

Article

Top2Vec Topic Modeling to Analyze the Dynamics of Publication Activity Related to Environmental Monitoring Using Unmanned Aerial Vehicles

Vladimir Albrekht¹, Ravil I. Mukhamediev^{1,2} , Yelena Popova³ , Elena Muhamedijeva² and Asset Botaibekov^{1,*}

¹ Institute of Automation and Information Technologies, Satbayev University (KazNRTU), 22 Satpayev Street, Almaty 050013, Kazakhstan; sluminat2@gmail.com (V.A.); ravil.muhamedyev@gmail.com (R.I.M.)

² Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan; muhamedijeva@gmail.com

³ Transport and Management Department, Transport and Telecommunication Institute, Lauvas iela 2, LV-1003 Riga, Latvia; yelenagp@gmail.com

* Correspondence: asset101@gmail.com

Abstract: Unmanned aerial vehicles (UAVs) play a key role in the process of contemporary environmental monitoring, enabling more frequent and detailed observations of various environmental parameters. With the rapid growth of scientific publications on this topic, it is important to identify the key trends and directions. This study uses the Top2Vec algorithm for topic modeling algorithm aimed at analyzing abstracts of more than 556 thousand scientific articles published on the arXiv platform from 2010 to 2023. The analysis was conducted in five key domains: air, water, and surface pollution monitoring; causes of pollution; and challenges in the use of UAVs. The research method included data collection and pre-processing, topic modeling, and quantitative analysis of publication activity using indicators of the rate (D1) and acceleration (D2) of change in the number of publications. The study allows concluding that the main challenge for the researchers is the task of processing data obtained in the course of monitoring. The second most important factor is the reduction in restrictions on the UAV flight duration. Among the causes of pollution, agricultural activities will be considered as a priority. Research in monitoring greenhouse gas emissions will be the most topical in air quality monitoring, while erosion and sedimentation—in the area of land surface control. Thermal pollution, microplastics, and chemical pollution are most relevant in the field of water quality control. On the other hand, the interest of the scientific community in topics related to soil pollution, particulate matter, sensor calibration, and volatile organic compounds is decreasing.

Keywords: UAVs; topic modeling; Top2vec; environmental monitoring; trend analysis; scientific publications; scientometrics



Academic Editors: Andrew Kirby and Tindaro Cicero

Received: 20 November 2024

Revised: 10 March 2025

Accepted: 12 March 2025

Published: 25 March 2025

Citation: Albrekht, V., Mukhamediev, R. I., Popova, Y., Muhamedijeva, E., & Botaibekov, A. (2025). Top2Vec Topic Modeling to Analyze the Dynamics of Publication Activity Related to Environmental Monitoring Using Unmanned Aerial Vehicles.

Publications, 13(2), 15. <https://doi.org/10.3390/publications13020015>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background and Significance

In recent years, unmanned aerial vehicles (UAVs) have been increasingly used in various spheres of manufacturing, entertainment, and monitoring (Shakhatareh et al., 2019). UAV technologies are used to improve the efficiency of agriculture (Mogili & Deepak, 2018), in mining and mineral exploration (Park & Choi, 2020), for magnetic exploration of volcanoes (Gailler et al., 2021), for organization of emergency communications (Erdelj et al., 2017), monitoring of urban agglomerations (Hu et al., 2022), etc. In many cases, UAVs solve the tasks of environmental monitoring, including in combination with the satellite

systems of remote sensing of the Earth's surface. Applications are diverse, ranging from monitoring forest fires (Hodgson et al., 2016) and wildlife (Yuan et al., 2023) to studying the effects of pollution on marine ecosystems (Gulf News, 2018). UAV monitoring of environmental disturbances is an important area of research and practical applications (Fascista, 2022). According to the literature, there is an urgent need for global large-scale monitoring to counteract environmental pollution, as well as for the integration of various monitoring technologies to establish comprehensive environmental surveillance systems (Hu et al., 2022).

With the growing number of scientific publications devoted to the application of UAVs for environmental monitoring, there is a need to systematize them and, if possible, to assess the trends in this area of scientific research. The volume of scientific publications is constantly growing, which requires the adaptation of analysis methods and the use of new metrics for evaluating publication activity (Gupta et al., 2022). One of the methods of such analysis is topic modeling, which allows us to identify hidden thematic structures in the corpora of scientific texts and to assess the dynamics of their change (Muhamedyev et al., 2018). There are several thematic modeling algorithms, of which the Top2Vec algorithm is chosen for application in this paper.

The aim of this research is to identify the main topic groups of scientific papers devoted to the application of UAVs in environmental monitoring, as well as to estimate and predict changes in their volumes using the metrics of growth rate and acceleration (Mukhamedyev et al., 2019; R. I. Mukhamediev et al., 2021). A total of 556,000 abstracts from computer science papers published on arXiv between 2010 and 2023 were collected and analyzed. We focused on abstracts rather than full papers since they provide concise representations of research content while being well-suited for large-scale topic modeling analysis. This comprehensive dataset allows identifying both rapidly growing thematic groups and those where publication activity is declining.

The main distinguishing feature of this research is a quantitative assessment of trends in publication activity. This analysis allows recognizing both the fastest-growing thematic groups and those groups where the number of publications is not growing. We examined five main domains (air pollution, water pollution, surface pollution, causes of pollution, and monitoring challenges) and identified the most popular research areas in terms of publication activity growth.

The results of the analysis can serve as a guideline for selecting science-intensive research areas or for selecting technologies where the results suitable for practical application have already been obtained.

The article is arranged as follows (see Figure 1):

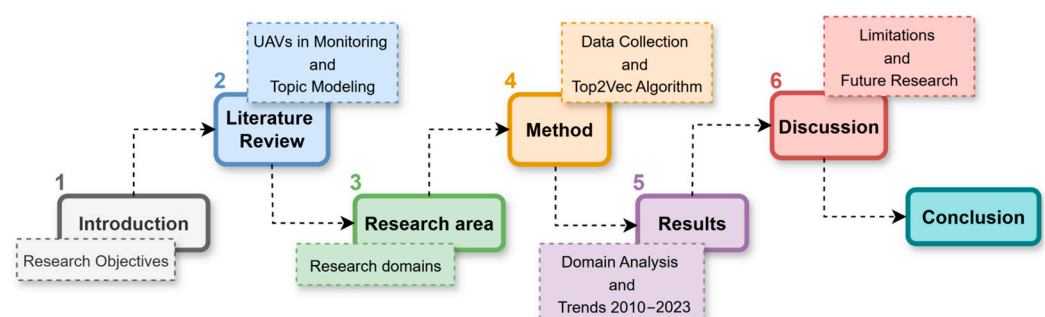


Figure 1. Article arrangement.

Section 2 provides an overview of the research articles on the application of UAVs for environmental monitoring and thematic modeling in the analysis of scientific publications and briefly summarizes the research area.

Section 3 describes the methods and approaches used in the study, including the data collection procedure, keyword selection, and principles of topic modeling with the employment of the Top2Vec algorithm.

Section 4 presents the results of the domain analysis (see Figure 1), including the dynamics of interest in different aspects of UAV applications for environmental monitoring for the period from 2010 to 2023.

Section 5 discusses the results of the study and its limitations.

Conclusions (Section 6) summarizes the findings and suggests directions for future research.

1.2. Literature Review

Monitoring tasks have become one of the most common applications of UAVs. Among the applications of UAV-based monitoring, we can mention the monitoring of geophysical processes (Erdelj & Natalizio, 2016; Hervás et al., 2003; Aljehani & Inoue, 2019), wildlife (Hodgson et al., 2016), technical and engineering structures (Jenssen & Roverso, 2019; Jordan et al., 2018; Ham et al., 2016), road traffic (Khan et al., 2020) and urban environment in general (Bayomi & Fernandez, 2023), water bodies (Medvedev et al., 2020), environmental pollution (Gulf News, 2018; R. Mukhamediev et al., 2020a), and many others. A generalized scheme of UAV applications formed on the basis of the analysis of several review publications (Mohsan et al., 2023; Telli et al., 2023; Erkec & Hajiyeve, 2022; Mohsan et al., 2022) is shown in Figure 2.

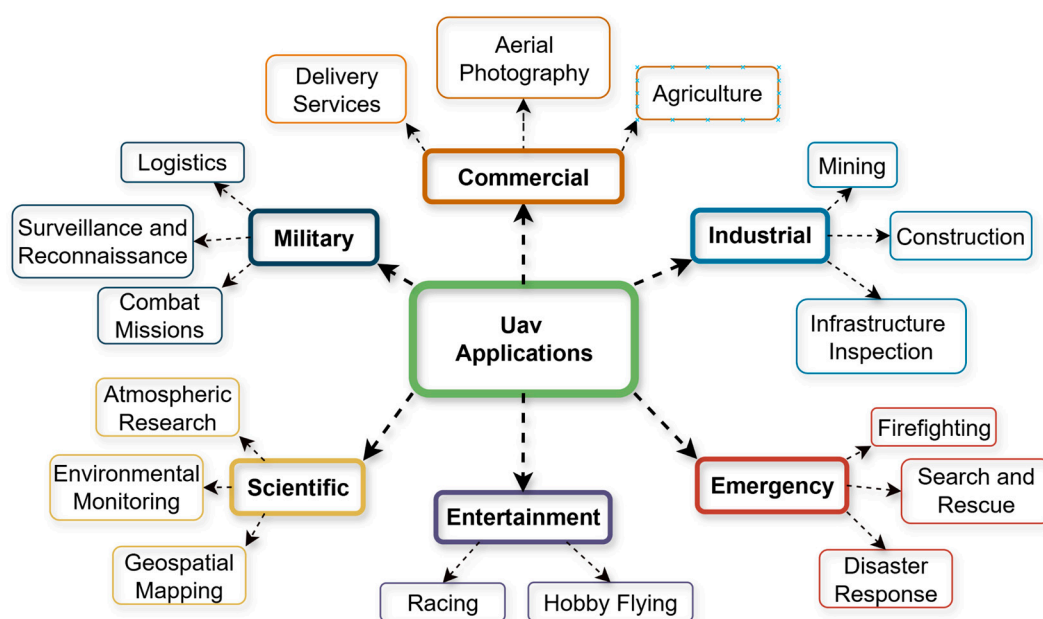


Figure 2. Employment of UAV applications.

Each of these areas of monitoring raises some corresponding scientific tasks. Some areas demonstrate a rapid growth of publication activities, while others, on the contrary, reflect a decline in the interest of the scientific community; as a rule, it happens due to the completion of research and the transition to wide practical use of their results. Reviews of scientific papers are widely used to assess such processes. The review articles on UAVs discuss UAV monitoring applications in various environments such as air (Lambey & Prasad, 2021), water (Yuan et al., 2023), urban environments (Mohamed et al., 2020), transportation flows (Butilă & Boboc, 2022), and agriculture (Barbedo, 2019). One of the most cited review papers considers UAVs as monitoring systems of general purpose, and the associated challenges, use cases, and research trends are analyzed for each application (Shakhatreh

et al., 2019). In general, it can be stated that the most cited review papers are aimed at the analysis of the existing application options, remaining challenges and limitations of use, and future research directions based on a rather limited list of scientific papers.

However, in a situation when the number of scientific publications increases dramatically, the manual processing of even the most significant of them becomes very difficult. In this regard, the automatic analysis of publication corpora is useful, for example, using various topic modeling models (Kherwa & Bansal, 2020). Topic modeling is one of the popular methods of the generalization and structuring of large text archives, which is used in media monitoring (R. I. Mukhamediev et al., 2020b) and bibliometrics. Several algorithms are used to implement the topic model of a corpus of texts (see Figure 3).

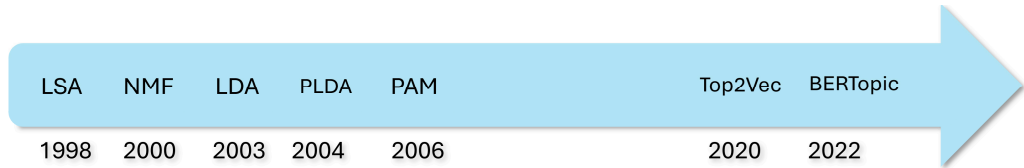


Figure 3. Topic modeling timeline.

Latent Dirichlet Allocation (LDA) is an undisputed leader in this area (Blei et al., 2003; Jelodar et al., 2019). As noted by Vayansky and Kumar (2020), although there are many approaches to topic modeling that take into account different relations and constraints, LDA is the most often used one in practice. An extension of LDA, BigARTM (Vorontsov et al., 2015), is used as a software implementation of the method. LDA is based on the statistical analysis of words in a corpus of texts. However, since the emergence of large language models that use vector representations of words and texts, such as word2vec (Mikolov et al., 2013), BERT (Devlin et al., 2019), etc., the possibility of using deeper semantic models has emerged. For example, in (Zengul et al., 2023) the authors compare three topic modeling methods: latent semantic indexing (LSA), LDA, and Top2Vec (Angelov, 2020) using a corpus of 65,292 abstracts devoted to COVID-19. The study shows a high level of correlation between the results of LDA and Top2Vec, indicating that they may be inter-changeable. It is observed that when dealing with large datasets and the need for automatic topic detection, both methods can produce similar results. However, when working with limited computational resources or when rapid analysis is required, researchers may prefer LSA. The authors argue that researchers should consider factors such as the ease of use, computational resources required, and availability of analytical tools when choosing between the two methods. In the context of the study, powerful research computing resources were required to work with LDA and Top2Vec since traditional desktop computers were insufficient. The authors recommend the use of Python 3.9 and research computing resources for projects related to LDA and Top2Vec, while LSA can be run on a regular desktop computer and requires significantly less time to process data.

In (García et al., 2024) the authors compare LDA, Top2Vec, and BERTopic (Grootendorst, 2022). The study is based on analyzing the texts from Weibo and Twitter social networks focused on the topic #ChatGPT. According to the results, BERTopic showed the best topic separation ability, providing a high degree of independence between topic clusters and clarity of semantics. These findings point to the potential effectiveness of BERTopic for analyzing social media data, where texts are often less formalized and more diverse. Nevertheless, when transferring these methods to the analysis of scientific texts, it is necessary to take into consideration the peculiarities of academic language and data structure, which may affect the choice and adjust the topic modeling method for the best result achievements. The interest in the development of topic modeling methods causes the emergence of new approaches that can significantly reduce computational costs in some

cases. In particular, in (Bretsko et al., 2023) the problem of clustering topics of scientific articles is investigated. The authors apply two approaches: a traditional method based on network science using the Combo algorithm to detect communities in the network by subjects, and a modern method based on transformers using the Top2Vec algorithm for embedding the content of articles. The study was conducted on a dataset obtained from the scientific papers devoted to computer science and mathematics and published in the journals indexed in the SCOPUS database, and various coherence measures were used to evaluate performance. Coherence in this context refers to the degree of coherence and cohesion of the extracted topics, in other words, how clearly and logically coherent are the topics represented as a result of clustering. The Combo algorithm was able to achieve a level of coherence comparable to the transformer-based Top2Vec. Hence, the community detection method can be a viable alternative for topic clustering, especially when high coherence and fast processing time are required.

The issue of the effectiveness of topic modeling methods in the analysis of short texts is very relevant due to the limited access to large text archives of scientific publications and the limited power of computing systems. In (Muthusami et al., 2024) the authors combine classical methods LDA for probabilistic topic modeling and non-negative matrix factorization (NMF) for non-probabilistic topic modeling. In (Egger & Yu, 2022), the advantages and disadvantages of four popular topic modeling methods: LDA, NMF (Lee & Seung, 1999), Top2Vec, and BERTopic are discussed. The authors of the paper studied tweets and found that BERTopic and NMF are particularly effective for processing short texts typical for social networks. BERTopic is highly accurate in extracting and interpreting the topics due to its deep understanding of context, while NMF demonstrates fast partitioning of data into topic clusters. At the same time, LDA and Top2Vec, despite their widespread use, showed limitations in accuracy and speed in processing short messages.

Identifying thematic clusters or groups is only one of the steps toward understanding large text corpora. After identifying the major topics of a text corpus, it is useful to identify their change trends. This will provide an understanding of the changing attention of the scientific community to certain thematic groups. Such a task is formulated in (Gulf News, 2018), devoted to the dynamics of publication activity in AI and ML, for which we introduced and analyzed the rate of change (D1) and acceleration of change (D2) of publication activity using Google Scholar data. The negative correlation between D1, D2, and the number of articles emphasizes the growing attention to understudied research areas. The method proposed in the aforementioned work identified rapidly growing research areas and made short-term predictions of publication activity with a mean prediction error of 6% and a standard deviation of 7%. However, in this work, the selection of articles was performed using keywords in the Google search engine, which does not fully take into account the semantics of articles. Obviously, it is possible to obtain more accurate results if we use thematic modeling to identify thematic groups of texts close in meaning.

This paper considers the abstracts of publications related to the application of UAVs for environmental monitoring as an object of research; these publications should be indexed in one of the large open databases of scientific publications—arXiv. It means that the corpus of texts contains documents much larger than tweets. Therefore, it seems useful to apply Top2Vec to analyze this corpus. Then, there can be evaluated the dynamics of changes in publication activity of the obtained topic clusters using the approach described in (Gulf News, 2018).

2. Study Area

In any of the studies dealing with the review of scientific publications, there arises the difficult issue of limiting the field of the study. In the case of this study, since it is

planned to use automated analysis tools, this field can be considered as broadly as possible without being limited by the number of publications. However, even in this case, some systematization of issues of monitoring with UAV usage is necessary. Of course, one could choose arXiv categories such as machine learning, Computer Vision, Artificial Intelligence, and others as research domains. However, these categories are not directly semantically related to the objectives of this research. Moreover, the monitoring issues are often at the intersection of these categories. Instead of this, we would like to analyze the corpus of publications firstly, from the point of view of monitoring applications, secondly, from the point of view of problems and challenges that generate new scientific ideas, and thirdly, from the point of view of possible reasons raising the need for monitoring. In these areas, similar to the classical review papers, we would like to identify the most promising (rapidly growing) research topics.

In this regard, it is assumed to consider these issues, first of all, from the point of view of the monitored environment. As abstract as possible, this is the surface of the earth, air, and water environments. Secondly, it is important to identify the causes of pollution and, finally, to consider the limitations of such monitoring. Based on these considerations, a hierarchical scheme of the study area was formed (see Figure 4).

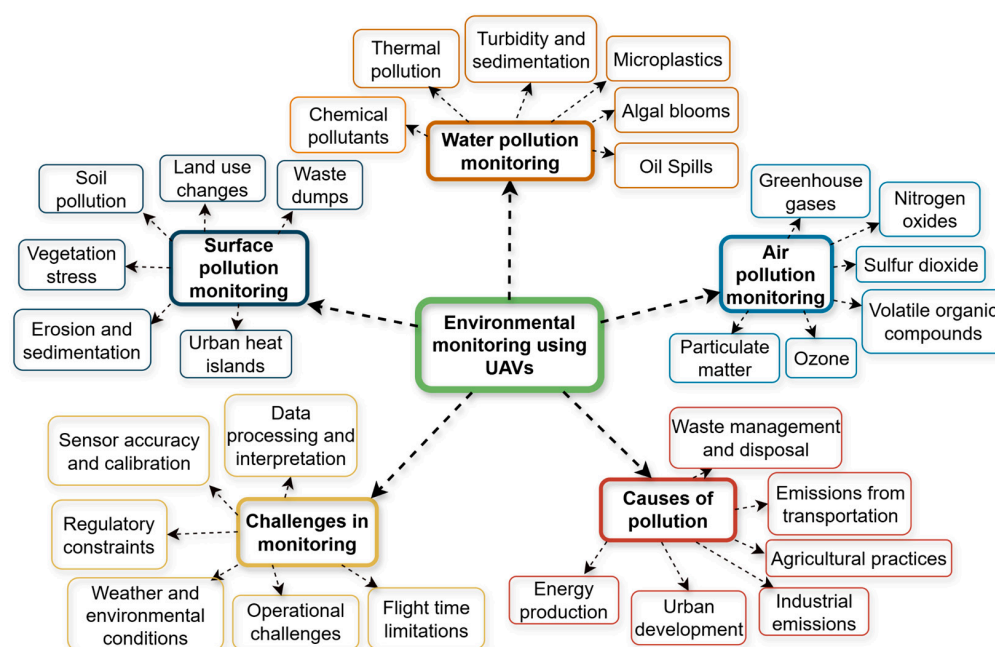


Figure 4. Study area.

The schematic representation in Figure 4 includes five main domains:

1. Water pollution monitoring;
2. Air pollution monitoring;
3. Surface pollution monitoring;
4. Causes of pollution;
5. Problems in pollution monitoring.

The abovementioned five key domains are the target of the analysis. The choice of these five domains is rather intuitive and is based especially on common sense and the key areas of analysis noted in the highly cited review papers mentioned above (Shakhatreh et al., 2019).

Each domain contains several thematic groups, presented by rectangles in the diagram. These groups reflect specific aspects or problems related to the respective domain. For example, the domain “Water Pollution Monitoring” includes the following thematic

groups: “Microplastics”, “Algal Blooms”, “Oil Spills”, “Turbidity and Sedimentation”, “Thermal Pollution”, and “Chemical Pollutants”. Topic groups are formed from semantic clusters obtained using the Top2Vec algorithm. The method of thematic group formation is described below in Section 4.

3. Materials and Methods

The methodological design of the study includes 4 main steps (see Figure 5):

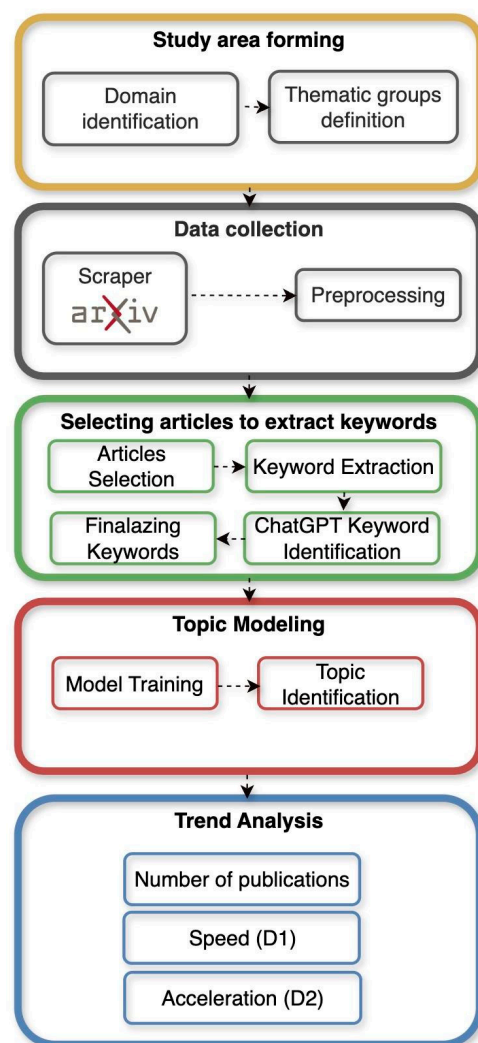


Figure 5. Methodological design of the study.

The research begins with “Study area forming” (see Section 2), where the five key domains and their respective thematic groups are defined as presented in Figure 4. This step establishes the conceptual framework for the entire study, identifying the specific areas of environmental monitoring to be analyzed. The second includes “Data Collection”, where a scraper is used to extract the abstracts of scientific articles placed on the arXiv platform. After selecting the articles, too short abstracts containing only service information (preprocessing) are removed. The next stage is “Selection of articles for keyword extraction”, which selects relevant articles for further analysis and extracts keywords that serve as a basis for thematic modeling. The main stage of the research is “Topic Modeling”, which includes “Model Training” using the Top2vec algorithm and “Topic Identification” to identify the key topic areas in the corpus of research articles. The final stage “Trend Analysis” includes the construction of regression models of changes in the number of publications in thematic

groups and estimation of the “Rate of Change” (D1), “Acceleration of Change” (D2), and “Average Number of Publications” of each thematic group. The mentioned stages of the research are described in more detail below.

3.1. Data Collection

The arxivscraper library ([Sadjadi, 2017](#)) for Python was used to extract article abstracts and metadata from arXiv. For each article, we collected its abstract, title, unique identifier, categories, DOI identifier, submission and last update dates, authors’ list, and a link to the full text. This tool allowed for setting the time frame (from 2010 to 2023) and filtering the data by the category “Computer Science (cs)”, corresponding to the focus of this study.

3.1.1. Data Preprocessing

Top2vec does not require complex text preprocessing such as stemming, lemmatization, or stop word removal. The model is able to handle “raw” texts efficiently by extracting semantic relations directly from natural language texts. The only preprocessing step was the exclusion of articles with abstract lengths less than 320 characters since short abstracts contain only service information that is not relevant to the purpose of this analysis.

3.1.2. Text Corpus

We collected all computer science abstracts from arXiv between 2010 and 2023 rather than pre-filtering for UAV-specific papers, as Top2Vec performs better with a broader corpus that captures the full semantic context of the field. This approach allows the model to better understand relationships between UAV-related and other technical concepts while also capturing relevant papers that might discuss drone technologies or environmental monitoring without explicitly using UAV terminology.

The final corpus consists of 556,000 documents (abstracts) with the following characteristics:

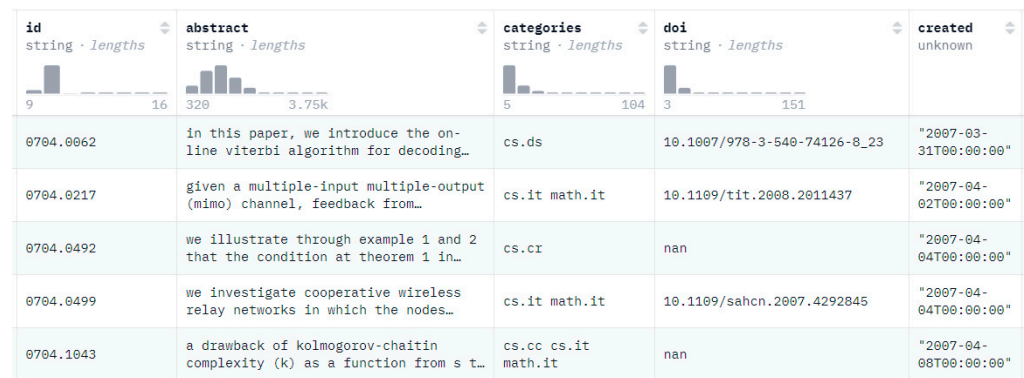
- Time period: 2010–2023;
- Average abstract length: 956 characters.

Primary category distribution: Computer Vision (cs.cv, 15%), machine learning (cs.lg, 13.4%), and Computational Linguistics (cs.cl, 7.5%). The top 20 categories are visualized in Appendix A, Figure A1, and the complete distribution of all 156 categories is provided in the Supplementary Materials.

Each record in the corpus is structured with the following metadata:

- Article title;
- Unique identifier;
- Abstract text;
- arXiv categories;
- DOI identifier;
- Submission and last update dates;
- Authors’ list;
- Link to full text;
- Abstract length;
- Sequential identifier.

Figure 6 shows an example entry from the corpus. The complete dataset, including all metadata and abstracts, is available on the Hugging Face platform ([CCRss, 2024a](#)) to ensure the reproducibility of our analysis.



id	abstract	categories	doi	created
string · lengths	string · lengths	string · lengths	string · lengths	unknown
9	320	5	3	
16	3.75k	104	151	
0704.0062	in this paper, we introduce the on-line viterbi algorithm for decoding...	cs.ds	10.1007/978-3-540-74126-8_23	"2007-03-31T00:00:00"
0704.0217	given a multiple-input multiple-output (mimo) channel, feedback from...	cs.it math.it	10.1109/tit.2008.2011437	"2007-04-02T00:00:00"
0704.0492	we illustrate through example 1 and 2 that the condition at theorem 1 in...	cs.cr	nan	"2007-04-04T00:00:00"
0704.0499	we investigate cooperative wireless relay networks in which the nodes...	cs.it math.it	10.1109/sahcn.2007.4292845	"2007-04-04T00:00:00"
0704.1043	a drawback of kolmogorov-chaitin complexity (k) as a function from s t...	cs.cc cs.it math.it	nan	"2007-04-08T00:00:00"

Figure 6. Example of an entry in a corpus of texts.

3.2. Keyword Extraction Process

Our keyword extraction involved a hybrid approach combining author-defined keywords with semantic analysis. The initial examination of author-defined keywords revealed three key challenges:

- Inconsistent availability—many papers lacked author-defined keywords entirely.
- Format variations—similar concepts were expressed differently across papers, for example:
 - ‘UAV monitoring’, ‘drone monitoring’, ‘UAV-based monitoring’;
 - ‘environmental sensing’ versus ‘environmental monitoring’;
 - Various forms of hyphenation and compound terms.
- Semantic gaps—important related concepts were often missing.

To address these challenges, we developed a standardization approach using ChatGPT (gpt-4-0125-preview). While traditional keyword extraction methods like RAKE (Rose et al., 2010) and YAKE (Campos et al., 2020) excel at statistical analysis of term frequency, we chose ChatGPT for its ability to understand semantic context and relationships between terms, which aligns better with the semantic nature of our Top2Vec model.

For standardizing and extracting keywords, we provided ChatGPT with the following instructions:

- Analyze the provided keywords from documents within the topic.
- Identify the most representative and recurring keywords.
- Exclude any variations of ‘UAV’, ‘unmanned aerial vehicle’, or ‘drones’. For multi-word keywords, split them into single words that are commonly recognized in the field.
- Select the top five keywords that best capture the essence of the topic, ensuring they are single words.
- Format Keywords: List the selected keywords in this format: [“keyword1”, “keyword2”, “keyword3”, “keyword4”, “keyword5”].
- Provide Keywords Only: Respond with the formatted list of keywords, without any additional sentences or explanations”.

After receiving the list of keywords from ChatGPT, additional manual analysis was conducted to refine and complete the list by identifying common keywords for each of the domains. Keyword relevance was determined based on the following criteria:

- Relation to the research topic: keywords directly related to the main aspects of UAV applications for environmental monitoring were prioritized.
- Uniqueness: rare words and phrases that occurred in the abstracts of the selected articles were considered as more meaningful and more accurately reflecting the features of the document.

- Semantic cohesion: keywords that had a strong semantic connection with other keywords in the group were considered as more relevant.

As a result, a final list of keywords for each domain was generated:

3.2.1. Monitoring Air Pollution Using UAVs

Common keywords for the domain: [uav, monitoring, air, pollution, emissions, atmospheric, sensor, environmental, concentration, detection]

- **Particulate Matter (PM2.5, PM10):** [particulate, aerosol, dust, smoke, particles, pm25, pm10, coarse, respirable];
- **Greenhouse Gases (CO₂, Methane):** [co2, methane, carbon, greenhouse, warming, climate, radiative, absorption, anthropogenic, flux];
- **Ozone:** [ozone, smog, photochemical, tropospheric, stratospheric, precursors, ultraviolet, depletion, formation, oxidant];
- **Nitrogen Oxides:** [nox, nitrogen, combustion, photochemical, nitric, dioxide, acid, traffic, industrial, catalytic];
- **Sulfur Dioxide:** [so2, sulfur, acid, industrial, fossil, smelting, scrubber, volcanic, coal, power];
- **Volatile Organic Compounds (VOCs):** [vocs, hydrocarbons, solvents, evaporative, chemical, benzene, toluene, xylene, formaldehyde, acetone].

3.2.2. Monitoring Water Pollution Using UAVs

Common keywords for the domain: [uav, monitoring, water, pollution, aquatic, environmental, sensor, detection, quality, contamination]

- **Algal Blooms:** [algae, eutrophication, cyanobacteria, chlorophyll, nutrients, phosphorus, nitrogen, hypoxia, toxins, phytoplankton];
- **Chemical Pollutants:** [pollutants, toxins, heavy, metals, pesticides, industrial, runoff, leaching, bioaccumulation, organic];
- **Turbidity and Sedimentation:** [turbidity, sediment, suspended, solids, clarity, erosion, particles, light, penetration, deposition];
- **Oil Spills:** [oil, petroleum, hydrocarbon, slick, dispersant, crude, tanker, spill, marine, coastal];
- **Thermal Pollution:** [thermal, temperature, cooling, discharge, power, plants, ecosystem, fish, stratification, biodiversity];
- **Microplastics:** [microplastics, plastic, debris, fragments, fibers, polymers, marine, freshwater, ingestion, accumulation].

3.2.3. Monitoring Surface Pollution Using UAVs

Common keywords for the domain: [uav, monitoring, surface, pollution, environmental, sensor, detection, imaging, remote, sensing]

- **Soil Contamination:** [soil, contamination, heavy, metals, pesticides, herbicides, toxicity, remediation, leaching, erosion];
- **Vegetation Stress:** [vegetation, stress, chlorophyll, biomass, spectral, drought, nutrient, disease, crop, forest];
- **Urban Heat Islands:** [urban, heat, thermal, temperature, urbanization, albedo, energy, concrete, asphalt, microclimate];
- **Waste Dumps:** [waste, dumping, landfill, litter, debris, illegal, hazardous, municipal, industrial, recycling];
- **Land Use Changes:** [landuse, deforestation, urbanization, agriculture, desertification, habitat, fragmentation, ecosystem, biodiversity, development];

- **Erosion and Sedimentation:** [erosion, sedimentation, topsoil, runoff, gully, rill, wind, water, coastal, desertification].

3.2.4. Causes of Environmental Pollution

Common keywords for the domain: [pollution, environmental, human, impact, emissions, waste, contamination, degradation, anthropogenic, sources]

- **Industrial Emissions:** [industrial, factory, manufacturing, smokestacks, effluents, chemical, production, processing, refineries, smelting];
- **Agricultural Activities:** [agriculture, pesticides, fertilizers, runoff, livestock, irrigation, monoculture, deforestation, agrochemicals, erosion];
- **Transportation and Vehicle Emissions:** [transportation, vehicles, exhaust, traffic, fossil, fuels, particulates, combustion, diesel, gasoline];
- **Urban Development:** [urban, development, construction, infrastructure, impervious, sprawl, housing, roads, demolition, landfills];
- **Waste Management and Disposal:** [waste, disposal, landfills, incineration, sewage, recycling, hazardous, electronic, plastic, leachate];
- **Energy Production:** [energy, power, plants, fossil, fuels, coal, oil, gas, nuclear, thermal].

3.2.5. Challenges in Environmental Monitoring Using UAVs

Common keywords for the domain: [uav, drone, challenges, monitoring, environmental, limitations, obstacles, constraints, issues, problems]

- **Flight Time Limitation:** [battery, endurance, power, consumption, range, duration, energy, efficiency, payload, weight];
- **Sensor Accuracy and Calibration:** [accuracy, calibration, precision, sensor, drift, error, correction, validation, measurement, reliability];
- **Weather and Environmental Conditions:** [weather, wind, rain, temperature, humidity, atmospheric, conditions, stability, interference, performance];
- **Regulatory Restrictions:** [regulations, laws, restrictions, permits, airspace, safety, privacy, certification, compliance, authorization];
- **Data Processing and Interpretation:** [data, processing, interpretation, analysis, algorithms, software, visualization, integration, modeling, classification];
- **Operational Challenges:** [logistics, deployment, maintenance, training, piloting, navigation, planning, coordination, communication, recovery].

3.3. Topic Modeling

Topic modeling with the usage of Top2vec produces multiple clusters of abstracts, which can then be combined into topic groups by matching them to the keywords defined above (Section 3.2). The Top2vec workflow involves the following steps (see Figure 7):

1. **Text Vectorization:** Top2Vec can use pre-trained models of vector text representation such as Doc2Vec (Le & Mikolov, 2014), BERT Sentence Transformer (Reimers & Gurevych, 2019) or universal sentence encoder (USE) (Cer et al., 2018) to convert the documents and words into vectors. Formally, for example, $D = \{d_1, d_2, \dots, d_n\}$ is the set of documents, and $W = \{w_1, w_2, \dots, w_m\}$ is the set of unique words in these documents. The Top2vec algorithm first creates a joint vector space for documents and words. In this space, each document d_i and each word w_j are represented by vectors v_{d_i} and v_{w_j} , respectively. These vectors are generated so that semantically similar documents and words are close to each other in the vector space. In the current implementation, USE has been used to obtain the vector representations.
2. **Dimensionality reduction:** After text vectorization, Top2Vec uses the UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) algorithm to reduce

the dimensionality of document vectors $\{v_{d_1}, v_{d_2}, \dots, v_{d_n}\}$. This simplifies the spatial structure of the data, which facilitates subsequent clustering and visualization.

3. Clustering: The final stage applies the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering algorithm (McInnes et al., 2017) to identify dense regions in the reduced document space. HDBSCAN functions by computing a hierarchical cluster tree based on density-connectivity, extracting clusters using a density threshold λ , and assigning points to clusters or marking them as noise. Each dense region represents a group of documents that are closely related to each other in terms of meaning (topical or semantic cluster). For each dense region determined in the previous step, Top2Vec computes the centroid of document vectors in the original vector space. This centroid serves as a topic vector of the semantic cluster c_k representing the average position of documents within a topic group. For each topic vector, the nearest words in the vector space of words are found, which become the keywords of that topic. The set of topic keywords T_k can be defined as follows:

$$T_k = \{ \text{dist}(v_{w_j}, c_k) \leq \varepsilon \}, \quad (1)$$

where $\text{dist}(\cdot, \cdot)$ is distance metric in vector space, and ε is a threshold value for selecting the closest words. These words help to interpret the content of the semantic cluster.

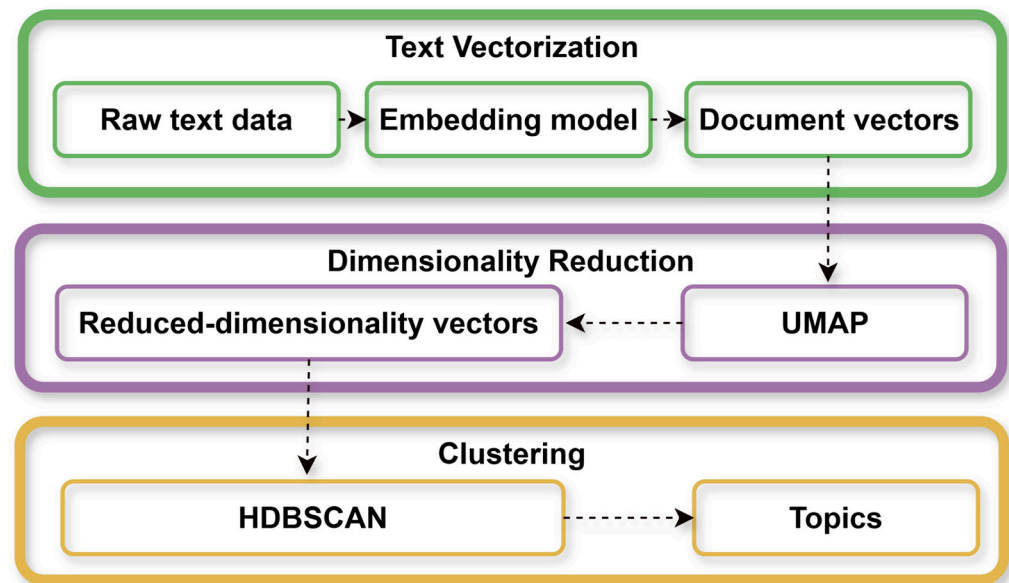


Figure 7. Schematic diagram of the Top2vec algorithm.

The parameters of the Top2vec model were chosen based on the preliminary experiments and recommendations in (McInnes, 2024; McInnes & Healy, 2024) and are given in Appendix B. The trained model is available on Hugging Face (CCRss, 2024b). As a result of the Top2vec application, 690 topics (semantic clusters) are obtained, from which those that are closest to topic groups are then selected.

3.4. Analysis of the Dynamics of Publication Activity Within the Topic Groups

At the final stage of the analysis, the most relevant semantic clusters (topics) were selected for each of the 30 topic groups defined in Section 3.3. For their selection, the cosine distance between the vector representation of the keyword combination of the topic group

and the vectors of the semantic clusters was calculated. The cosine similarity ($\cos(\theta)$) between vectors x and y of the same dimensionality n is calculated as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^n x_i y_i}{\left(\sqrt{\sum_{i=1}^n x_i^2}\right) \left(\sqrt{\sum_{i=1}^n y_i^2}\right)} \quad (2)$$

Using the obtained values for each topic group, 20 topics with the minimum cosine distance were selected. In essence, these topics form the content of the topic group.

To quantitatively evaluate how research interests in these topics change over time, we employed polynomial regression modeling. This approach was chosen because it allows us to perform the following:

- Capture non-linear trends in publication activity;
- Calculate the rate (D1) and acceleration (D2) of change in the number of publications through derivatives;
- Make predictions about future research trends.

A code snippet is given in Appendix C.

A polynomial regression model was formed for each group to numerically evaluate the dynamics of publication activity of the identified topic groups. The optimal regression order was determined as follows:

- The maximum degree of the polynomial in the search process was limited to the number 15.
- Data were split into 5 parts using the TimeSeriesSplit method to estimate models using cross-validation.
- Accuracy for all regression degrees from 1 to 15 was estimated as the mean of the mean square error (MSE) across all 5 parts.
- The value of the degree with the minimum obtained MSE was selected.

The obtained regression models reflecting the dynamics of publication activity within the topic groups were then evaluated by calculating D1 and D2 indicators, which correspond to the rate of change (D1) and acceleration of change (D2) in the number of publications (Gulf News, 2018). Table 1 shows how the combination of D1 and D2 values can be interpreted when assessing the publication activity.

Table 1. Interpretation of D1 and D2 indicators.

D1	D2	Interpretation of Indicators
"+"	"+"	accelerated growth
"+"	"−"	slowing growth
"−"	"−"	slowing decline
"−"	"+"	accelerated decline

4. Results

Quantitative indicators of the dynamics of publication activity in the topic groups are given in Appendix D. Figure 8 summarizes the quantitative indicators of each topic group for the end of 2023.

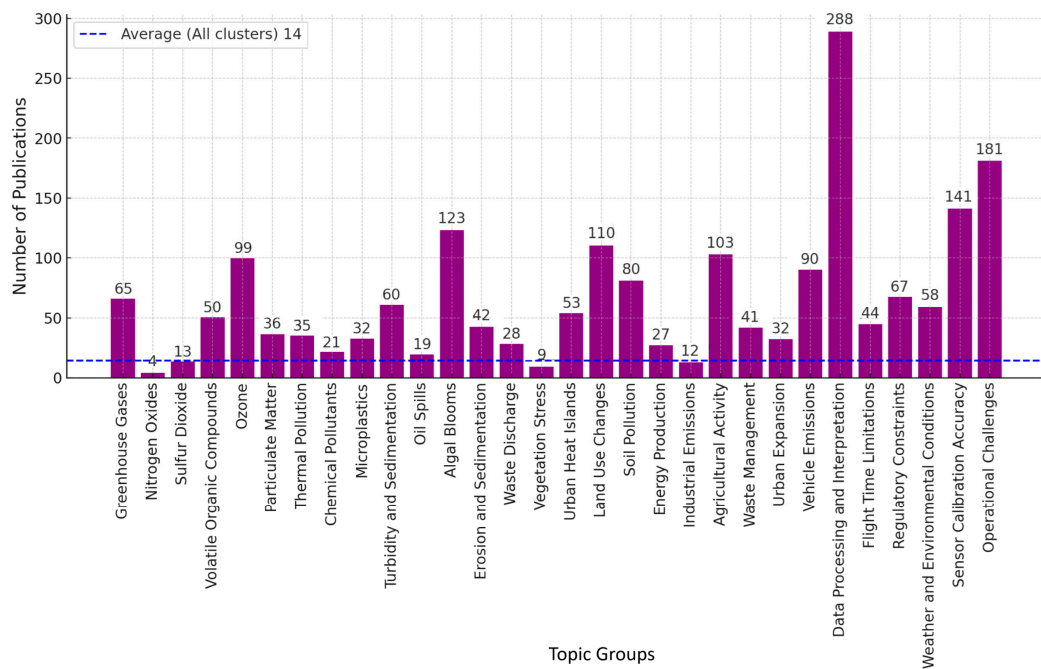
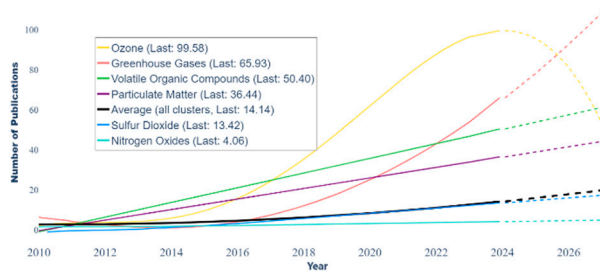
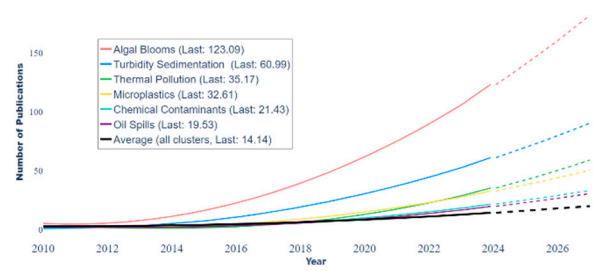


Figure 8. Number of publications relevant to the respective topic groups.

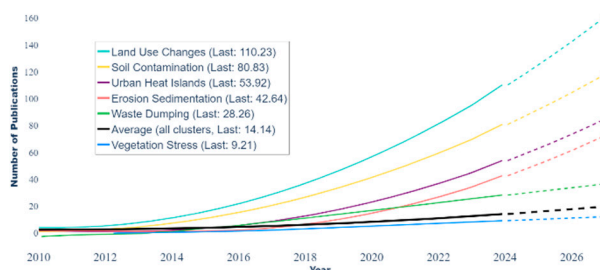
The results of regression models describing changes in publication activity in all five domains are shown in Figure 9.



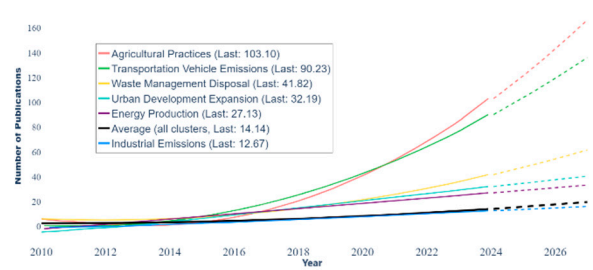
(1) Air pollution monitoring using UAVs



(2) Water pollution monitoring using UAVs

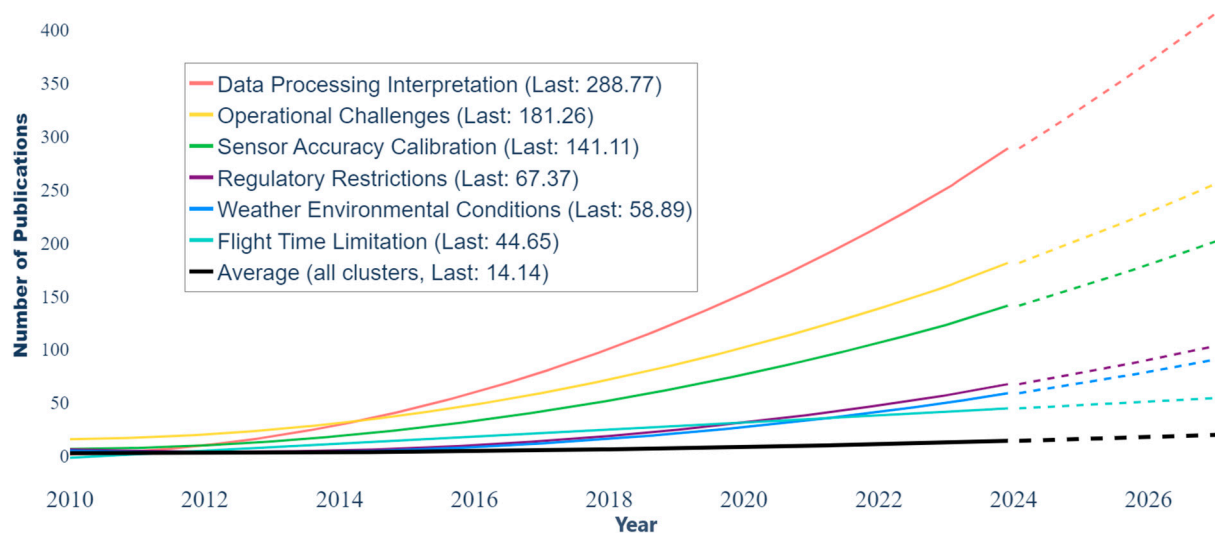


(3) Surface pollution Monitoring using UAVs



(4) Causes of environmental pollution

Figure 9. Cont.



(5) Challenges in environmental monitoring using UAVs

Figure 9. Models of changes in the number of publications in the field of environmental monitoring using UAVs for the period from 2010 to 2024. Last-estimated number of publications (according to the regression model) in the topic group for 2023.

In Figure 10, the fastest-growing topic groups with positive D1 and D2 values (marked with a star) are labeled. The topic groups marked with a muted color and a gray circle have a negative D2 value. The remaining groups marked with a green square have a D2 value close to 0.

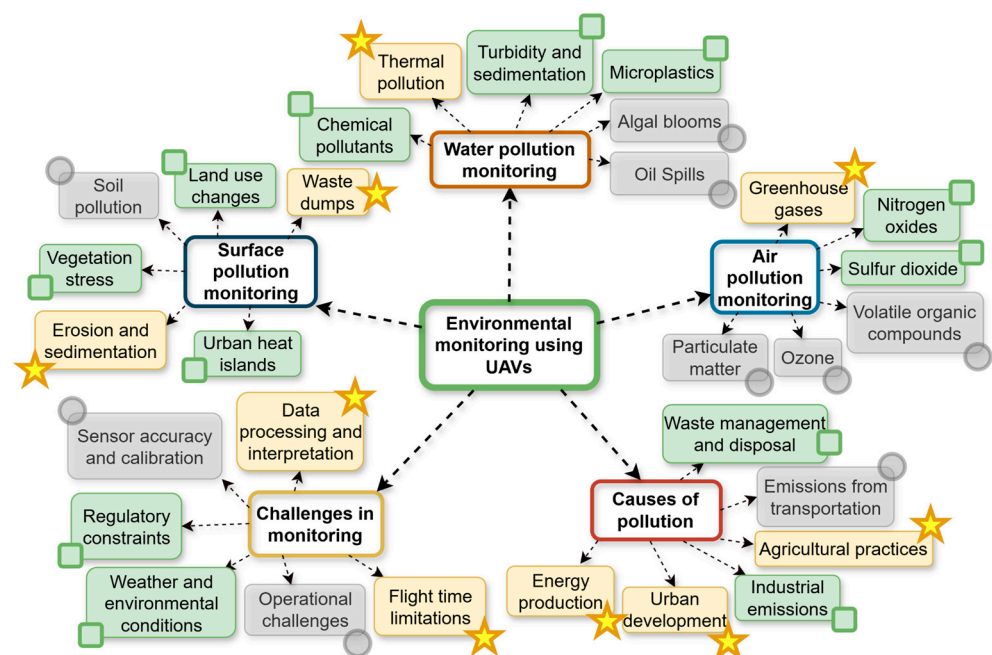


Figure 10. Classification of topic groups. The groups showing accelerated growth in the number of publications are marked with a star. A green square marks the groups demonstrating normal growth. The gray circle is the group showing a decrease in the growth of the number of publications.

5. Discussion

D1 and D2 indicators reflect significant differences in the dynamics of publication activity in different topic groups. If we consider the tasks of monitoring challenges research, the highest growth rate of the number of publications is observed in the field of data

processing and interpretation ($D1 = 3.016$, $D2 = 0.0448$) with a large (288) for 2023 number of publications (see table in Appendix D, the values are highlighted in bold). This topic group is the leading one both in terms of growth rate and absolute number of publications. This seems to be due to the fact that the technical elements of data collection systems are largely developed. At the same time, the processing of the obtained data, in particular, using machine learning methods is a promising scientific task. The second in terms of growth rate is “agricultural activities”, which is included in the “causes of pollution” domain. It can be assumed that this is due to the significant growth of research in the area of precision agriculture, where UAVs are considered as an important monitoring tool (supposedly 26% of the market of all UAVs is in agricultural models, or agrodrones (Insider Intelligence, 2020)). In the same domain, the topic groups “energy production” and “industrial emissions” are also receiving increased attention. This is also expected since the use of UAVs for monitoring pollution resulting from industrial plants and power plants is highly demanded (Asadzadeh et al., 2022; Jońca et al., 2022), as it can be operational, three-dimensional, and quite large-scale. In the “air pollution” domain, the topic group “greenhouse gases” is the most growing one. The decisions on the reduction in greenhouse gas (CO_2 and CH_4) emissions adopted by the UN (Boesch et al., 2021) initiated the research on methods for their monitoring, especially in the urban environment (Han et al., 2024). In the “water pollution” domain, the topic groups “thermal pollution”, “microplastics”, and “chemical pollutants” demonstrate high growth. In the domain “surface pollution” the groups “erosion and sedimentation” and “waste dumping” are the leaders in terms of the growth rate of the number of publications.

Based on $D1$ and $D2$ indicators, it is possible to rank the topic groups (see Figures 11 and 12). Based on the dynamic indicators, we can conclude that the topics with positive $D1$ and $D2$ values are the most promising for research. In other words, these topic groups will see an accelerated growth in the number of publications. On the other hand, in topic groups with negative $D1$ values, the number of publications will decrease. The topic groups “particulate matter” and “volatile organic compound” are good examples of these trends. The industry is already producing fixed and wearable sensors to monitor airborne particulate matter, organic compounds, and some common gases (sulfur dioxide). In other words, the problem has moved from research to the production technology area. As a result, the number of scientific publications devoted to this topic is decreasing.

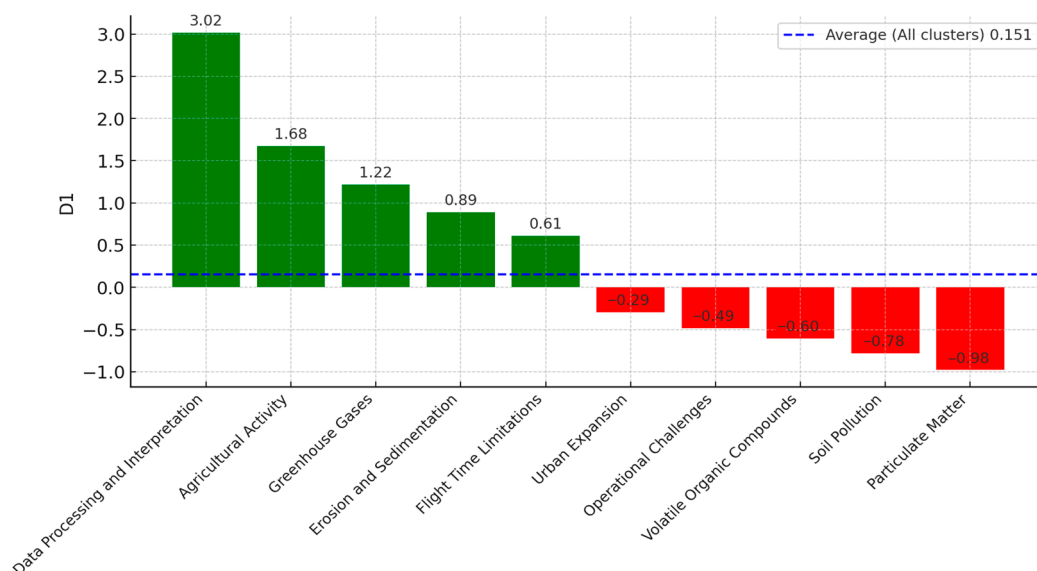


Figure 11. Top 5 positive (in green) and top 5 negative (in red) $D1$ indicators.

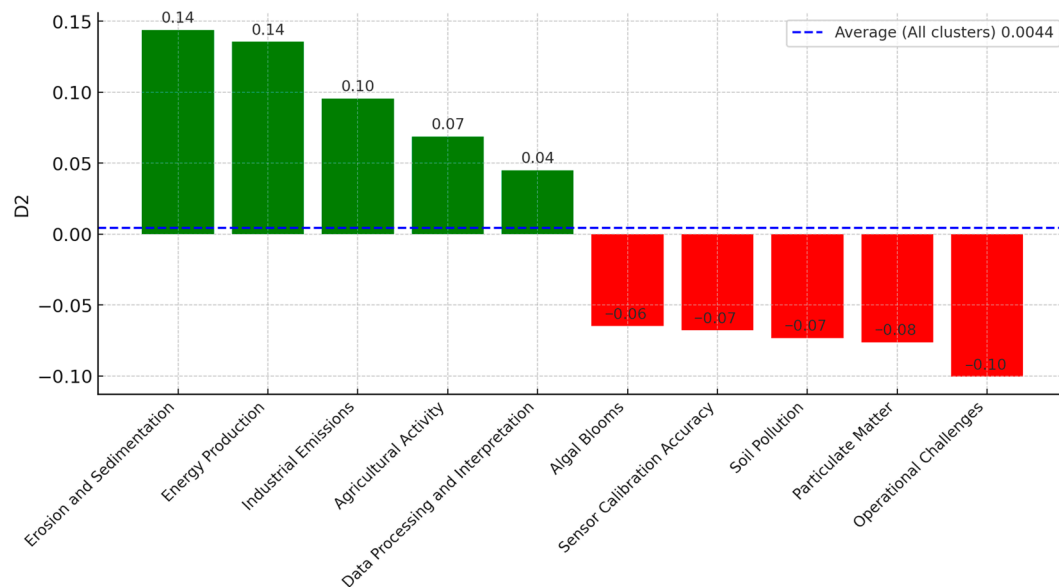


Figure 12. Top 5 positive (in green) and top 5 negative (in red) D2 indicators.

In general, it can be stated that the main challenge for researchers is the task of processing data obtained in the process of monitoring. Second in importance is the reduction in restrictions on the duration of UAV flight. The agricultural activities will be considered as a priority among the causes of pollution. Research in monitoring greenhouse gas emissions will be most needed in the air quality monitoring field, while erosion and sedimentation—in the area of land surface control. Thermal pollution, microplastic pollution, and chemical compounds will be most relevant in the area of water quality control.

Despite the understandable results, it is important to consider the limitations of the method:

1. The method is based only on the calculation of the dynamics of publication activity without taking into account the number of citations of articles, authors' characteristics, and their interrelation.
2. The original corpus of texts was limited to abstracts of articles obtained from the arXiv platform. While this provided a good basis for analysis, expanding the corpus to include data from other scientific journals and databases could have improved representativeness.
3. Using Top2vec for topic modeling leads to the following challenges:
 - a. Too many topics may arise, each requiring detailed examination and analysis.
 - b. The algorithm may generate outliers, in other words, topics with no clear semantic cohesion or relevance.
 - c. Each document in the model relates to only one topic, making it difficult to evaluate cross-disciplinary research.
 - d. There are no objective metrics to evaluate the quality of the resulting topics, making it difficult to verify the results.
4. There is a risk of subjectivity in the interpretation of topic groups, since the selection and description of themes may depend on the expert's opinion.
5. The study did not formally validate the topics using external data or comparisons with other studies, which limits confidence in their reliability and universality.
6. Top2vec was not compared with other topic modeling methods.

Despite the listed limitations, it should be noted that unlike systematic reviews of research areas (Butilă & Boboc, 2022; Shakhathreh et al., 2019), the proposed method allows analyzing a large number of publications that are inaccessible to manual analysis. Studying

the dynamics of publication activity can provide a clearer picture of the scientific community topics of interest. The method allows obtaining quantitative values of the rate of change in publication activity, as well as predicting their changes. In particular, our study made it possible to quantitatively substantiate the intuitively expected conclusion about the great interest of researchers in the tasks of processing data obtained using UAVs, and about interest in the issues of energy production, industrial emissions, greenhouse gas emissions, waste disposal, and environmental pollution in the process of agricultural production. At the same time, not-so-obvious rapidly growing domains of UAV application were also identified: thermal pollution, chemical contaminants, and microplastics.

6. Conclusions

This paper presents the numerical analysis of the dynamic performance of publication activity in the field of pollution monitoring with UAV employment. Using the Top2Vec topic modeling algorithm and regression analysis, there is analyzed the dynamics of publication activity in key areas including air, water, and surface pollution monitoring, causes of air, water, and surface pollution, and challenges associated with the use of UAVs in environmental monitoring. The study covered more than 556 000 abstracts of scientific articles published on the arXiv platform in the period from 2010 to 2023. The analysis revealed the following key trends:

1. Air pollution monitoring:
 - a. The greatest growth is seen in research devoted to greenhouse gases. Other areas, such as the monitoring of volatile organic compounds and particulate matter, show slower growth in publications.
2. Water pollution monitoring:
 - a. Research on thermal pollution continues to gain progress, showing a steady increase in publications. Topics related to microplastics and chemical pollutants are also evolving. Such areas as oil spills show declining interest.
3. Surface contamination monitoring:
 - a. Erosion and sedimentation are the fastest-growing areas among surface pollution studies. Land use change and soil contamination remain active, but the rate of growth of publications in these areas is declining.
4. Causes of environmental pollution:
 - a. Agricultural activities, energy production, and industrial emissions continue to attract the attention of researchers, showing a steady increase in the number of publications. Emissions from vehicles show a slight decrease in interest.
5. Challenges in environmental monitoring with UAV employment:
 - a. Data processing and interpretation and increasing flight time remain the most actively developing areas. Regulatory constraints and weather conditions remain relevant, but the pace of publications in these areas has slowed.

A distinctive feature of this work is that a quantitative assessment of the dynamics of publication activity has been carried out, which can be used to predict the number of publications on the relevant topic.

In order to overcome the limitations listed in Section 5 (Discussion), we plan in future studies to perform the following:

1. Expand the corpus of texts to increase the representativeness of the analysis by including publications from other scientific journals and databases, and also use full-text articles in addition to abstracts.

2. Apply various topic modeling methods to compare their performance, such as LDA, BERTopic, and others.
3. Perform topic validation to improve the reliability of the results using external data or expert judgment.
4. Explore the relationships between topics for a better comprehension of the topic structure of scientific publications. This will help to identify the potential interdisciplinary research areas.
5. Analyze the impact of new technologies on topic trends and research directions in environmental monitoring.

Supplementary Materials: The following supporting information can be downloaded at <https://www.dropbox.com/scl/fo/dcuu8p7vtw8yjdhpky3t/ALu0Xe6HUDa1gdTWZniocmg?rlkey=uizw6hs1vgdhv88vjqqis9lw3&dl=0>, (accessed on 18 January 2025).

Author Contributions: Conceptualization, R.I.M. and V.A.; methodology, R.I.M.; software, V.A.; validation, E.M., Y.P., and A.B.; investigation, A.B. and E.M.; resources, R.I.M. and V.A.; data curation, V.A. and A.B.; writing—original draft preparation, R.I.M., V.A., and Y.P.; writing—review and editing, Y.P., E.M., and A.B.; visualization, V.A. and E.M.; supervision, R.I.M.; project administration, Y.P.; funding acquisition, R.I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant BR21881908 “Complex of urban ecological support (CUES)”.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Additional Dataset Statistics

Table A1. Top 20 Categories.

Rank	Category	Count	Percentage (%)
1	cs.cv	84,656	15.24
2	cs.lg	74,650	13.44
3	cs.cl	41,216	7.42
4	cs.it	30,811	5.55
5	cs.ro	19,967	3.59
6	cs.cr	18,929	3.41
7	cs.ai	18,247	3.28
8	math.na	14,335	2.58
9	cs.ni	13,997	2.52
10	cs.ds	13,469	2.42
11	stat.ml	13,181	2.37
12	eess.sy	11,607	2.09
13	cs.dc	11,583	2.08
14	cs.se	11,066	1.99
15	eess.iv	10,186	1.83
16	cs.si	9331	1.68
17	cs.cy	8845	1.59
18	cs.lo	8705	1.57
19	cs.hc	8700	1.57
20	cs.ir	8583	1.54

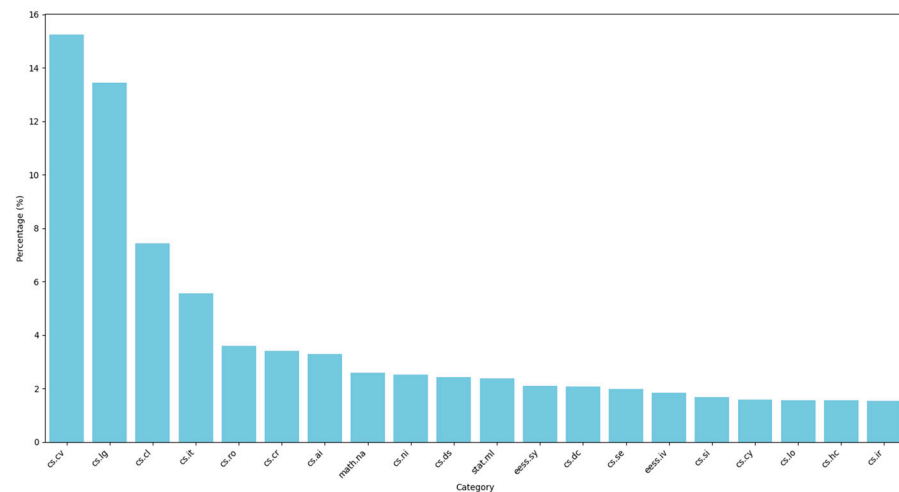


Figure A1. Distribution of arXiv CS Subcategories in the Dataset (2010–2024).

Appendix B. Tuning the Parameters of Top2vec Algorithm

The Top2vec algorithm has been customized using the following parameters:

1. **Learning speed:** In the course of experiments, it was decided to use the “learn” mode for the learning speed parameter, since no significant differences were observed between the “fast-learn”, “learn”, and “deep-learn” modes in the context of this study. The “learn” mode represents a balanced option that provides sufficient vector quality with acceptable learning time.
2. **Embedding model:** The “universal sentence encoder” model was used.
3. Additional parameters for dimensionality reduction (UMAP) and clustering (HDBSCAN) algorithms were configured as follows:
 1. UMAP:
 - a. `n_neighbors = 15`: Number of nearest neighbors used to evaluate local and global data structure.
 - b. `n_components = 5`: Number of dimensions in the reduced dimensionality space.
 - c. `Metric = 'cosine'`: Distance metric used to compute distances between points in the original space.
 - d. `min_dist = 0.0`: Minimum distance between points in the reduced dimension space.
 - e. `random_state = 42`: Value for reproducibility of results.
 2. HDBSCAN:
 - a. `min_cluster_size = 15`: Minimum cluster size, which defines the threshold value for cluster formation.
 - b. `Metric = 'cosine'`: A distance metric used to determine the proximity of points during clustering.
 - c. `cluster_selection_method = 'eom'`: The cluster selection method that defines the way to select a representative for each cluster.

Appendix C. Snippet of the Polynomial Model Code

```

...
def calc_polynomial_model(df, degree, x_range):
    # Prepare data
    x = df['year'].values.reshape(-1, 1)
    y = df['counts'].values

    # Create polynomial features
    poly = PolynomialFeatures(degree)
    x_poly = poly.fit_transform(x)

    # Fit linear regression on polynomial features
    model = LinearRegression().fit(x_poly, y)
    y_fit = model.predict(x_poly)

    # Calculate derivatives using numpy's polyfit and polyder
    coefs = np.polyfit(x.flatten(), y_fit, degree)
    poly_der1 = np.polyder(coefs) # First derivative
    poly_der2 = np.polyder(poly_der1) # Second derivative

    return poly_der1, poly_der2, y_fit, model, poly
...

```

Appendix D. Indicators of Publication Activity and Number of Publications in 2023

Domain	Topic Group	D1	D2	Number of Publications
Mean value (all 690 clusters)		0.151	0.0044	14.14
Air pollution	Greenhouse gases	1.218	0.0397	65.93
	Nitrogen oxides	0.003	−0.0005	4.06
	Sulfur dioxide	−0.241	−0.0200	13.42
	Volatile organic compounds	−0.603	−0.0483	50.40
	Ozone	−0.013	−0.0547	99.58
	Particulate matter	−0.979	−0.0762	36.44
Water pollution	Thermal pollution	0.577	0.0047	35.17
	Chemical contaminants	0.299	0.0023	21.43
	Microplastics	0.372	0.0010	32.61
	Turbidity and sedimentation	0.497	−0.0034	60.99
	Oil spills	−0.110	−0.0045	19.53
	Algal blooms	−0.077	−0.0649	123.09

Domain	Topic Group	D1	D2	Number of Publications
Surface pollution	Erosion and sedimentation	0.887	0.1438	42.64
	Waste disposal	0.220	0.0277	28.28
	Vegetation stress	0.097	−0.0005	9.21
	Urban heat islands	0.312	−0.0181	53.92
	Land use change	0.407	−0.0209	110.23
	Soil contamination	−0.779	−0.0733	80.83
Causes of pollution	Energy production	0.312	0.1354	27.13
	Industrial emissions	0.236	0.0956	12.67
	Agricultural activities	1.675	0.0688	103.10
	Waste disposal	0.201	−0.0147	41.82
	Urban expansion	−0.294	−0.0260	32.19
	Vehicle emissions	0.151	−0.0379	90.23
Monitoring challenges	Data processing and interpretation	3.016	0.0448	288.77
	Flight time restrictions	0.608	0.0058	44.65
	Regulatory restrictions	0.541	−0.0137	67.37
	Weather and environmental conditions	0.307	−0.0230	58.89
	Sensor calibration accuracy	−0.161	−0.0679	141.11
	Operational difficulties	−0.486	−0.1006	181.26

Background colours in the table correspond to the colours of the domains of the study area, presented in Figures 4 and 10. The extreme values mentioned in the text are highlighted in bold.

References

- Aljehani, M., & Inoue, M. (2019). Performance evaluation of multi-UAV system in post-disaster application: Validated by HITL simulator. *IEEE Access*, 7, 64386–64400. [\[CrossRef\]](#)
- Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv*, arXiv:2008.09470.
- Asadzadeh, S., de Oliveira, W. J., & de Souza Filho, C. R. (2022). UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. *Journal of Petroleum Science and Engineering*, 208, 109633. [\[CrossRef\]](#)
- Barbedo, J. G. A. (2019). A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses. *Drones*, 3(2), 40. [\[CrossRef\]](#)
- Bayomi, N., & Fernandez, J. E. (2023). Eyes in the sky: Drones applications in the built environment under climate change challenges. *Drones*, 7(10), 637. [\[CrossRef\]](#)
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Boesch, H., Liu, Y., Tamminen, J., Yang, D., Palmer, P. I., Lindqvist, H., Cai, Z., Che, K., Di Noia, A., Feng, L., Hakkarainen, J., Ialongo, I., Kalaitzi, N., Karppinen, T., Kivi, R., Kivimäki, E., Parker, R. J., Preval, S., Wang, J., . . . Chen, H. (2021). Monitoring greenhouse gases from space. *Remote Sensing*, 13(14), 2700. [\[CrossRef\]](#)
- Bretsko, D., Belyi, A., & Sobolevsky, S. (2023). Comparative Analysis of Community Detection and Transformer-Based Approaches for Topic Clustering of Scientific Papers. In *International conference on computational science and its applications* (pp. 648–660). Springer Nature Switzerland.
- Butilă, E. V., & Boboc, R. G. (2022). Urban traffic monitoring and analysis using unmanned aerial vehicles (UAVs): A systematic literature review. *Remote Sensing*, 14(3), 620. [\[CrossRef\]](#)

- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 257–289. [CrossRef]
- CCRss. (2024a). *ArXiv papers CS dataset*. Available online: https://huggingface.co/datasets/CCRss/arxiv_papers_cs (accessed on 11 September 2024).
- CCRss. (2024b). *Topic modeling Top2Vec scientific texts*. Available online: https://huggingface.co/CCRss/topic_modeling_top2vec_scientific-texts (accessed on 11 September 2024).
- Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, C., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strophe, B., & Kurzweil, R. (2018, October 31–November 4). *Universal sentence encoder for English*. 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 169–174), Brussels, Belgium.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June 2–7). *Bert: Pre-training of deep bidirectional transformers for language understanding*. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186), Minneapolis, MN, USA.
- Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7, 886498. [CrossRef]
- Erdelj, M., & Natalizio, E. (2016, February 15–18). *UAV-assisted disaster management: Applications and open issues*. 2016 International Conference on Computing, Networking and Communications (ICNC) (pp. 1–5), Kauai, HI, USA.
- Erdelj, M., Król, M., & Natalizio, E. (2017). Wireless sensor networks and multi-UAV systems for natural disaster management. *Computer Networks*, 124, 72–86. [CrossRef]
- Erkec, T. Y., & Hajiyev, C. (2022). Swarm architecture of UAVs. In *Progress in sustainable aviation* (pp. 15–36). Springer International Publishing.
- Fascista, A. (2022). Toward integrated large-scale environmental monitoring using WSN/UAV/Crowdsensing: A review of applications, signal processing, and future perspectives. *Sensors*, 22(5), 1824. [CrossRef] [PubMed]
- Gailler, L., Labazuy, P., Régis, E., Bontemps, M., Souriot, T., Bacques, G., & Carton, B. (2021). Validation of a new UAV magnetic prospecting tool for volcano monitoring and geohazard assessment. *Remote Sensing*, 13(5), 894. [CrossRef]
- García, Y. E., Villa-Pérez, M. E., Li, K., Tai, X. H., Trejo, L. A., Daza-Torres, M. L., Montesinos-López, J. C., & Nuño, M. (2024). Wildfires and social media discourse: Exploring mental health and emotional wellbeing through Twitter. *Frontiers in Public Health*, 12, 1349609.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*, arXiv:2203.05794.
- Gulf News. (2018, November 5). *Drone inspections help cut pollution by half* (Staff Report). UAE. Available online: <https://gulfnews.com/uae/environment/drone-inspections-help-cut-pollution-by-half-1.2263928> (accessed on 14 October 2024).
- Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of research trends using LDA based topic modeling. *Global Transitions Proceedings*, 3(1), 298–304.
- Ham, Y., Han, K. K., Lin, J. J., & Golparvar-Fard, M. (2016). Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): A review of related works. *Visualization in Engineering*, 4, 1. [CrossRef]
- Han, T., Xie, C., Yang, Y., Zhang, Y., Huang, Y., Liu, Y., Chen, K., Sun, H., Zhou, J., Liu, C., Guo, J., Wu, Z., & Li, S. M. (2024). Spatial mapping of greenhouse gases using a UAV monitoring platform over a megacity in China. *Science of The Total Environment*, 951, 175428. [CrossRef]
- Hervás, J., Barredo, J. I., Rosin, P. L., Pasuto, A., Mantovani, F., & Silvano, S. (2003). Monitoring landslides from optical remotely sensed imagery: The case history of Tessina landslide, Italy. *Geomorphology*, 54(1–2), 63–75.
- Hodgson, J. C., Baylis, S. M., Mott, R., Herrod, A., & Clarke, R. H. (2016). Precision wildlife monitoring using unmanned aerial vehicles. *Scientific Reports*, 6(1), 22574.
- Hu, J., Niu, H., Carrasco, J., Lennox, B., & Arvin, F. (2022). Fault-tolerant cooperative navigation of networked UAV swarms for forest fire monitoring. *Aerospace Science and Technology*, 123, 107494.
- Insider Intelligence. (2020, February 10). Commercial Unmanned Aerial Vehicle (UAV) Market Analysis—Industry trends, forecasts and companies. In *Business insider*. Available online: <https://www.businessinsider.com/commercial-uav-market-analysis> (accessed on 1 September 2024).
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169–15211.
- Jenssen, R., & Roverso, D. (2019). Intelligent monitoring and inspection of power line components powered by UAVs and deep learning. *IEEE Power and Energy Technology Systems Journal*, 6(1), 11–21.
- Jońca, J., Pawnuk, M., Bezyk, Y., Arsen, A., & Sówka, I. (2022). Drone-assisted monitoring of atmospheric pollution—A comprehensive review. *Sustainability*, 14(18), 11516. [CrossRef]
- Jordan, S., Moore, J., Hovet, S., Box, J., Perry, J., Kirsche, K., Lewis, D., & Tse, Z. T. H. (2018). State-of-the-art technologies for UAV inspections. *IET Radar, Sonar & Navigation*, 12(2), 151–164.

- Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Usmani, R. S. A., & Nayyar, A. (2020). Smart traffic monitoring system using unmanned aerial vehicles (UAVs). *Computer Communications*, 157, 434–443.
- Kherwa, P., & Bansal, P. (2020). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), e2.
- Lambey, V., & Prasad, A. D. (2021). A review on air quality measurement using an unmanned aerial vehicle. *Water, Air, & Soil Pollution*, 232, 109.
- Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196). PMLR.
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791.
- McInnes, L. (2024). *UMAP parameters documentation*. Available online: <https://umap-learn.readthedocs.io/en/latest/parameters.html> (accessed on 21 October 2024).
- McInnes, L., & Healy, J. (2024). *HDBSCAN parameter selection guide*. Available online: https://hdbscan.readthedocs.io/en/latest/parameter_selection.html (accessed on 14 October 2024).
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, arXiv:1802.03426.
- Medvedev, A., Telnova, N., Alekseenko, N., Koshkarev, A., Kuznetchenko, P., Asmaryan, S., & Narykov, A. (2020). UAV-derived data application for environmental monitoring of the coastal area of Lake Sevan, Armenia with a changing water level. *Remote Sensing*, 12(22), 3821. [CrossRef]
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, arXiv:1301.3781.
- Mogili, U. R., & Deepak, B. B. V. L. (2018). Review on application of drone systems in precision agriculture. *Procedia Computer Science*, 133, 502–509. [CrossRef]
- Mohamed, N., Al-Jaroodi, J., Jawhar, I., Idries, A., & Mohammed, F. (2020). Unmanned aerial vehicles applications in future smart cities. *Technological Forecasting and Social Change*, 153, 119293.
- Mohsan, S. A. H., Khan, M. A., Noor, F., Ullah, I., & Alsharif, M. H. (2022). Towards the unmanned aerial vehicles (UAVs): A comprehensive review. *Drones*, 6(6), 147. [CrossRef]
- Mohsan, S. A. H., Othman, N. Q. H., Li, Y., Alsharif, M. H., & Khan, M. A. (2023). Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. *Intelligent Service Robotics*, 16(1), 109–137. [CrossRef]
- Muhamedyev, R. I., Aliguliyev, R. M., Shokishalov, Z. M., & Mustakayev, R. R. (2018). New bibliometric indicators for prospectivity estimation of research fields. *Annals of Library and Information Studies*, 65(1), 62–69.
- Mukhamediev, R., Kuchin, Y., Yakunin, K., Symagulov, A., Ospanova, M., Assanov, I., & Yelis, M. (2020a). Intelligent unmanned aerial vehicle technology in urban environments. In *International conference on digital transformation and global society* (pp. 345–359). Springer International Publishing.
- Mukhamediev, R. I., Symagulov, A., Kuchin, Y., Yakunin, K., & Yelis, M. (2021). From classical machine learning to deep neural networks: A simplified scientometric review. *Applied Sciences*, 11(12), 5541. [CrossRef]
- Mukhamediev, R. I., Yakunin, K., Mussabayev, R., Buldybayev, T., Kuchin, Y., Murzakhmetov, S., & Yelis, M. (2020b). Classification of negative information on socially significant topics in mass media. *Symmetry*, 12(12), 1945. [CrossRef]
- Mukhamedyev, R. I., Kuchin, Y., Denis, K., Murzakhmetov, S., Symagulov, A., & Yakunin, K. (2019). Assessment of the dynamics of publication activity in the field of natural language processing and deep learning. In *International conference on digital transformation and global society* (pp. 744–753). Springer International Publishing.
- Muthusami, R., Mani Kandan, N., Saritha, K., Narenthiran, B., Nagaprasad, N., & Ramaswamy, K. (2024). Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports*, 14(1), 12003. [CrossRef] [PubMed]
- Park, S., & Choi, Y. (2020). Applications of unmanned aerial vehicles in mining from exploration to reclamation: A review. *Minerals*, 10(8), 663. [CrossRef]
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*, arXiv:1908.10084.
- Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. In *Text mining: Applications and theory* (pp. 1–20). John Wiley & Sons.
- Sadjadi, M. (2017). *ArXivScraper: A Python package for scraping arXiv.org*. Available online: <https://github.com/Mahdisadjadi/arxivscraper> (accessed on 27 October 2024).
- Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N. S., Khreishah, A., & Guizani, M. (2019). Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. *IEEE Access*, 7, 48572–48634. [CrossRef]
- Telli, K., Kraa, O., Himeur, Y., Ouamane, A., Boumehraz, M., Atalla, S., & Mansoor, W. (2023). A comprehensive review of recent research trends on unmanned aerial vehicles (uavs). *Systems*, 11(8), 400. [CrossRef]

- Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582. [[CrossRef](#)]
- Vorontsov, K., Frei, O., Apishev, M., Romov, P., & Dudarenko, M. (2015). Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *Analysis of images, social networks and texts: 4th international conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, revised selected papers 4* (pp. 370–381). Springer International Publishing.
- Yuan, S., Li, Y., Bao, F., Xu, H., Yang, Y., Yan, Q., Zhong, S., Yin, H., Xu, J., Huang, Z., & Lin, J. (2023). Marine environmental monitoring with unmanned vehicle platforms: Present applications and future prospects. *Science of The Total Environment*, 858, 159741. [[CrossRef](#)]
- Zengul, F., Bulut, A., Oner, N., Ahmed, A., Yadav, M., Gray, H. G., & Ozaydin, B. (2023, January 3–6). *A practical and empirical comparison of three topic modeling methods using a COVID-19 corpus: LSA, LDA, and Top2Vec*. 56th Hawaii International Conference on System Sciences, Maui, HI, USA.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.