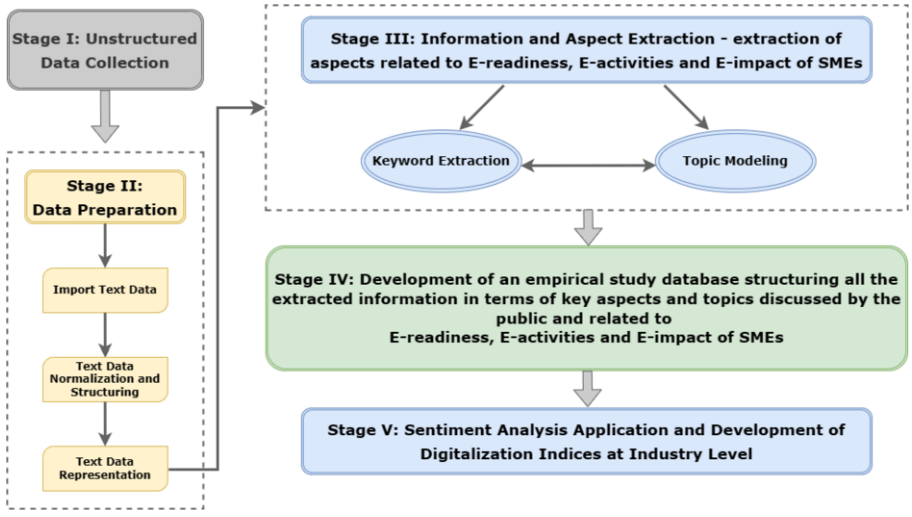

Figure 4. Analytical process behind the proposed conceptual framework for assessing digitalization in SMEs

4.1. Stage I – Unstructured Data Collection

The framework suggests that unstructured publicly available (online) data discussing small and medium sized enterprises is collected through web scraping. Such data could provide information on digitalization KPIs from an independent observer's perspective. Instead of scraping text data based on general keywords like "small and medium-sized enterprises" or abbreviations of this term, we broaden the scope of our search by extracting public comments for particular SMEs. To accomplish this, we first develop a sample of selected SMEs operating in different industries. This sample is derived with the help of official lists of SMEs names provided by government services. Then, using this information, text data is collected using carefully selected key words (for example, names of the companies, abbreviations of the names etc.). Thus, we are able to capture a wider spectrum of public opinion on SMEs performance and digitalization efforts. In addition, such data collection strategy enables empirical analysis not only at the "industry level", but also at "individual company" level (provided that sufficient amount of data is found for a specific company).

In the proposed conceptual framework, we put the most emphasis on the extraction and utilization of related to the selected SMEs online interactions in the form of reviews or public comments in social media and community forums. Similar to what have been already accomplished in the e-government domain [9], the application of sentiment analysis techniques on such data will enable the identification of public opinions regarding digital initiatives, processes and services and provide valuable insights into public perceptions of digital experiences with a given company. Such insights could highlight areas where improvements are required in the digitalization endeavors or where there might be deficiencies in digitalization development among SMEs in a particular industry. By applying different NLP tools on the publicly available online data, we are also able to extract key aspects related to SMEs digitalization on industry level. The aggregation of the results from information extraction and sentiment analysis, enables us to assess the digitalization efforts in each of these aspects from the customer (public) viewpoint.

4.2. Stage II – Data Preparation

This subsection describes the methodology for text data preparation (Stage II). The first step consists of the application of several important analytical techniques for text data normalization. After ensuring that text data crawled from the Internet are imported and read correctly in the software used for data analysis, we apply necessary NLP techniques for text data processing. The final aim is to derive text data fit for further analysis. Textual data collected from social networks and discussion forums exist in different formats and often include a significant amount of irrelevant or noisy content. Due to this reason, within Stage II, we implement essential procedures for data cleaning and employ necessary techniques to structure the data.

Let D denote the corpus of publicly available text data in the form of online reviews and public comments related to SMEs operating in a pre-defined set of K industries. Thus, each document d_i $(i = 1 \ldots N)$ is associated to a particular industry I_u, where $u = 1 \ldots K$. Given D consists of N text documents, then each document d_i will undergo the following text normalization techniques in the specified order:

1. Removal of HTML tags or analogous fragments presented in text data as a result of the data crawling process.

2. URL removal – URLs part of text data are considered irrelevant for subsequent analyses and therefore are removed.
3. Text data translation – this step might be skipped depending on the exact use case and choice of a particular country which SMEs will be under analysis. If the native language spoken in that particular country is considered a "low-resource" language (meaning that few language resources are available for processing and analysis of text data in this language), then text data translation might be applied. For example, performing sentiment analysis in unsupervised settings would require the availability of appropriate tools for carrying out the task on text data in the respective language. However, there might be the case that few such language resources exist for the particular language, while those that are available might not be appropriate for the domain of text data under analysis. In such scenario, data will be translated to English since there is an abundance of linguistic resources for this language.
4. Case normalization and digits/special characters removal.
5. Stop words removal – removal of the most frequently appearing words (stop words refer to the frequently used words in a particular language that contribute minimal information in data analysis). It is important to note that depending on the choice of text data representation technique this step might be skipped.

After the application of key text normalization methods, data will be put into more "structured" format through text tokenization applied on word level. Each document d_i consisting of z_i number of words ($z_i = \{1,2,\dots\}$) will be split to word tokens. At this point, it is important to mention that based on the specific characteristics of the text data sample at hand, z_i might be set to values larger than a given threshold value (and documents filtered out based on this criteria) since extremely low values of z_i for some text documents might imply lack of enough context and inability to extract useful insights in further stages of the analysis.

The final step in Stage II is text representation i.e., turning textual data into numerical format. The choice of an appropriate text representation technique hugely depends on the choice of machine learning algorithms that will be applied in subsequent data analysis. For example, traditional topic modeling algorithms like LDA [19] would require representing data using the vector space model which applies the bag-of-words assumption. However, more recent approaches like BERTopic [20] utilize document embeddings. The latter can effectively capture complex relationships within textual data, including semantic meanings. Both classical and novel approaches for text representation have their pros and cons [21]. However, an extensive discussion on this topic is outside the scope of the current study.

For the sake of maintaining simplicity within the context of the proposed framework, we briefly describe the text representation process when applying the classical vector space model. Utilizing the latter means that each document d_i will be represented as a fixed-length vector v_i ($v_i \in \mathbb{V}^S$) of word weights in the vector-space $\mathbb{V}^S$. Each vector v_i contains S elements (where S denotes the total number of unique words/tokens in corpus D) of which exactly y_i are non-zero (equals the number of unique words/tokens in d_i and $y_i \leq z_i$). Some established approaches for turning tokenized textual data to numerical format [22] that might be applied include integer vectorization and term frequency-inverse document frequency (TF-IDF). The latter is preferred in the application of classical topic models like LDA. Furthermore, TF-IDF is also employed

in most recent approaches like BERTopic where class-based TF-IDF weights are utilized in the development of more efficient topic representations.

The final output from Stage II is a document-term matrix $F_{(N\times S)}$ which contains the numerical representation of all text data related to SMEs.

4.3. Stage III – Information and Aspect Extraction

This subsection describes the methodology for application of NLP techniques for information extraction on publicly available (online) data related to SMEs. Stage III is crucial since we aim at the extraction of key factors/aspects/concepts related to SMEs digitalization and presented in text data. For the sake of illustration, we define three such aspects - E-Readiness, E-Activities and E-Impact of SMEs operating in a given industry I_u $(u = 1\ldots K)$. It is important to note, that these aspects might be subject to change during empirical analyses.

First, we employ a simple keywords extraction technique. A list of carefully selected keywords/phrases associated to each component of SMEs digitalization (E-Readiness, E-Activities and E-Impact) is developed based on domain knowledge. Each keyword/phrase is denoted by p_j, where $j = 1\ldots P$. Information about the presence of p_j in document d_i is stored in the form of a dummy variable, where the value of 0 indicates the absence of p_j, while the value of 1 indicates its presence in a given document. This idea is illustrated in Figure 5 - see Keyword 1, Keyword 2 etc. at each section related to E-readiness, E-activities and E-impact. As might be seen, each dummy variable associated to a particular keyword is a column in the dataframe, developed to structure all the extracted knowledge from text data.

			E-Readiness						E-Activities						E-Impact					
Industry	**Company**	**Text**	**Keyword 1**	**Keyword 2**	**…**	**Topic 1**	**Topic 2**	**…**	**Keyword 1**	**Keyword 2**	**…**	**Topic 1**	**Topic 2**	**…**	**Keyword 1**	**Keyword 2**	**…**	**Topic 1**	**Topic 2**	**…**
Industry I_1	Company 1	…	1	0	…	1	0	…	0	1	…	0	0	…	1	0	…	0	0	…
Industry I_1	Company 1	…	0	0	…	0	1	…	1	1	…	0	1	…	1	0	…	0	1	…
Industry I_1	Company 2	…	0	1	…	0	1	…	1	1	…	0	1	…	0	0	…	1	1	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Industry I_K	Company C	…	1	1	…	1	0	…	0	0	…	1	0	…	1	0	…	1	0	…
Industry I_K	Company C	…	1	0	…	0	0	…	1	0	…	0	0	…	1	1	…	1	0	…

Figure 5. Output from Stage IV – an empirical study database, that structures all the extracted information from the publicly available text data (for all SMEs part of the sample). For the sake of illustration, keywords and topics are related to three main aspects of SMEs digitalization level - E-Readiness, E-Activities and E-Impact.

The second technique for information extraction applied in an attempt to extract more aspects related to SMEs digitalization is topic modeling. This is a statistical technique used to discover latent topics or themes presented in a collection of text documents [23]. As mentioned earlier, classical topic modeling approaches like LDA or Non-negative Matrix Factorization (NMF) would require a text representation based on the vector space model. Such algorithms can be directly applied on the document-term matrix $F_{(N\times S)}$ to extract T main topics of interest discussed in reviews and comments related to SMEs. Each discovered topic t $(t \in \{1,\ldots,T\})$ is represented by a set of keywords considered as being most important in providing its general context.

After manual review of the extracted topics by domain experts, each of them is being assigned to one of the three main components of SMEs digitalization - E-Readiness, E-Activities and E-Impact. Since main focus is put on these three aspects of digitalization,

topics that cannot be related to any of these aspects are disregarded. Finally, the presence of a particular topic in a given text document is encoded using dummy variables (Figure 5 - see Topic 1, Topic 2 etc. at each section related to E-readiness, E-activities and E-impact), where the value of 0 indicates the absence of the topic, while the value of 1 indicates its presence in a given document (similar to the logic applied during keywords extraction). It is important to note that the topic assignment process hugely depends on the specifics of the finally chosen topic modeling algorithm and details could be provided in future research. As noted earlier, recent approaches for topic modeling as BERTopic would require a text representation based on the development of document embeddings rather than document-term matrices. Nevertheless, rest of the general logic behind the described analytical process for information extraction remains the same.

			E-Readiness						E-Activities						E-Impact					
Industry	**Company**	**Text**	**"website"**	**"security"**	**…**	**Topic 1**	**Topic 2**	**…**	**"chatbot"**	**"delivery"**	**…**	**Topic 1**	**Topic 2**	**…**	**"ads"**	**"instagram"**	**…**	**Topic 1**	**Topic 2**	**…**
Low-tech (I_1)	Company 1	Text 1	1	1	…	1	1	…	0	0	…	0	0	…	0	0	…	1	0	…
Low-tech (I_1)	Company 1	Text 2	0	0	…	0	0	…	0	1	…	0	1	…	0	0	…	0	0	…
Low-tech (I_1)	Company 1	Text 3	0	0	…	0	1	…	0	0	…	0	0	…	0	1	…	1	0	…

Figure 6. Example of the empirical study database (output from Stage IV) for "Company 1"

Coming back from our example set of customer feedback in Figure 3, we illustrate with Figure 6 how the first three rows of the empirical study database would look like. As might be seen from Figure 6, we have assumed that the company's name is "Company 1" and that it operates in the low-tech sector (denoted by I_1). Furthermore, in the first customer feedback the most predominantly discussed aspect of digitalization is E-readiness of "Company 1". We sum the dummy variables for each aspect in order to find out which aspect is most dominant. Sentiment is negative. In the second customer feedback, the most predominantly discussed aspect is E-activities of "Company 1". Sentiment is positive. In the third customer feedback, the most predominantly discussed aspect is E-impact of "Company 1". Sentiment is positive.

Topic	Keywords
Topic 1 (E-readiness)	GDPR, security, breach, confidence, privacy
Topic 2 (E- readiness)	easy, information, contact, fast, navigate
…	…
Topic 1 (E-activities)	assistant, digital, prompt, question, information
Topic 2 (E-activities)	digital, delivery, fast, track, information
…	…
Topic 1 (E-impact)	social, instagram, facebook, message, media
Topic 2 (E- impact)	ads, google, media, month, new
…	…

Figure 7. Illustrative example of topics

Figure 7 illustrates further the concept of topics and how they will be formed after the topic analysis is performed over the full sample of customer feedback for all companies. Each topic is represented by a set of keywords considered as being most important in providing its general context.

4.4. Stage IV – E-Readiness, E-Activities and E-Impact of SMEs at Industry Level

Stage IV is crucial for any subsequent analysis since it aims at developing an empirical study database containing in structured format all the extracted information related to each of the three components of SMEs digitalization. This idea is illustrated in Figure 5.

We develop the matrix $M_{(N\times(P+T))}$ structuring the extracted information about all aspects of the three main components of SMEs digitalization (E-Readiness, E-Activities and E-Impact) that are presented in the text documents in corpus D. As depicted on Figure 5, each row in the dataframe corresponds to a particular review/comment about a given SME, while each column is a dummy variable indicating the presence of a particular aspect of digitalization discussed in the given review/comment. Since each SME in the sample operates in a particular industry I_u ($u = 1 \ldots K$), the empirical study database allows to draw important insights on digitalization efforts at both company and industry level.

4.5. Stage V – Sentiment Analysis

In Stage V of the proposed conceptual framework are applied sentiment analysis techniques aimed at understanding the sentiments expressed by the public towards the three main components of SMEs digitalization efforts - E-Readiness, E-Activities and E-Impact. The output from this stage are three main digitalization indices - $SI_{\mathrm{E-Readiness}}$, $SI_{\mathrm{E-Activities}}$ and $SI_{\mathrm{E-Impact}}$ calculated at industry level. Each of these indices measures the overall public sentiment expressed towards the digitalization efforts of SMEs operating in a particular industry. This idea is illustrated in Figure 8.

Industry	E_1 (E-Readiness)	E_2 (E-Activities)	E_3 (E-Impact)
I_1	SI_{11}	SI_{12}	SI_{13}
I_2	SI_{21}	SI_{22}	SI_{23}
I_3	SI_{31}	SI_{32}	SI_{33}
…	…	…	…
I_K	SI_{K1}	SI_{K2}	SI_{K3}

Figure 8. Output from Stage V – a matrix of (digitalization) indices measuring the overall public sentiment expressed towards the digitalization efforts of SMEs operating in a particular industry. Each index SI_{ug} provides information on the overall public sentiment about SMEs digitalization efforts towards E_g ($g = \{1,2,3\}$) in industry I_u.

To derive the digitalization indices depicted in Figure 8, we employ the following methodology. First, we apply a suitable sentiment analysis model on each document d_i ($d_i \in D$) in which one (or more) of the three main components of SME digitalization is predominantly discussed (component dominance is calculated using the dummy variables in the matrix $M_{(N\times(P+T))}$ – see Figure 5). In case more than one component is equally represented in a given document d_i, sentiment analysis will be applied on aspect level [24], rather than document level, in order to capture the sentiments towards each of these main components of digitalization.

Sentiment analysis will enable us to evaluate the opinions expressed by the author of the text towards one (or more) of the main components of SMEs digitalization.

Sentiments are broadly defined in two categories - "positive" and "negative". For example, in the following public comment "I really like their digital delivery management process", the mainly discussed component of digitalization is "E-Activities" and the expressed sentiment is positive. Sentiment analysis is applied in an unsupervised manner by utilizing the pre-trained language model SiEBERT [25]. The latter is a sentiment analysis model designed for general-purpose application, trained on an extensive corpus of text data spanning diverse domains, including tweets, social media posts, and reviews of products and services. Its robustness and suitability for our use case stem from the extensive fine-tuning on a large volume of texts from various domains. It is important to note that natively SiEBERT operates on document level and additional fine-tuning might be performed in order to apply the model on the aspect level when necessary.

As explained earlier, our methodology for collecting text data results in the creation of a sample that permits empirical analysis on both "individual company" and "industry" level. Nevertheless, we are mostly interested at capturing insights on SMEs digitalization efforts at the industry level. For that reason, the final step in Stage V is to aggregate the extracted information about expressed opinions at company level and derive digitalization indices that measure the overall public sentiments expressed towards E-Readiness, E-Activities and E-Impact of SMEs operating in a particular industry.

For clarity, we denote each of the three main components of SMEs digitalization efforts (E-Readiness, E-Activities and E-Impact) with E_g, where $g = \{1,2,3\}$. Let SI_{ug} denote the overall public sentiment regarding E_g expressed towards companies operating in a particular industry I_u $(u = 1 \ldots K)$. SI_{ug} is an aggregated measure calculated by applying the following formula:

$$SI_{ug} = \frac{Po_{ug} - Ne_{ug}}{A_{ug}} \quad (1)$$

Eq. (1) calculates the overall public sentiment towards SMEs digitalization efforts E_g in industry I_u, where:

Po_{ug} - total number of positive sentiments expressed towards digitalization efforts E_g of companies operating in industry I_u. For example, Po_{11} is the total number of positive sentiments towards E-Readiness of SMEs operating in industry I_1.

Ne_{ug} - total number of negative sentiments expressed towards digitalization efforts E_g of companies operating in industry I_u. For example, Ne_{11} is the total number of negative sentiments towards E-Readiness of SMEs operating in industry I_1.

A_{ug} – total number of sentiments expressed towards digitalization efforts E_g of SMEs operating in industry I_u.

$SI_{ug} \in [-1,1]$, where the value of (-1) indicates strong negative public sentiments, while the value of 1 indicates strong positive public sentiments towards SMEs digitalization efforts E_g in industry I_u. Applying Eq. (1) leads to the development of the matrix $H_{(K \times G)}$ illustrated in Figure 8.

The constructed indices SI_{ug} serve as objective ("through the lens of the customer") measures of SMEs digitalization efforts at industry level. In addition, these indices combined with other available information about digitalization level and efforts in a particular industry might be used to reveal many new insights on the processes, degree,

effects, and problems of SMEs digitalization as well as on the scope of digital entrepreneurship in a particular country.

5. Discussion and Conclusions

In this paper we proposed an ML-based AI conceptual framework for assessing digitization of SME's data. We combine ML and NLP techniques in an attempt to create a robust methodology that could be used to assess SMEs digitalization level and efforts through the lens of customer preferences. One of the major contributions of our study is the development of a framework that integrates the analysis of unstructured data related to SMEs digitalization initiatives. Our literature review reveals the almost complete lack of studies in this research direction. Another important contribution of the paper is that our methodology relies on sentiment analysis thus capturing tendencies in customer preferences.

Such a framework has the potential to allow the extraction of valuable insights that are unobservable when conventional methodologies are being applied. Our approach for SMEs digitalization assessment in a given country is applicable not only at industry level, but also at "individual company" level (provided that sufficient amount of data is available for a given SME). We believe that the suggested framework might prove particularly valuable not only in assessing the digitalization level, but also in facilitating the development of measures aimed at supporting the digitalization of SMEs.

One of the major difficulties in the development of the framework stems from the noisy nature of unstructured textual data in the form of public comments freely expressed in social media. Acquirement, processing, and extraction of structured knowledge from such data pose considerable challenges because of the lack of standardized formats, diversity of platforms and the usage of informal language, abbreviations, slang, and misspellings.

One future refinement of the proposed framework includes the application of topic modeling analysis separately for each industry under analysis. Another direction for future research is studying the possibilities for improving and automating some of the manual tasks in the process of digitalization aspects extraction from text data. We believe that the current study will be valuable for researchers, policy-makers and other authorities involved in the development of digitalization policies and methodologies for assessment of the adaption of digital technologies in the business.

References

[1] Naradda Gamage SK, Ekanayake EM, Abeyrathne GA, Prasanna RP, Jayasundara JM, Rajapakshe PS. A review of global challenges and survival strategies of small and medium enterprises (SMEs). Economies. 2020; 8(4): 79. doi: https://doi.org/10.3390/economies8040079

[2] Gherghina ȘC, Botezatu MA, Hosszu A, Simionescu LN. Small and medium-sized enterprises (SMEs): The engine of economic growth through investments and innovation. Sustainability. 2020; 12(1): 347. doi: https://doi.org/10.3390/su12010347

[3] SME Performance Review 2022-2023 – EU27 countries sheet. European Commission; 2023 Jun 27, available at: https://single-market-economy.ec.europa.eu/smes/sme-strategy/sme-performance-review_en.

[4] Ismail Albalushi K, Naqshbandi MM. Factors affecting success and survival of small and medium enterprises in the middle east. Knowledge. 2022; 2(3): 525-38. doi: https://doi.org/10.3390/knowledge2030031

[5] Thomas G, Douglas E. Small firm survival and growth strategies in a disrupted declining industry. Journal of Small Business Strategy. 2021; 31(5): 22-37. doi: 10.53703/001c.29814
[6] Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012; 5(1): 1-167. doi: https://doi.org/10.1007/978-3-031-02145-9
[7] Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez PM. A review on sentiment analysis from social media platforms. Expert Systems with Applications. 2023; 223: 119862. doi: https://doi.org/10.1016/j.eswa.2023.119862
[8] Hristova G, Bogdanova B, Netov N. Design of ML-based AI system for mining public opinion on e-government services in Bulgaria. In AIP Conference Proceedings; 2022, September: AIP Publishing LLC. p. 020005. doi: https://doi.org/10.1063/5.0100870
[9] Hristova G, Netov N. Utilization of Transformer-Based Language Models in Understanding Citizens' Interests, Sentiments and Emotions Towards Public Services Digitalization. In Digitalization and Management Innovation II; 2023: IOS Press. p. 1-13. doi: https://doi.org/10.3233/FAIA230711
[10] Rongxuan S, Bin Z, Jianing M. End-to-End Aspect-Level Sentiment Analysis for E-Government Applications Based on BRNN. Data Analysis and Knowledge Discovery. 2022; 6(2/3): 364-375.doi: 10.11925/infotech.2096-3467.2021.0945
[11] Muliawaty L, Alamsyah K, Salamah U, Maylawati DSA. The concept of big data in bureaucratic service using sentiment analysis. In Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines; 2022: IGI Global. p. 1189-1202.
[12] Yue A, Mao C, Chen L, Liu Z, Zhang C, Li Z. Detecting changes in perceptions towards smart city on Chinese social media: A text mining and sentiment analysis. Buildings. 2022; 12(8): 1182. doi: https://doi.org/10.3390/buildings12081182
[13] Mostafa L, Beshir S. Using Gamification in Egyptian E-Government. In International Conference on Advanced Intelligent Systems and Informatics; 2022 Nov 18: Cham: Springer International Publishing. p. 344-353. doi: https://doi.org/10.1007/978-3-031-20601-6_31
[14] Rodriguez Bolivar MP, Alcaide Munoz L. Identification of research trends in emerging technologies implementation on public services using text mining analysis. Information Technology & People. 2022. doi: https://doi.org/10.1108/ITP-03-2021-0188
[15] Righettini MS, Ibba M. Cultural Heritage Digitalisation Policy as a Co-creation of Public Value. Evaluation of the Participatory Digital Public Service of Uffizi Galleries in Italy During the COVID-19. In INTERNATIONAL SYMPOSIUM: New Metropolitan Perspectives; 2022 May 24: Cham: Springer International Publishing. p. 257-267. doi: https://doi.org/10.1007/978-3-031-06825-6_25
[16] Alqaryouti O, Siyam N, Abdel Monem A, Shaalan K. Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics. 2024; 20(1/2): 142-61. doi: https://doi.org/10.1016/j.aci.2019.11.003
[17] Al-Qudah DA, Ala'M AZ, Castillo-Valdivieso PA, Faris H. Sentiment analysis for e-payment service providers using evolutionary extreme gradient boosting. IEEE Access. 2020 Oct 19; 8: 189930-44. doi: 10.1109/ACCESS.2020.3032216
[18] Arku RN, Buttazzoni A, Agyapon-Ntra K, Bandauko E. Highlighting smart city mirages in public perceptions: A Twitter sentiment analysis of four African smart city projects. Cities. 2022; 130: 103857. doi: https://doi.org/10.1016/j.cities.2022.103857
[19] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003; p. 993-1022.
[20] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL]. 2022. doi: https://doi.org/10.48550/arXiv.2203.05794
[21] Hristova G, Netov N. Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic: A Topic Modeling Approach Based on BERTopic. In 2022 IEEE International Conference on Big Data (Big Data); 2022 Dec 17: IEEE. p. 2259-2264. doi: 10.1109/BigData55660.2022.10020466
[22] Miner G. Practical text mining and statistical analysis for non-structured text data applications. Academic Press; 2012. doi: 10.1016/C2010-0-66188-8
[23] Kherwa P, Bansal P. Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems. 2019; 7(24). doi: http://dx.doi.org/10.4108/eai.13-7-2018.159623
[24] Cambria E, Das D, Bandyopadhyay S, Feraco A. (Eds.). A practical guide to sentiment analysis. Cham: Springer International Publishing; 2017. doi: https://doi.org/10.1007/978-3-319-55394-8
[25] Hartmann J, Heitmann M, Siebert C, Schamp C. More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing. 2022. doi: https://doi.org/10.1016/j.ijresmar.2022.05.005