Актуальні проблеми розвитку економіки регіону. Випуск 22. Т. 2 **ISSN (Print) 2313-8246**
The actual problems of regional economy development. Issue 22. Vol. 2 **ISSN (Online) 2518-7589**

РОЗВИТОК ТУРИСТИЧНОГО, ГОТЕЛЬНОГО ТА РЕСТОРАННОГО ГОСПОДАРСТВА
THE DEVELOPMENT OF TOURISM, HOTEL AND RESTAURANT BUSINESS

UDK 338.48:004.89
doi: https://doi.org/10.15330/apred.2.22.72-88

Дехтяр Н.А.

ВИКОРИСТАННЯ МОДЕЛЕЙ ШІ У КОНТЕНТНОМУ АНАЛІЗІ ПОТЕНЦІЙНОГО ТУРИСТИЧНОГО ПОПИТУ

Харківський національний економічний
університет імені Семена Кузнеця,
кафедра міжнародної економіки і менеджменту
пр. Науки 9-А, м. Харків,
61165, Україна,
тел.: +380663056847
e-mail: nadiya.dekhtyar@hneu.net
ORCID: https://orcid.org/0000-0001-7932-8620

Анотація. У статті досліджується використання сучасних моделей штучного інтелекту для контентного аналізу потенційного туристичного попиту в умовах глобальної цифрової трансформації та необхідності стратегічного відновлення туристичної галузі України; аналізується процес визначення завдань моніторингу відкритих вебресурсів і формування вимог до оптимальних великих мовних моделей на основі методів монографічного аналізу, синтезу та моделювання процесів узагальнення неупорядкованих цифрових даних. Одним із ключових результатів є експериментальне порівняння екстрактивних (BERT, SBERT) та абстрактивних (GPT, BART, T5, Pegasus) моделей, яке довело, що модель BART забезпечує найбільш збалансоване поєднання контекстуального розуміння та генеративної плавності, тоді як інші моделі можуть викривляти реальні настрої аудиторії через надмірну акцентуацію на окремих групах або генерацію неіснуючих ідей. Надано детальний опис технологічної послідовності дослідження відгуків – від збору даних за допомогою YouTube Data API та методів вебскрапінгу до токенізації, лематизації та застосування предиктивної аналітики для розпізнавання людських афектів у межах концепції "affective computing". Наукова новизна полягає в обґрунтуванні спеціалізованої дослідницької інфраструктури для стратегічного управління в туризмі, що інтегрує коригуючі вузли у вигляді експертів-аналітиків, які здійснюють верифікацію результатів роботи машинних алгоритмів для підвищення достовірності моніторингу ринку та ідентифікації гібридних загроз в інформаційному просторі. Практична значущість дослідження визначається можливістю трансформації якісних споживчих уподобань у кількісні аргументи для стратегічного планування рекреаційної інфраструктури та ребрендингу національного туристичного продукту на основі інформаційної підтримки з боку державних і приватних інституцій. Запропонований підхід, що базується на використанні доступних бібліотек з відкритим кодом, забезпечує малим та середнім підприємствам можливість впровадження інтелектуальних систем аналізу ринку, що виходять за межі традиційного маркетингу й дозволяють проактивно управляти туристичними потоками в умовах трансформації глобальної економіки, а на рівні держави — запобігати впливу гібридних загроз.

Ключові слова: штучний інтелект, контентний аналіз, туристичний попит, великі мовні моделі (LLM), аналіз настроїв, моніторинг потреб споживачів, стратегічне управління в туризмі.

Dekhtyar N.A.

USING AI MODELS IN THE CONTENT ANALYSIS OF POTENTIAL TOURISM DEMAND

Simon Kuznets Kharkiv National University of Economics,
Department of International Economics and Management
Nauky Ave., 9A, Kharkiv,
61165, Ukraine,
tel.: +380663056847
e-mail: nadiya.dekhtyar@hneu.net
ORCID: https://orcid.org/0000-0001-7932-8620

Abstract. This article examines the use of modern artificial intelligence (AI) models for content analysis of potential tourism demand in the context of global digital transformation, and the need to revitalise Ukraine's tourism sector strategically. It analyses the process of defining tasks for monitoring open web resources, and of formulating requirements for optimal large language models. This is based on monographic analysis and the synthesis and modelling of processes for generalising unstructured digital data. A key finding is the experimental comparison of extractive (BERT and SBERT) and abstractive (GPT, BART, T5 and Pegasus) models. This demonstrated that the BART model provides the most balanced combination of contextual understanding and generative fluency. Other models may, however, distort the audience's actual sentiments due to an excessive focus on specific groups or the generation of non-existent ideas. The technological workflow for analysing reviews is described in detail, from data collection using the YouTube Data API and web scraping methods, to tokenisation and lemmatisation, and the application of predictive analytics to recognise human emotions within the framework of "affective computing". The scientific novelty lies in justifying a specialised research infrastructure for strategic management in tourism that integrates corrective nodes in the form of expert analysts who verify the results of machine algorithms, thereby enhancing the reliability of market monitoring and identification of hybrid threats in the information space. The research's practical significance lies in its ability to transform qualitative consumer preferences into quantitative arguments for the strategic planning of recreational infrastructure and the rebranding of the national tourism product, based on information support from public and private institutions. The proposed approach, which is based on the use of open-source libraries that are readily available, enables small and medium-sized enterprises to implement intelligent market analysis systems that go beyond the scope of traditional marketing. These systems allow for the proactive management of tourist flows amidst the transformation of the global economy. At the state level, they enable the prevention of the impact of hybrid threats.

Key words: Artificial intelligence, content analysis, tourism demand, large language models (LLMs), sentiment analysis, monitoring consumer demands, strategic management in tourism.

Introduction. When developing a new digital product, it is necessary to consider not only the technological requirements but also data security, privacy, and the psychological, emotional, and social impact the product will have on individual users and society as a whole. Ukrainian society is currently facing an increased threat from hybrid attacks and disinformation in the digital environment; therefore, developers of digital products must implement effective methods to protect personal data and prevent their products from being used for illegal purposes, which unfortunately occurs on a massive scale on social media.

The latest internet infrastructure envisages a transition to almost complete decentralisation of data at all levels of storage and management. New models, ranging from fully centralised to fully decentralised, challenge traditional organisational and infrastructure management structures and the transfer of expertise. However, sectoral management lacks coordinated state strategies for monitoring consumer sentiment, and does not fully understand the sequence of stages and hierarchy of process integration. This research is relevant because of the need to revive Ukraine's tourism sector, which under martial law focuses primarily on digital services. Before embarking on sentiment analysis, which is already used by multinational companies based on global big data, Ukrainian experts must define the information infrastructure and ensure coordination between different expert groups. Hybrid interfaces may be used, incorporating elements of sentiment analysis in the initial stage of semantic analysis, or models specifically designed to analyse visitors' emotions may be selected (as will be evident from the examples given below, different categories of AI models place varying degrees of emphasis on the presence of emotions in the text of primary sources).

The issue of using artificial intelligence in the tourism sector is a very popular topic in academic publications; most researchers view it as a multifaceted process that most often encompasses marketing, customer service and operational management. Antczak B. et al. [1] argue that the integration of AI into marketing strategies is revolutionising the sector through the use of chatbots, virtual assistants and recommendation systems, which provide a personalised experience, thereby increasing customer satisfaction and loyalty. A similar approach is proposed by Dyduch W. and Brzozowska A. [2], who state that to improve the efficiency and competitiveness of businesses, the "SmartTour" model should be utilised, which involves process automation, personalised offers and in-depth data analysis using machine learning. Katsanakis I. et al. [3] note that AI technologies, in particular predictive analytics, enable tourism companies to forecast demand, dynamically adjust prices and identify new growth opportunities.

Particular attention is paid in the publications to direct interaction with tourists and content creation. Zhang M. et al. [4] believe that the emergence of virtual influencers and streamers created using generative AI (GenAI) opens up new marketing opportunities for engaging younger audiences and promoting cultural destinations. Seyfi S. et al. [5] note that when implementing GenAI for travel planning, it is important for companies to take tourists' personality traits (the "Big Five" model) into account, offering adaptive interfaces and trial programmes to overcome barriers to adoption. Štilić A. et al. [6] investigated AI-generated travel itineraries (AITI) and are confident that they are an effective tool to assist travel agents, as they enable the rapid creation of basic plans, which specialists can then refine to ensure uniqueness and cultural appropriateness.

Researchers also highlight ethical considerations and user trust. Koo I. et al. [7] emphasise that ethical practices in the use of AI, such as algorithmic transparency and the protection of personal data, are critical to building trust, which directly influences tourists' willingness to adopt new technologies. López-Naranjo A. L. et al. [8] note that AI can optimise risk management by helping to identify potential threats, ranging from weather conditions to socio-political changes, enabling companies to take proactive measures to ensure tourists' safety.

At the same time, the researchers call for a balanced approach, where technology complements rather than completely replaces human involvement. Hall C. M. and Cooper C. [9] note that, despite rapid automation and the introduction of service robots, tourism remains a sector based on human interaction and empathy; it is therefore important to preserve the human touch in the tourist experience. Thus, the authors of these publications emphasise, in one way or another, the idea that the future of tourism depends on the harmonious combination of AI's analytical capabilities with human experience and ethical standards. On the other hand, most researchers focus on the behaviour of the end consumer of tourism services and, more generally, on society's perception of innovations influenced by AI, but do not provide recommendations on how recreational infrastructure might be transformed – including the monitoring of the tourism market for government bodies regulating tourism, nor do they emphasise the need to incorporate the relevant information space – where travellers gain knowledge about tourism resources and exchange experiences – into the market research methodologies of tourism companies. This is extremely important today, as active internet users are increasingly planning their trips using AI technologies, which in turn draw largely on an unstructured pool of information. Consequently, issues relating to traveller safety are of particular relevance, not only in terms of personal data protection but also regarding the security and reliability of the information that forms the content basis of AI models.

Task statement. The aim of the study is to identify the main tasks involved in analysing open-access resources containing data on enquiries from consumers of tourism services, and the requirements for language models based on artificial intelligence technologies that are optimal for specific expert monitoring tasks. The study utilised methods of monographic analysis – to review publications on the subject of the article; synthesis – to propose a procedure for content assessment; and modelling – to select the optimal large language model based on the generalisation of comments. The scientific novelty lies in justifying the transformation of the state regulatory system for tourism activities by shifting from fragmented consumer behaviour analysis to comprehensive strategic monitoring of sentiment within unstructured information pools using large language models. This enables institutions responsible for tourism development – which indirectly influence major tourism companies – to adapt recreational infrastructure to actual demand and forecast the stages of the sector's recovery under martial law in Ukraine.

Results. Content creators are keen to increase user loyalty and retain existing subscribers and visitors, as well as attract new ones. To this end, they utilise technologies that track consumer behaviour on their platforms and process the information (including sensitive personal data) that consumers provide. Such technologies have long been used by various services and applications, with even the most popular search engines offering results based on stored browsing history and preferences. Users can turn off this feature — for instance, by blocking cookies — but internet users generally do not consider where their device and search history information is sent, particularly as not everyone possesses sufficient digital literacy.

However, it is precisely the ability to collect and process de-identified data that forms the basis of big data processing technology. If strict restrictions are imposed, the resource owner will receive inaccurate, incomplete and fragmentary information, which will reduce the quality of the services offered. Hotel owners, for example, may find that data on guests' preferences, taken from personal forms completed upon check-in and post-stay reviews, is sufficient. However, transport and leisure service providers require data on traveller flows, their average profiles, preferences regarding various leisure activities and the criteria used to choose destinations for short and long-term stays. Even the most basic profile requires more than standard questionnaire data; results of at least a semantic analysis of search queries related not only to travel are required. In many countries, local legislation does not require guests to verify their identity with a passport or equivalent document upon check-in. Consequently, informal surveys and extracts from booking system records are often the only source of information about guests, provided these are available, and the hotel is connected to such a platform. Social media is invaluable for obtaining personalised analytics, but many users have multiple profiles, some of which may be private. Not all travellers are willing to share their account details with every service provider they come into contact with, even if it is only for one or two days. This problem arises not only for owners of hotels and other tourist facilities, where at least some information can be obtained during a face-to-face meeting, but also for providers of purely digital services, such as virtual tours, websites, and apps offering infotainment content, particularly if the consumer accesses the page via a desktop computer. In such cases, the most reliable information that can be obtained without breaching privacy is the technical specifications of the devices used and the geographical location of the node from which the request originated (i.e. where the consumer was located), as determined by the IP address. Again, this assumes that the virtual tourist has not enabled a VPN.

Research into emotions in the digital environment also utilises artificial intelligence, for example methods for extracting emotions from text messages (sentiment analysis), which are integrated into large language models (LLMs). This field is known as 'affective computing' – the study and development of systems and devices capable of recognising, interpreting, processing and simulating human emotions (from the Latin 'affectus' meaning passion or emotional agitation). These are understood to be strong, relatively short-lived emotional processes during which the degree of self-control is reduced. In other words, when a consumer is 'in a state of affect' — under the influence of one of the basic emotions — they unconsciously assess the situation subjectively and, depending on their level of self-control, may immediately take the action the marketer is counting on. The method of incorporating emotion into the sales process has indeed been used successfully in advertising for many years and is considered indispensable regardless of technological or societal development.

Most emotionally balanced people are happy to leave positive reviews of tourist services if they have the time; however, if they did not like something but it is not a major issue, they usually leave an impersonal, neutral review or do not respond at all. In such cases, both the visitor and the service provider retain a positive impression of one another. However, the service provider may not realise that the customer was dissatisfied, and will not know which aspect of the product needs improvement. It is quite likely that rectifying the element that annoyed the tourist would require minimal effort. Tourists often leave comments about past trips and future preferences later than expected, or in places that service providers do not anticipate. Furthermore, relatively few tourists return to the same destination where they have previously holidayed, particularly young people – they tend to try visiting different countries or even continents with every holiday to diversify their experiences. Therefore, all tourism service providers and producers must be proactive in creating destinations and generating digital content while understanding the sentiments and expectations of potential visitors regarding how they spend their leisure time. Consequently, analysing social media and forums where travellers leave comments and share experiences should begin at the product development stage. While the decentralisation of the tourism market towards spontaneous individual travel and sporadic booking of package holiday components makes it difficult to identify target segments for analysis, this chaos and abundance of personal post content provides a resource for meaningful semantic analysis. Even if bots generate most reviews and are negative (a common modern tool of competitive warfare), text analysis can identify fundamental trends – the most relevant statements and topics for visitors to a given resource and the issues that currently resonate with people on a specific topic. In free-form comments on platforms not associated with official service provider or tourist attraction channels, travellers tend to express their opinions more openly. For example, if tourists generally enjoyed their stay at a hotel but were extremely dissatisfied with the location's other leisure facilities, they are unlikely to mention these issues in their hotel reviews. This does happen, but rarely, as people generally try to respond more quickly to 'official' enquiries, even if they enjoyed the service. The downside of automated responses to questionnaires with closed questions and perhaps a final open-ended question is that travellers will answer concisely, whilst hotel owners are not particularly interested in how establishments that are not direct competitors operate. If a city has several museums that are not particularly well known, meaning tourists do not choose the city solely to visit them, a hotel owner will be more interested in entertainment infrastructure and transport links within walking distance, as a transfer from the airport to the hotel can be a key competitive advantage. However, tourists who join discussions about particular events in the country's cultural life or seemingly unrelated topics on local museum pages may provide ideas for rebranding tourist destinations in that country or highlight issues that are important to tourists from other countries but not to residents, such as duplication of information on street signs in different languages or suggestions to expand payment methods for digital services beyond just tourist-related ones.

The process of managing user data consists of the following stages. For new visitors to the resource, it is linear; however, for an existing database, some components may be rearranged depending on the purpose of the query. (Fig. 1).

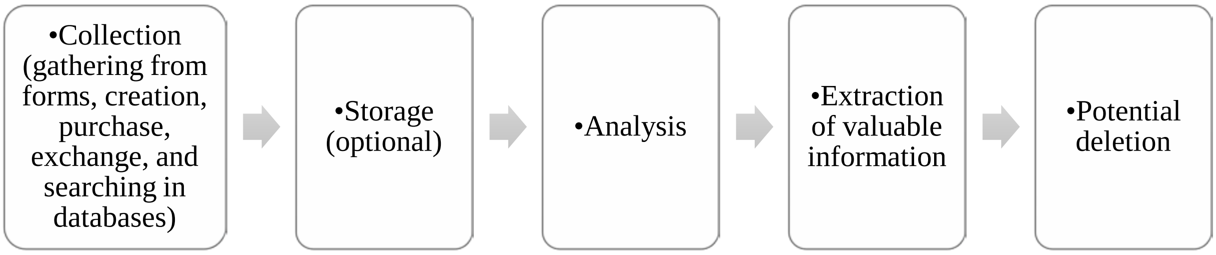

Fig. 1. Data processing operations [10]

As can be seen from the diagram, this process consists of several stages at which important information may be lost. Secondary data, on the other hand, may take up valuable storage space. When it comes to thousands of subscribers across hundreds of web pages that need constant monitoring for feedback and ideas, the task becomes virtually impossible for experts using traditional analysis methods. While a hotel manager may be able to remember the requirements of regular customers, even a team of administrators would be unable to keep track of the dozens of sources that influence a tourist's choice of accommodation in one way or another.

Where resources permit, it is advisable to use AI agents for the preliminary semantic analysis of text messages relating to various aspects of shaping Ukraine's tourism image on the internet (it is also worth incorporating the analysis of information in audio and video formats). Earlier generations of agents allowed for the monitoring of comments on social media and simple text embedded in HTML code, and this technology is now widespread. Such agents operate using "web spider" technology (also known as web scrapers or web crawlers) – the first programmes created to save human analysts time when searching for and collecting information on a given topic hosted on various resources (much like a mini-search engine). Search engines (such as Google) use search robots, also known as crawlers or simple bots, to index websites. As well as their search function, spiders can be programmed to extract information from specific websites, for example, when search bots do not identify unpopular pages in the search query rankings or when it is necessary to filter and download specific information from a website. The standard procedure for configuring an agent is as follows: let us consider the keyword phrase 'virtual tourism' and the content analysis of comments on YouTube under videos dedicated to this topic. (Fig. 2).

The process of researching content users' opinions

Stage 1 →	1.1. Identifying the list of sources to be studied and collecting data. 1.2. Identifying the original languages. Data cleaning and normalisation 1.3. Tokenisation and lemmatisation
Stage 2 →	2.1. Analysis of user sentiments and emotions. Summary of sentiments across different channels 2.2. Identification of dominant themes
Stage 3 →	Synthesis (generalisation of information): 3.1. Extractive generalisation – identifying comments that reflect the content of each cluster 3.2. Abstract generalisation based on MLM – writing a concise summary that generalises the various points of view 3.3. Comparative analysis of differences and similarities between channels (if several sources are being analysed, or if significant discrepancies have been identified between clusters) 3.4. Temporal analysis – how the topics and tone of discussions change over time
Stage 4 →	Preparing results for practical use 4.1. Writing a brief summary 4.2. Preparing visualisations (optional)

Fig. 2. Sequence of stages in the process of analysing reviews/comments
(compiled by the author)

Researching consumer expectations of virtual tourism services using comments requires a systematic approach that begins with collecting data from various sources. Comments posted beneath videos on several YouTube channels initially constitute a disparate collection of impressions and discussions (which often begin with the video's topic but may end up discussing completely unrelated subjects), which can be consolidated using the YouTube Data API or other data collection mechanisms. Once collected, these comments must be cleaned and normalised. Spam, duplicates and irrelevant characters are removed; it is also necessary to identify the original languages of the account holders to ensure high-quality processing of multilingual comments and preserve their semantic meaning. (The task of adapting off-the-shelf algorithms, designed for English, to Ukrainian is in fact quite non-trivial, as grammatically unrelated language groups contain different components, or conversely, they lack elements that are key to the algorithm, and the source code often has to be significantly modified, sometimes even altering the logic of the tokenisation process. A. Glybovets and V. Tochytskyi [11] highlighted this problem in one of their early publications on the subject, suggesting the creation of bespoke libraries for analysing texts in Ukrainian, and gave an example of adapting Porter's algorithm for tokenisation and stemming. Tokenisation and lemmatisation further refine the text, preparing it for more in-depth analysis. The developers of the national LLM model, 'Syaivo', also encountered significant differences in the structure of the Ukrainian language compared to English. This joint project was launched at the end of 2025 by specialists from the Ministry of Digital Transformation and the company "Kyivstar". Google Gemma 3 was chosen as the training foundation (with the transition to Gemma 4 now). The main issue is that large language models were developed for English, so an intermediate step is required to process Ukrainian text. The input text is first translated into English, then processed to generate responses in English. Finally, the finished response is translated back into Ukrainian for the final user. But the difficulties do not end there. Even when a model is specifically tailored to Ukrainian, a significant volume of data is required to fine-tune it. The larger the volume of content used to train the model, the better it performs. However, there are far fewer high-quality Ukrainian language sources on the internet than for other widely spoken languages. This is why experts are currently working on digitising sources, which will then be included in the training model pool. Plans are also in place to migrate the model from Google's cloud infrastructure to the "AI Factory", the state-run AI infrastructure located in Ukraine [12].

The next stage involves analysing users' sentiments and emotions. Lexicon-based methods can quickly assign polarity values, but transformational models such as BERT or RoBERTa provide a more nuanced interpretation, distinguishing between emotions such as enthusiasm, scepticism and neutrality. When applied to the context of the key phrase, these models reveal whether followers are expressing enthusiasm for virtual travel, concern about the condition of the real-world site on which the tour is based, or whether they are at all influenced by what they have seen – or, conversely, are expressing indifference towards any form of tourism. Summarising sentiments across different channels provides an initial understanding of how the audience perceives the topic.

In parallel with sentiment analysis, topic extraction methods are applied to identify the dominant themes (headings) in the comments. Keyword extraction techniques, such as TF-IDF or RAKE, identify recurring terms, whilst topic modelling approaches, such as Latent Dirichlet Allocation, group comments into coherent clusters. Clustering based on embeddings using Sentence-BERT and k-means further improves the grouping of semantically similar comments. In the context of virtual tourism, these clusters can reveal discussions centred on digital technologies, virtual reality devices, exploring post-conflict territories, or integrating science into travel to learn about cultural heritage.

Generalisation takes place during the synthesis stage. Extractive generalisation identifies representative comments that reflect the essence of each cluster, whilst abstractive generalisation uses large language models to generate coherent, concise summaries that synthesise different viewpoints into a single narrative. When applied to virtual tourism, generalisations can highlight subscribers' emphasis on the educational value of such trips, recognition of the role such trips play in promoting environmental awareness, given that many sites of tourist interest have suffered damage and destruction due to excessive recreational pressure, content users' concerns regarding ethical considerations, or the high cost of the equipment required to immerse oneself in a virtual tour.

A comparative analysis allows us to identify differences and similarities between the channels. Some audiences may focus on the potential of virtual tourism to foster international cooperation in raising awareness of destroyed natural sites and historical and cultural heritage landmarks, whilst others may emphasise its economic implications or its role in the recovery of Ukraine's tourism sector following the end of the armed conflict. A temporal analysis adds another layer, showing how discussions evolve over time and whether enthusiasm or scepticism is growing in response to new developments.

Finally, the results can be visualised and presented in a format suitable for practical use. Dashboards can display trends in sentiment, thematic clusters, and comment volumes, whilst word clouds and diagrams highlight the most frequently used terms. Comparative tables can show how perceptions of the topic under study differ across channels, offering a structured overview of the discourse. Through this process, AI agents transform raw streams of comments into actionable insights, enabling a deeper understanding of how virtual tourism is perceived and discussed in the digital environment, and whether content users are ready to accept it as a commercial product.

Using the topic of virtual tourism as an example, let's examine the differences between various models of abstract generalisation. Imagine a set of comments on YouTube under a video about virtual tourism. The input comments, in simplified form, are as follows: let's assume that we have received a file containing a list of comments, ready for analysis. Each line in the uploaded text file must look like this: 'user000': 'This is a comment'.

```
'user001': "With virtual tours, I can visit places that I can't afford to travel to in person."
'user002': "What I see doesn't feel as real as actual travel."
'user003': "I like that these tours are accessible to people with physical disabilities."
'user004': "The technology still needs improving to provide a fully immersive virtual reality experience."
```

Researchers can choose one or more AI models to analyse text, but there are significant differences between them. Currently, three main groups of models are distinguished (Table 1):

Table 1

CATEGORIES OF AI MODELS

Category	Model, year of development	Features
Extractive	BERT (2018), RoBERTa, XLNet, SBERT (2019), Longformer, ELECTRA (2020)	Encoder-only architectures; perform well for classification, key sentence extraction and semantic embeddings; lack a generative decoder, and are therefore used for extractive generalisation and analysis.
Abstract	GPT (2018), BART, T5 (2019), Pegasus, ProphetNet, mBART (2020), FLAN-T5 (2022)	Generative or encoder-decoder models; generate new text and are capable of paraphrasing and generalisation; used for dialogue systems, translation and automatic summarisation.
Hybrid / extended	BigBird (2020), Switch Transformer (2021), UL2, PaLM (2022), LLaMA (2023)	Scalable architectures combining different approaches; optimised for long contexts, multilingualism and general-purpose NLP tasks; used in research and commercial products.

Compiled by the author based on official sources from the developers

Table 2 presents the results of summarising the four comments provided using different models.

Table 2

EXAMPLES OF COMMENT PROCESSING

Model	Features of the method	Generalisation result (without additional text processing)
Extractive models		
BERT (Bidirectional Encoder Representations from Transformers)	Used for extractive generalisation; assesses the importance of sentences. Not suitable for abstractive generalisation due to its encoder-only architecture and the absence of a generative decoder.	*The result is identical to that of SBERT (see below), but this model is now rarely used. SBERT, on the other hand, has been specifically trained to generate high-quality sentence embeddings optimised for semantic similarity. Unlike 'pure' BERT, SBERT produces more meaningful results because it has been trained on NLI and paraphrasing tasks, rather than simply on token masking.*

Table 2 (Continued)

SBERT (Sentence-BERT)	A modification of BERT for generating semantic sentence embeddings; used for extractive summarisation and finding similar texts. It does not generate new text, but merely selects the most relevant sentences.	With virtual tours, I can visit places I can't afford to travel to in person. What I see doesn't feel as real as actually travelling there. I like the fact that these tours are accessible to people with physical disabilities. The technology still needs improving to provide a fully immersive virtual reality experience.
Abstract models		
GPT (Generativ(Pre-trained Transformer)	A generative model that sequentially predicts words; it produces smooth, narrative-style output.	With virtual tours, I can visit places I cannot afford to travel to in person. What I see doesn't feel as real as actually travelling there. I like the fact that these tours are accessible to people with physical disabilities. The technology still needs improving to provide a fully immersive virtual reality experience. The first step is for users of VR headsets to be able to interact directly with their own devices and experiences without having to worry about how they will react when you walk around or use your headset whilst walking through an area where there is no need at all.
BART (Bidirectional and Auto-Regressive Transformers)	A hybrid encoder-decoder; combines contextual understanding with generative fluency, and works well for abstract tasks.	With virtual tours, I can visit places I cannot afford to travel to in person. What I see doesn't feel as real as actually travelling there. I like that these tours are accessible to people with physical disabilities
T5 (Text-to-Text Transfer Transformer)	Formulates generalisation as a text-to-text transformation task; offers high flexibility and versatility for various NLP tasks.	Virtual tours are accessible to people with physical disabilities. These tours are designed for those with physical limitations – and will provide an immersive experience if this is achieved.
Pegasus (Pre-training with Gap Sentences)	Specialises in abstract generalisation; learns to reconstruct missing key sentences, making it highly effective for creating concise summaries.	I'm a big fan of virtual reality tours.

Table 2 (Continued)

Hybrid (advanced) models		
BigBird (Big and Bold Transformer)	A sparse attention encoder model designed to handle very long texts; it efficiently processes documents containing thousands of tokens, identifying key sentences and structure, and is suitable for extractive tasks and the analysis of large corpora.	With virtual tours, I can visit places that I cannot afford to travel to in person. What I see does not feel as real as actual travel. I like that these tours are accessible to people with physical disabilities. The technology still needs improving to provide a fully immersive virtual reality experience.

Compiled by the author using open-source libraries; the code can be found in Appendix 1
https://gist.github.com/nadiya-dekhtyar/b1b26a40581d3312ca39ddc4c7795108

As can be seen from the table, even the length of the resulting text differs significantly. The GPT model expanded on the existing sentences in the correct context, but the actual comments did not contain half of the suggested ideas; therefore, experts may form a false impression of the true sentiments of potential tourists. T5 placed too much emphasis on people with disabilities, whereas only 25 per cent of the real reviews mentioned this aspect. Pegasus produced only a ‘like/dislike’ type of response (analogous to the Boolean variable ‘yes/no’), meaning it could be inferred that everyone, without exception, finds virtual tours interesting (although one comment was negative, with the visitor writing that virtual tours cannot replace physical travel). The BART model proved to be the closest to the context, but it essentially combined disjointed sentences into a coherent paragraph. Extractive models, on the other hand, repeated all the sentences. However, these comments differ in meaning, so if a researcher considers this particular form of generalisation to be optimal, the model’s reliability must be tested on different volumes of text with varying styles and frequencies of key ideas. It should be noted that free, pre-trained versions of the models were used; therefore, it would be incorrect to choose one over another based solely on such a simple test. However, it is possible to gain a basic understanding of the generalisation results.

An abstractive model functions as an encoder – it analyses text bidirectionally and determines the importance of sentences or fragments. Its main applications are classification, search, question-answering and the selection of key sentences for summarisation. Unlike abstractive models (GPT, BART, T5, Pegasus), the extractive models BERT and SBERT do not generate new text, but merely evaluate and select existing text. Hybrid and extended models, on the other hand, are designed for scaling and handling long or complex contexts.

As the table shows, each model produces different results for the same set of comments, except for BERT, SBERT and BigBird, which gave identical responses. On short articles or comments, SBERT and BigBird will appear almost identical. Nevertheless, on long texts, such as books, BigBird will begin to diverge, because its sparse attention mechanism allows it to model long-range dependencies that SBERT cannot.

To date, there are up to two dozen models that are most widely used. A real boom among developers took place in 2020, and since 2025, a separate sub-sector dedicated to their training, testing and fine-tuning has been operating on a massive scale.

GPT (OpenAI, 2018). The Generative Pre-trained Transformer is the first large-scale autoregressive model for text generation. It aimed to learn to predict the next word based on a large corpus of data, which made it possible to generate text fluently in conversations, creative writing and code. There are free, open-source alternatives (such as DistilGPT-2), but the latest commercial versions, GPT-3.5/4, are available via paid APIs.

BERT (Google AI, 2018). Bidirectional Encoder Representations from Transformers – designed for language understanding tasks such as classification, search and question-answering. It is trained bidirectionally, which allows for better consideration of context, but due to its encoder-only architecture, it is not used for abstract generalisation. The model has open-source, free variants (e.g. bert-base-uncased), and these are widely used in research.

XLNet (Google Brain & Carnegie Mellon, 2019) – an alternative to BERT and GPT, combining autoregressive pre-training with permutations to better model dependencies between words. The aim is to overcome the limitations of BERT in generation and GPT in bidirectional context. Open-source models are freely available, although their use in commercial applications may entail additional costs.

RoBERTa (Facebook AI, 2019) is an optimised version of BERT that removed the NSP task and was trained on a larger dataset. It demonstrated that BERT was "under-trained" and could achieve significantly better results. The model is available free of charge in open-source repositories.

BART (Facebook AI, 2019). The Bidirectional and Auto-Regressive Transformer combines a BERT-like encoder and a GPT-like decoder, making it a hybrid model for abstract generalisation, translation and text generation. It has gained popularity due to its balance between contextual understanding and smooth generation. Free open-source versions are available (e.g., facebook/bart-large-cnn).

T5 (Google Research, 2019). The Text-to-Text Transfer Transformer is a general-purpose model that formulates all NLP tasks in a ‘text-to-text’ format. This provides flexibility for translation, summarisation, classification and QA. The base T5 models are available free of charge, but instruction-tuned variants (such as FLAN-T5) also have open-source versions.

Longformer (Allen Institute for AI, 2020) – designed to handle long documents thanks to sparse attention with linear computational complexity. It aims to process large texts efficiently in QA, classification and summarisation. The model is available free of charge in open-source repositories.

ELECTRA (Google Research, 2020) introduced a new approach to pre-training – a ‘replaced or not’ discriminator – which enables models to be trained more quickly and cost-effectively whilst maintaining high quality. Developed by Google, it is available free of charge for research purposes.

Pegasus (Google Research, 2020) specialises in abstract generalisation, learning to reconstruct missing key sentences. This makes it highly effective for generating concise summaries. The open-source models are available free of charge, although commercial APIs may incur a fee.

ProphetNet (Microsoft Research Asia, 2020) – designed to improve generation by predicting future n-grams, allowing the model to plan the text more effectively. It is used in summarisation and question generation. Free open-source versions are available.

mBART (Facebook AI, 2020) – a multilingual variant of BART, developed for translation and summarisation across multiple languages. It was trained as a denoising autoencoder on multilingual corpora. The model is available free of charge.

BigBird (Google Research, 2020) utilises sparse attention (global, local, random) for long sequences, making it suitable for QA, classification and even genomics. Open-source models are available free of charge.

Switch Transformer (Google Brain, 2021) is a mixture-of-experts architecture with simple routing, which has enabled scaling up to a trillion parameters. It is used primarily in research; open-source versions are available, but commercial applications are subject to a fee.

UL2 (Google Research, 2022). Unified Language Learner, which combines various pre-training paradigms (span corruption, causal LM, prefix LM). It is versatile for fine-tuning and few-shot tasks. Free, open-source models are available.

PaLM (Google Research, 2022). Pathways Language Model – a large-scale model (540B parameters) designed for multi-step reasoning, translation and coding. It is a commercial product, accessible via paid APIs, but the research papers are open access.

FLAN-T5 (Google Research, 2022) is an instruction-tuned variant of T5, trained on thousands of tasks to improve zero-shot/few-shot performance. Open-source, free models are available.

LLaMA (Meta AI, 2023). Meta AI's Large Language Model was created as an open resource for researchers, demonstrating that smaller models can outperform GPT-3. It has become the basis for hundreds of derivative open-source models. LLaMA is available free of charge for research, but commercial applications may be subject to licensing restrictions.

There is a wide range of models and features to choose from. To start with, you can stick to the free options, which is ideal if a travel company does not have in-house machine learning specialists or has limited funds to commission external development. However, in addition to the cost of the model itself, you should factor in the cost of accessing the APIs of the resources where the content to be analysed is stored.

For example, to connect to the YouTube Data API, you need Google Cloud Console keys, which incur a charge once the free trial credits have been used up. While it is possible to retrieve YouTube comments via web scraping without using the official API, direct scraping can be unreliable due to frequent changes in the functionality and code of YouTube pages, and it may breach the terms of service if data is extracted on a large scale. This method analyses the HTML code of the video page or uses a 'headless' browser to load dynamic content (a web browser without a graphical user interface). Clearly, this method is not suitable for ongoing commercial use. However, for research or prototyping purposes, you can use Python code based on the open-source library *youtube-comment-downloader*.

The result of summarising comments using the BART method based on the SAMSum community model for the video 'The Rijksmuseum's Treasures – A Virtual Tour from Rembrandt to Van Gogh' (a virtual tour of the Rijksmuseum in Amsterdam, the Netherlands; URL: https://www.youtube.com/watch?v=mlcqYb-SCpY). The original text of the comments, which was loaded into the model, and the code are contained in Appendix 2 (https://gist.github.com/nadiya-dekhtyar/b1b26a40581d3312ca39ddc4c7795108).

```
"Jessica gave a tour of the Rijksmuseum in Amsterdam. She would like to visit the Van Gogh Museum as well. Her family came to South Africa from Holland, and she loved the tour. She wants Jessica to make a video about the museum."
```

This conclusion is not ideal, as it immediately creates confusion as to which sentences, starting from number 2, refer to the author of the video (it gives the impression that the name "Jessica", which is mentioned, belongs to the author, although in fact this is not the case), and which refer to the users. In this instance, this is important because the country from which the author presumably came is merely an additional biographical detail; however, if tourists from South Africa travelled to the Netherlands specifically as a family to visit a particular museum, this is a signal for tour operators and organisations responsible for promoting the Netherlands' image as a tourist destination worldwide. But overall, it is clear that the commenters were interested in the topic and would have liked to take a virtual tour of another museum (for example, the Van Gogh Museum).

The result obtained using the T5-Large model is slightly different. This model provides better generalisation quality but requires more resources (there are options: t5-small – fast but lower quality; t5-base – a balance between speed and quality; and t5-large). However, this model is not perfect either, as the resulting text contains syntactic errors and is not logically structured.

"The Van Gogh Museum is my favourite artist, and has been for more than two decades since I studied art at university. Please do an episode on the Van Gogh Museum as well. The effort you put into this is evident."

The output generated using the GPT-2 Large model (the freely available version was selected for testing) is already a coherent, logically structured text; however, it lacks general information about foreign tourists who have visited Amsterdam. None of the models tested noted the presence of a commenter from Canada, which, in the long term, might suggest the popularity of virtual tours of European museums amongst North Americans; however, this user's comment did not contain a comment as such (a review of the video or an expression of impressions), and was therefore likely overlooked.

"Viewers express great admiration for the museum tour and the presenter's engaging and educational style. Many people praise the quality of the video, the beautiful portraiture, and how enjoyable and informative the experience was. Several commenters mention their personal connections to Amsterdam and the Rijksmuseum, sharing how meaningful the visit felt to them. A large number of viewers have requested a future episode about the Van Gogh Museum and expressed excitement at the prospect of learning more. Overall, the comments demonstrate strong appreciation, an emotional connection and enthusiasm for future tours."

Thus, the use of AI models to track the expectations of consumers of tourism services can be carried out in parallel with monitoring potential negative practices aimed at discrediting Ukraine's image amongst global internet users. It is possible to economise during the initial data collection stage, as it is advisable to analyse foreign resources (or foreign-language sources) where Ukraine is mentioned first and foremost – the country's name will serve as one of the keywords. If a large list of keywords is not entered at the outset, this corpus of text data will consist of posts concerning assessments of the country's image, likely its foreign policy, prospects for recovery, compilations of historical facts, descriptions of cultural heritage, and so on. In other words, any resources on such topics will be useful for understanding the sentiments held by foreigners towards our country. For tourism industry professionals, the priority is to study the spontaneous process of a country's image formation and to filter out the most popular attractions, phenomena and regions, which are mentioned most frequently, and from which a ready-made offering can be developed – initially in the form of virtual tours and other recreational digital services, with itineraries planned for physical visits once hostilities have ceased. For security experts, this set of information is needed to identify potential manipulation of users' opinions regarding content or manipulation of facts published in reputable media. The language models, which will be used in subsequent stages for detailed semantic analysis of texts, are likely to differ for the two groups of specialists and will be trained on different examples; however, where possible, it is worth saving time and the cost of API access keys if several research groups require an almost identical input dataset.

Ukraine already has significant potential for AI application. In the 2025 Government AI Readiness Index, the country ranked 41st out of 195 countries, scoring 60.84. The USA ranked 1st with a score of 88.36, and the lowest score was 12.12 [13].

Table 3

PILLARS OF THE GOVERNMENT AI READINESS INDEX

Pillar	Ukraine	Top-10 / leaders*
Total score	60.84	(1) United States of America – 88.36; (2) France – 80.81; (3) United Kingdom – 77.75; (4) Netherlands – 77.18; (5) South Korea – 76.89; (6) Germany – 76.78; (7) Singapore – 76.42; (8) China – 76.27; (9) Australia – 75.73; (10) Norway – 74.84
Policy capacity	73,5	100.0 – Australia, Egypt, Serbia, United Kingdom; 96.0 – Estonia, India, Poland, South Korea; 92.5 – Canada, Chile, China, Denmark, USA
AI Infrastructure	52,4	(1) USA – 91.77; (2) China – 78.36; (3) France – 76.27; (4) Singapore – 74.46; (5) UAE – 73.77; (6) Germany – 73.28; (7) United Kingdom – 72.49; (8) Canada – 72.32; (9) Japan – 70.62; (10) Australia – 70.02
Governance	83	(1) Ireland – 94.57; (2) Netherlands – 94.19; (3) Norway – 94.02; (4) Denmark – 94.02; (5) Portugal – 93.55; (6) Saudi Arabia – 93.21; (7) United Kingdom – 92.27; (8) Lithuania – 91.67; (9) Germany – 91.17; (10) China – 91.14
Public sector adoption	80,24	(1) Estonia – 99.63; (2) UAE – 97.27; (3) France – 97.18; (4) Brazil – 94.78; (5) Netherlands – 94.01; (6) USA – 92.87; (7) Singapore – 92.50; (8) Peru – 90.40; (9) Saudi Arabia – 88.62; (10) Thailand – 85.86
Development & diffusion	40,34	(1) USA – 88.16; (2) China – 74.67; (3) Canada – 69.03; (4) United Kingdom – 68.46; (5) France – 68.20; (6) Germany – 66.24; (7) Japan – 66.13; (8) South Korea – 65.88; (9) Israel – 64.66; (10) Singapore – 64.15
Resilience	58,21	(1) Denmark – 90.01; (2) France – 88.63; (3) United Kingdom – 83.30; (4) Australia – 82.94; (5) South Korea – 82.71; (6) Spain – 81.70; (7) Japan – 80.75; (8) Austria – 79.69; (9) USA – 78.28; (10) Norway – 77.50

* Sometimes, countries share the same scores, so either the top ten national economies or the greater number of leaders is indicated

Source: compiled by the author on the rankings data

The key success indicators for Ukraine were the development of a sovereign AI infrastructure, the integration of AI into core government platforms (including Diia), and the adoption of MilTech. As for consumers/citizens' AI usage, there is no unique source of "official" statistics. It can only literally be collected by service providers, who obtain IP addresses and/or subscription details as the primary source of users' characteristics. However, this can distort a user's citizenship. Furthermore, the MFA assessed that 8 million Ukrainians were living abroad as of May 2026 [14], who are probably using the digital infrastructure of their current country of residence. In this respect, the GovTech sector collects the most reliable statistics. Nevertheless, AI services extend far beyond governmental services, so market observations revert to questionnaires as the easiest way to gain "first-hand access" to people's choices. According to experts, the number of people using AI in Ukraine increased from 28% in 2024 to 49% in 2025. Of these users, 23% use AI for work, 17% for study, and 28% for other purposes. 42% of Ukrainian adults and 70% of Ukrainian teenagers use it regularly [15]. Further detailed research is required on statistics regarding the use of artificial intelligence by sector. For instance, tourism professionals would be well advised to survey their target audience, since content about foreign travel and culture is always of interest to those who have limited opportunities to travel or pay for virtual tours.

The findings obtained from the analysis of consumer preferences, media reports and trends expressed by authoritative bodies in the tourism sector, as well as proposals and reports from travel companies, can either form the basis of strategies in the form of qualitative recommendations (following consultation with experts) or be converted into quantitative inputs for other models. For example, consumer feedback on watching themed videos about tourist attractions can be adapted to a rating scale and used to develop the storyline, duration, emotional tone, content, the use of formal/high-quality language (rather than relying solely on automatic machine translation) in the release, and other characteristics, for which recommendations can be presented in a simplified rating format. The results of analysing statistical macroeconomic data, fed into econometric or language models trained using the necessary mathematical procedures, can be summarised as a set of recommendations for entrepreneurs planning to export tourism services abroad. It is important to understand that, for the smooth transfer of sets of resulting data and the retraining of models, constant communication is required between various groups of specialists: some of whom work, for example, within city tourism authorities; others form part of the Ministry of Foreign Affairs team responsible for image-building initiatives; whilst others work at the State Tax Service of Ukraine and must calculate the impact of reducing income tax for a selected priority sector as part of a multi-year government initiative which is entirely unrelated to the activities of tourism enterprises, but is linked to other sectors whose development already has a direct impact on the quality of tourism infrastructure and the number of travellers.

Conclusion. As demonstrated, various AI models can be used to carry out specific tasks involving the analysis of the opinions of potential consumers of tourism products. However, such monitoring is intended to address more important objectives – the search for ideas for national rebranding and the identification of existing hybrid threats. Therefore, strategic planning for the creation of a research infrastructure for analysing the content of relevant resources must take into account not only the technical aspects of the interaction between information flows generated by various databases, but also the deployment of so-called ‘corrective nodes’ – analysts who will verify the quality and comprehensiveness of incoming information and the results of machine algorithm calculations, confirm the urgency of measures and recommend optimal ways to optimise sectoral strategies.

References

1. Antczak, B. "The impact of artificial intelligence on the tourism industry: A marketing perspective." *Journal of Modern Science,* 2025, issue 62(2), pp. 450–472, https://doi.org/10.13166/jms/208128. Accessed 13 July 2026.
2. Dyduch, W., & A.Brzozowska. "Artificial intelligence in tourism management: Theoretical underpinnings, empirical tests and the SmartTourAI framework." *Central European Management Journal,* 2026, issue 34(2), pp. 242–258, https://doi.org/10.1108/CEMJ-07-2024-0227. Accessed 13 July 2026.
3. Katsanakis, I., Zouni, G., Spasos, S., Lykoudi, D. M., & N.Akriotis. "Exploring the impact of artificial intelligence on the tourism industry: Opportunities and challenges." *In GMA-GAMMA Joint Symposium. Springer Nature Switzerland,* 2025, pp. 154–178, https://doi.org/10.1007/978-3-031-96930-0_9. Accessed 13 July 2026.
4. Zhang, M., Yi, F., & D. Gursoy. "The effects of generative artificial intelligence on consumers in hospitality and tourism: A systematic review and future research directions." *International Journal of Hospitality Management,* 2026, issue 133, art. 104452, https://doi.org/10.1016/j.ijhm.2025.104452. Accessed 13 July 2026.
5. Seyfi, S., Kim, M. J., Nazifi, A., Murdy, S., & T.Vo-Thanh. "Understanding tourist barriers and personality influences in embracing generative AI for travel planning and decision-making." *International Journal of Hospitality Management,* 2025, issue 126, art. 104105, https://doi.org/10.1016/j.ijhm.2025.104105. Accessed 13 July 2026.
6. Štilić, A., Puška, A., & M. Nicić. "The role of artificial intelligence in shaping the future of the travel industry: An expert analysis of artificial intelligence-generated travel itineraries." *DETUROPE – The Central*

European Journal of Regional Development and Tourism, 2025, issue 16(2), pp. 57–79, deturope.eu/pdfs/det/2024/02/04.pdf. Accessed 13 July 2026.

7. Koo, I., Zaman, U., Ha, H., & S. Nawaz. "Assessing the interplay of trust dynamics, personalisation, ethical AI practices, and tourist behaviour in the adoption of AI-driven smart tourism technologies." *Journal of Open Innovation: Technology, Market, and Complexity*, 2025, Issue 11(1), art. 100455, https://doi.org/10.1016/j.joitmc.2024.100455. Accessed 13 July 2026.
8. López-Naranjo, A. L., et al. "Artificial intelligence in the tourism business: a systematic review." *Frontiers in Artificial Intelligence,* 2025, issue 8, art. 1599391, https://doi.org/10.3389/frai.2025.1599391. Accessed 13 July 2026.
9. Hall, C. M., & C. Cooper. "Making tourism smart in the age of artificial intelligence." *Current Issues in Tourism*, 2025, issue 28(1), pp. 1–5, https://doi.org/10.1080/13683500.2025.2460922. Accessed 13 July 2026.
10. *Metaverse Privacy and Safety*. July 2023 Edition. WEF Forum, www.weforum.org/publications/privacy-and-safety-in-the-metaverse/ Accessed 13 July 2026.
11. Glybovets, A., & V. Tochytskyi. “Tokenisation and stemming algorithm for the Ukrainian language.” *Scientific Notes of NaUKMA (Computer Science),* vol. 198, 2017, pp. 4–8, ekmair.ukma.edu.ua/server/api/core/bitstreams/a756c821-adae-4dda-8642-712e278bff3e/content. Accessed 13 July 2026.
12. Pylypiv, I. "‘Syayvo’ did not launch on schedule. How Ukraine is developing its national AI." *Ekonomichna Pravda,* 15 June 2026, 08:41, epravda.com.ua/tehnologiji/stvorennya-ukrajinskoji-llm-syayvo-progres-vikliki-ta-novi-termini-zapusku-822657/ Accessed 03 August 2026.
13. *Government AI Readiness Index*. 2025 ed. *Oxford Insights,* oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025/. Accessed 03 August 2026.
14. Prots, A. "Number of Ukrainians forced abroad by war reaches 8 million", Ukrainska Pravda, 29 May 2026, www.pravda.com.ua/eng/news/2026/05/29/8036931/ 03 August 2026
15. “Almost half of Ukrainians use artificial intelligence.” U*krinform, 1*3 March 2026, www.ukrinform.ua/rubric-society/4101218-majze-polovina-ukrainciv-koristuetsa-stucnim-intelektom.html. 03 August 2026

Submission date: 14.07.2026
Date of acceptance for publication: 04.08.2026