Sentiment Analysis for the Natural Environment: A Systematic Review

MUHAMMAD OKKY IBROHIM, CRISTINA BOSCO, and VALERIO BASILE, Dipartimento di Informatica, Università degli Studi di Torino, Italy

In this systematic review, Kitchenham's framework is used to explore what tasks, techniques, and benchmarks for Sentiment Analysis have been developed for addressing topics about the natural environment. We comprehensively analyze seven dimensions including contribution, topical focus, data source and query, annotation, language, detail of the task, and technology/algorithm used. By showing how this research area has grown during the last few years, our investigation provides important findings about the results achieved and the challenges that need to be still addressed for making this technology actually helpful for stakeholders such as policymakers and governments.

CCS Concepts: • **Computing methodologies** → **Natural language processing;**

Additional Key Words and Phrases: Natural environment, data-driven policy, sentiment analysis, natural language processing (NLP), systematic review

ACM Reference format:
Muhammad Okky Ibrohim, Cristina Bosco, and Valerio Basile. 2023. Sentiment Analysis for the Natural Environment: A Systematic Review. *ACM Comput. Surv.* 56, 4, Article 88 (November 2023), 37 pages.
https://doi.org/10.1145/3604605

1 INTRODUCTION

The natural environment is one of the most important topics to be discussed in this era, and its issues need to be addressed by the population and governmental bodies with policies and regulations since it has an all-encompassing impact on all aspects of life. Its safety and health are crucial for the life of all beings that substantially depend on the natural environment. Environmental issues need to be discussed and solved together collectively, by the public as well as the governments and policymakers. Governments have an especially important role since they have the power to make and control the natural environment's policies and regulations. Nevertheless, they need to understand the main environmental issues in their country/region so they tackle them effectively, e.g., by means of data-driven policies. Exploring the relationship between humans and the environment through Sentiment Analysis may be an important step in this direction.

Sentiment Analysis (SA) has been widely applied in recent years on several broad topics like economy (e.g., SA on product reviews), healthcare (e.g., SA on patient feedback), government (e.g., SA on public facility reviews), and so on, to gauge how people feel about products, services, or other

Authors' address: M. O. Ibrohim, C. Bosco, and V. Basile, Dipartimento di Informatica, Università degli Studi di Torino, Turin, Italy; emails: {muhammadokky.ibrohim, cristina.bosco, valerio.basile}@unito.it.

0360-0300/2023/11-ART88
https://doi.org/10.1145/3604605

specific discourse topics. With respect to the environment, SA can help us understand the public perception of the state of the environment and government policies that impact the environment. This can allow governments to better fit citizens' expectations when they decide on priorities with regard to environmental issues in their country/region.

To explore SA applied to environment-related topics, some researchers have conducted reviews and surveys providing different perspectives according to the variety of addressed topics and analyzed text types. In [86], a review is conducted to explore the application of SA in the climate change debate. Meanwhile, Du et al. [23] explore the use of SA for analyzing opinions on several smart city environmental issues like climate change, urban policy, energy, and traffic. While Stede and Patz [86] explore papers that used various types of data sources (i.e., news articles, social media, etc.), Du et al. [23] explore only papers that analyze sentiment in social media. However, both [86] and [23] do not provide an in-depth exploration of the NLP techniques (from the creation of dataset to the evaluation of SA models) that researchers used applying SA on natural environment topics, since they only cover a few among the large variety of topics closely related to nature and environment, like food or carbon issues.

Through our novel systematic survey, we firstly want to complement the reviews provided in [86] and [23], in order to describe a more precise scenario about the exploitation of NLP techniques for addressing environmental topics. Moreover, we want to shed some light on the challenges that have to be addressed in future work, to improve the performance and the impact of SA techniques applied on natural environment topics.

For this systematic review, we, therefore, raise two main **research questions (RQs)** as follows:

(1) RQ1: What are the tasks, techniques, and benchmarks in SA on natural environment topics?
(2) RQ2: What are the challenges of SA in the natural environment domain compared to the general domain?

In conducting a systematic review to answer the research questions we raise, we utilize the Kitchenham framework [45]. In summary, the main contributions of this paper are as follows:

(1) To answer RQ1, we comprehensively analyze and summarize the findings of the following seven major dimensions on selected papers: contribution, topical focus, data source and query, annotation, language, detail of the task, and technology/algorithm used.
(2) To answer RQ2, we compare our findings in tasks, techniques, and benchmarks with other systematic reviews that discuss SA in the general domain [8, 51, 94].
(3) To fasten and make more systematic the selection process, we introduce an NLP framework implemented in Python for downloading and filtering papers that follows the Kitchenham framework. Making publicly available our Python implementation of this framework[1] we provide also a further contribution, not limited to the topic addressed in this paper, but rather useful to others that apply the Kitchenham strategy to systematic reviews.

More generally, a goal of this work is to provide structured and updated information on the one hand to the NLP scholar and practitioner interested in studying environmental topics with computational methods, and on the other hand to the social scientist, activist, and decision-maker interested in understanding and applying state-of-the-art language technology with focus on the natural environment. The rest of this systematic survey is organized as follows. Section 2 discusses the related works. Section 3 discusses the methodology we used, derived and adapted from [45]. Next, Section 4 proposes and explains the NLP framework, that follows Kitchenham's approach, to fasten the papers selection process and includes our result of paper selection for each stage. In

[1] https://github.com/okkyibrohim/kitchenscrap

Section 5, we comprehensively analyze the final selection of papers along seven dimensions. In Section 6, we wrap up the findings and discuss them to answer the research questions we raised. Finally, we give our conclusions and provide some future work suggestions in Section 7.

2 RELATED WORKS

SA has been one of the most popular NLP research fields in the last 20 years [100]. The goal of SA is to explore the human perception regarding a particular entity, such as an event, issue, service, product, public figure, or government, to name a few [52]. In recent years, several surveys surfaced on the topic of SA to explore what research has been done on this topic and what needs to be done in the future to obtain robust sentiment classifiers and other affective computational tools.

For the wide general term, [51] has conducted a survey for a tertiary study of SA in 2021. They collect and filter the papers using the Kitchenham framework to analyze three dimensions—tasks, approaches, and sentiment level classification in the selected papers. Still in 2021, Birjali et al. [8] conducted a survey that extends the results presented by [51] by further analyzing the application domains of SA, data collection, pre-processing, feature extraction, feature selection techniques, and the evaluation metrics used in the selected papers. In the following year, Wankhade et al. [94] proposed a study similar to [51] and [8], with the addition of a discussion of challenges and in particular the highlighting of structured SA, that is, the task of extracting additional knowledge about sentiment, such as the origin and target of the sentiment, or which span of text expresses the sentiment itself. Nonetheless, [94] found that structured SA is still not mature, with only a few papers discussing the research topic.

Following the trends in the broader NLP research community, in recent years, SA surveys focus on more specific topical focus. Among these, [102] explore the use of SA techniques for medical purposes, [13] examine previous works on SA in social media data for business intelligence, while [41] investigate what research has been done in SA for student feedback exploration. To the best of our knowledge, there is no deep and systematic review of the literature on SA on natural environment topics. As explained in Section 1, there are papers that discuss SA on natural environment topics [23, 86], but several areas are left unexplored, such as datasets building and models evaluation. Moreover, both [86] and [23] do not cover a broad spectrum of topics related to environmental issues, since they are mostly focused on the discussion around climate change and a few other topics (smart city, urban policy, energy and traffic).

In conducting this survey, we follow the Kitchenham systematic literature review framework for the paper selection process, how we define our query and how the selection stage is done, including the inclusion and exclusion criteria for each stage, similarly to [13, 51, 102]. For the dimensions that we analyzed in this survey, we are inspired by those proposed by [8, 51, 94], but we extend and adapt them to the research needs on the environment. For the final analysis, we focus on seven dimensions, namely contribution, topical focus, data source and query, annotation, language, detail of the task, and technology/algorithm used.

3 METHODOLOGY

In this section, we discuss how we follow the Kitchenham systematic literature review framework. This process includes various steps, namely the selection of database sources, the definition of the queries we applied on them, the paper selection stage itself, and the definition of the focuses of the analysis we want to provide for the selected papers.

3.1 Sources and Keywords

In this survey, we only consider computer science paper databases that we have the ability to access to download the full papers. We used six top computer science research publisher databases (ACL

Anthology,[2] ACM Digital Library,[3] IEEE Xplore,[4] MDPI,[5] Science Direct,[6] and Springer Open,[7]) one top computer science research pre-print database (Arxiv[8]) and one top computer science research indexer database (Scopus[9]).

As mentioned in Section 1, we conduct a systematic review that covers more natural environment topics than Stede and Patz [86] and Du et al. [23]. In the selection of the topics, we follow the *natural environment ontology* published by the **European Molecular Biology Laboratory (EMBL)**[10] and the list of sustainability research topics published by the **Worcester Polytechnic Institute (WPI)**[11]. For defining the query to be used to find the papers that may be relevant with respect to natural environment topics given by EMBL and WPI, we follow the **PICOC (Population, Intervention, Context, Outcome, Comparison)**. Before defining the PICOC query, firstly we did a small experiment to understand whether we need to add to our PICOC query some term variations. This is done by querying several databases using the topical keywords from EMBL and WPI, and performing a simple analysis of the search results, i.e., only analyzing the first page of results. For example, "waste" is actually not specifically listed in either EMBL or WPI. However, if we search using the "environment" term, we get "waste" among the results. Therefore, to be sure we retrieve also the papers about this relevant topic from all the selected sources, we added "waste" to our PICOC query. As another example, when we search "sustainable" as a variant of "sustainability", we get different results depending on the database. Therefore, we add both "sustainable" and "sustainability" to our PICOC query. If the results do not change by testing more variants of a term, only one variant is used for efficiency. In summary, based on this small experiment result, we defined our PICOC queries as follows:

- **Population**
  Our main general topic for this systematic review paper is SA. Therefore, we used *"Sentiment Analysis"* as our Population query.
- **Intervention**
  This systematic review is focused on the natural environment topic. We used queries that are related to the topic that follows from EMBL and WPI. We also add the variant terms resulting from the experiment described earlier. The final selection comprises *"Green, Nature, Environment, Chemical, Food, Plant, Organism, Climate Change, Sustainability, Sustainable, Carbon, Emission, Waste, Pollution"*, and *"Global Warming"* as our Intervention queries.
- **Context**
  Our systematic review paper is focused on exploring a series of NLP dimensions of the surveyed works, in line with previous work on the selected topic. Therefore, the Context queries used in this paper are queries that are relevant to NLP research. In this case, we used *"Corpora, Lexicon, Model, Algorithm"*, and *"Classifier"* as our Context queries.
- **Outcome**
  As stated in RQ1, the main goal of this systematic review paper is to explore the tasks, techniques, benchmarks, and challenges in SA on natural environment topics. This is done by analyzing seven NLP dimensions including contribution, topical focus, data source and

[2] https://aclanthology.org/
[3] https://dl.acm.org/
[4] https://ieeexplore.ieee.org/
[5] https://www.mdpi.com/
[6] https://www.sciencedirect.com/
[7] https://www.springeropen.com/
[8] https://arxiv.org/
[9] https://www.scopus.com/
[10] https://www.ebi.ac.uk/ols/ontologies/envo
[11] https://libguides.wpi.edu/c.php?g=355355&p=2396763

query, annotation, language, detail of the task, and technology/algorithm used. Therefore, there is no specific query defined for the Outcome part, rather it is already covered in Population, Intervention, and Context.

- **Comparison**
  As described in RQ2, we aim to compare the differences occurring in tasks, techniques, benchmarks, and challenges featuring the natural environment domain with respect to the general domain. However, instead of making two systematic review papers and comparing them, we will compare our findings in tasks, techniques, benchmarks, and challenges with another systematic review paper that discusses SA in the general domain. Therefore, there is no specific query defined for the Comparison part (already covered in Population, Intervention, and Context queries).

In implementing our PICOC queries, we used a boolean-based strategy to get the relevant papers where the Population, Intervention, and Context queries are joined using an *AND* operator, while the Intervention and Context parts are joined by *OR* operators. Based on these criteria, we construct our main boolean query as follows: *("Sentiment Analysis") AND ("Green" OR "Nature" OR "Environment" OR "Chemical" OR "Food" OR "Plant" OR "Organism" OR "Climate Change" OR "Sustainability" OR "Sustainable" OR "Carbon" OR "Emission" OR "Waste" OR "Pollution" OR "Global Warming") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier").*

Concerning the application of the query, we first focus on searching the papers' abstracts, in order to reduce false positives. However, not all sources provide means to effectively search terms inside the paper abstract (i.e., ACL Anthology, IEEE Xplore, Science Direct, and Springer Open). In those cases, we search the full papers following the default setting of the platforms, i.e., search from all metadata. Since each database source has a different advanced search facility, we adjust our main boolean query to each source accordingly, in three ways[12]:

(1) **Directly using the main boolean query**
For ACM Digital Library, IEEE Xplore, Springer Open, and Scopus, we can apply our main boolean query directly on their advanced search environment.

(2) **Splitting the Intervention part**
For Science Direct, there is a limit to the number of boolean operators that can be used in the advanced search box (up to eight operators). In this case, we split the Intervention part so that we have a combination of five queries:
- *("Sentiment Analysis") AND ("Green" OR "Nature" OR "Environment") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier")*
- *("Sentiment Analysis") AND ("Chemical" OR "Food" OR "Plant") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier")*
- *("Sentiment Analysis") AND ("Organism" OR "Climate Change" OR "Sustainability") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier")*
- *("Sentiment Analysis") AND ("Sustainable" OR "Carbon" OR "Emission") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier")*
- *("Sentiment Analysis") AND ("Waste" OR "Pollution" OR "Global Warming") AND ("Corpora" OR "Lexicon" OR "Model" OR "Algorithm" OR "Classifier")*

(3) **Generating the power-set for each combination part**
For the rest of the databases sources (ACL Anthology, Arxiv, and MDPI), we cannot combine AND and OR operators at once in their advanced search environment.[13] Therefore, we

[12]The complete result for the collected paper can be seen on this GitHub page: https://github.com/okkyibrohim/slr-sa-in-ne
[13]Actually, in the ACL Anthology we can use our main boolean query directly but they limit the results to up to 100 papers. We therefore apply the power-set modification to ACL in order to get a more complete result.

Table 1. Stages of Paper Selection

Stages	Inclusion Criteria	Exclusion Criteria
Stage 1: Initialization	–	• Not a research paper • Paper duplicated in a single source • More than ten years passed from publication
Stage 2: Title and Abstract Selection	Meets the relevant criteria	• Paper duplicated across different sources • Survey paper • Paper not satisfying relevance criteria
Stage 3: Full Text Selection	Meets the paper quality criteria	• Survey paper • Paper not satisfying the quality criteria

generate the power-set of Population, Intervention, and Context in order to create all their possible combinations in the form of simpler queries. Given that our Population part has one query, the Intervention part has 15 queries, and the Context part has five queries, this process results in a total of 75 queries.

3.2 Selection Stages

Following Kitchenham's guidelines, this survey implements the paper selection as a three steps process, as can be seen in Table 1.

The first stage of the paper selection process is the **initialization stage**. In this stage, we collected paper metadata from the eight selected databases, which include source name, paper link, year of publication, paper title, conference/journal name, author/s name, keywords and abstract. From several sources, we get some results that are not properly research papers, such as proceedings description/list, author's biography, tutorial session, and the like, and we removed them from our result list. Since we run multiple queries on several papers, (i.e., ACL Anthology, ArXiv, Science Direct, and MDPI), we may also retrieve duplicated results within the same source. In this case, we separately removed those duplicate results from each source. In this survey, we only consider papers published in the last ten years, therefore, papers published before 2012 are also removed in this stage.

Next, the second stage of the paper selection process is the **title and abstract selection**, that is, the assessment of the relevance to the papers based on their titles and abstracts. For this stage, we raised three questions that need to be answered by the title and/or abstract of the paper. The details of the questions including the inclusion and exclusion criteria can be seen in Table 2. In this stage, papers that fail one or more relevant criteria assessment questions are removed from our list.

Finally, the third stage of the paper selection process is **paper quality assessment** based on the full-text reading. For this assessment, we raised eight questions (four are *mandatory* and four *optional*) that were assessed using a binary scale (fulfilled or not). For each assessment question, we set the tolerance level differently based on the need for this systematic review. In this review, we emphasize that the paper must include a clear description of the topic and be relevant to our goals. For other assessment questions, we set high tolerance in order to be the most inclusive in our selection, also considering the limited number of papers about the topic we want to address. According to this tolerance strategy, the assessment questions are as follows:

(1) Did the paper clearly describe the research goals/problems? (*mandatory*)
(2) Did the paper show the related works? (*mandatory*)
(3) Did the paper explain the research methodology? (*mandatory*)

Table 2. Relevant Criteria Assessment

No	Questions	Inclusion Criteria	Exclusion Criteria
1	Is the paper discussing environmental topics? *(Mandatory, all inclusion criteria must be fulfilled)*	The paper discusses some environmental topic in accordance with the Intervention part and its similar objects	• The paper does not mention some of the terms in the **Intervention** part and its similar objects • The paper mentions some of the terms in the **Intervention** part and/or its similar objects, but it does not discuss environmental topics (e.g., food/restaurant review, prices/services of company energy review, etc.)
2	Is the paper proposing an NLP approach? *(Mandatory, all inclusion criteria must be fulfilled)*	The paper proposes some NLP approach, whether about dataset, shared task, or model building	• The paper provides a qualitative assessment • The paper is a survey • The paper provides basic text mining (e.g., top words analysis)
3	Is the paper discussing SA modeling? *(Mandatory, one or more inclusion criteria must be fulfilled)*	The paper discusses sentiment analysis in terms of • Corpus building • Corpus benchmarking • Lexicon building • Pre-trained language model building Tools/ • Library/pre-trained model application Classifier building (whether • a document, sentence, or aspect level)	The paper discusses topic modeling or other tasks without properly proposing/discussing SA

(4) Did the research results have relevance to the research goals/problems? (*mandatory*)
(5) Did the paper review the research background, literature review, and research context? (*optional*)
(6) Did the paper provide state-of-the-art results? (*optional*)
(7) Did the paper give future work recommendations? (*optional*)
(8) Did the paper come from a top conference/journal? (*optional*)

In conclusion, according to a quite tolerant strategy, papers that fail on one or more mandatory assessment questions will be excluded from our list. Meanwhile, for the optional assessment questions, a minimum of two of four questions must be fulfilled by the paper included in our survey. As in the second stage, if we still found a survey paper during the assessment in this final stage (because the title and abstract do not mention that it is a survey paper and subsequently it passed the Stage 2 selection process), we exclude it from our list.

3.3 Analysis of Selected Papers

In order to answer our research questions, we analyzed seven NLP dimensions in the papers that were selected according to the criteria described in the previous section. The dimensions that we analyzed in the final selection of papers are the following:

(1) **Contribution**
This dimension is about the type of contribution that the paper provides, i.e., whether its contribution is on corpus building, corpus benchmarking, lexicon/pre-trained model

building, analyzing sentiment on text using existing SA library/pre-trained model, or building a novel SA model.

(2) **Topical focus**
This dimension analyses what the paper focuses on, with respect to our Intervention part queries, i.e., whether the focus is on a specific topic (e.g., climate change, food problems, etc.), the paper combines several topic foci or covers all environmental topics in general.

(3) **Data source and query**
This dimension is about the dataset type and the query used for collecting the dataset described by the paper. For the dataset, we observed whether data come from microblogging/social media (e.g., Twitter, Facebook, Instagram, etc.), community forum discussion (e.g., Reddit), newspaper comments, or other sources. For the query, we observed instead how the queries used when crawling the dataset vary in different papers. For this dimension, we also analyzed how the authors filtered the dataset before labeling/using it for further analysis, when they scrap/collect the dataset, how big the dataset is, and also flagging whether the dataset is fully open for use by the research community or not.

(4) **Annotation**
This dimension is about the different aspects involved in the annotation: how the data were annotated, including what data were annotated (i.e., whether annotating their own scraped dataset or re-annotating an existing dataset), how many annotators are involved in the annotation process and how final labels were aggregated, who is/are the annotators (i.e., whether using experts or crowdsourcing), and how the annotation result has been evaluated (e.g., measuring inter-rater reliability). As far as the 'expert' definition goes, we follow that given by The Perspectivist Data Manifesto.[14] Experts can be the authors, domain experts, language teachers, native speakers, or simply someone who is trained to annotate the dataset by the authors. Meanwhile, crowdsourcing entails annotators hired through the Internet that are only given a guideline document without special training from the authors.

(5) **Language**
This dimension is related to the coverage of different languages in the datasets described in the selected papers, i.e., whether the datasets cover only one language (monolingual) or multiple languages (multilingual) and the characteristics of the datasets covering multiple languages (i.e., switch- or mixed-coding, or combination of datasets each including a single language). Moreover, we not only discuss the language in terms of the dataset, but also in the terms of approaches and models used i.e., what approaches and models are used, in handling monolingual and multilingual datasets.

(6) **Details of the task**
This dimension analyzes the task type described in the paper (e.g., only exploring SA or also combining it with topic modeling), how polarity is expressed (e.g., two or three standard polarity values, or another scheme of polarities), and how text is segmented for applying sentiment classification (i.e., whether the classification is referred to documents, paragraphs, sentence or to some more precise span within the text).

(7) **Technology/algorithm**
This dimension analyzes what technology or algorithm is used in the paper, e.g., existing SA classifier library/pre-trained model or some newly developed model. If the SA model is novel, we also analyzed the approach used (e.g., lexicon-based, classic **machine learning (ML)**-based, **deep learning (DL)**-based, or a combination of them), including algorithms and feature extraction, and also how the authors evaluated their model.

[14] https://pdai.info/

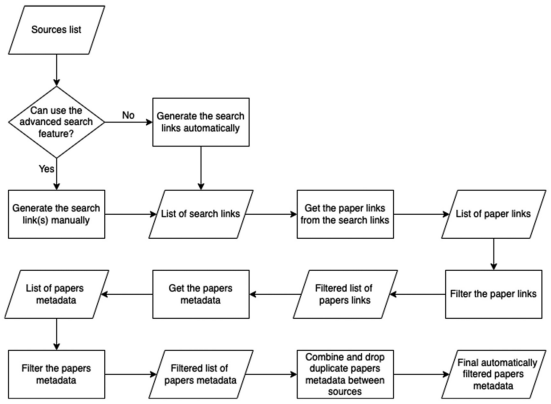

Fig. 1. The proposed framework flow.

4 PAPER SELECTION RESULT

As mentioned in Section 1, to fasten the paper selection process, we propose an NLP framework implemented in Python to automatically collect and filter the paper metadata following Kitchenham's framework. Our framework flow can be seen in Figure 1.

In the first step, we generate the search link(s) for each source that will be used as input in our paper links scraper. As explained in Section 3.1, for the sources that allow using the advanced search feature (ACM Digital Library, IEEE Xplore, Springer Open, Scopus, and Science Direct), we generate the search link manually, by entering our boolean query in their advanced search platform and setting some useful filter parameters (if applicable) such as publication year, paper type (research article from conferences or journals), and subject area (computer science). For some sources, an advanced search utility is either not present (ACL), or it is provided, but only in the form of a conjunction of search clauses (Arxiv and MDPI). In these cases, we build the search links by generating the power-set of our query parts (Population, Intervention, and Context) combination and putting them into the search link pattern on the source, finally saving the list of search links in a file.

After we get the list of search links for each source, the next step consists in collecting the results. In this case, we use Selenium[15] and Beautiful Soup[16] to scrap the paper links. Since the search result drawn from some sources contain links that are not research articles (e.g., front matters in the ACL Anthology), we need to remove them from our list. In this case, we find all links that are not research articles (e.g., proceedings description/list, author biographies, tutorial sessions) by following the link pattern and then removing it from our list.

The next step, after we get the list of paper links, consists in collecting the paper metadata from each paper link using the libraries Selenium, Beautiful Soup, and MetadataParser.[17] Nevertheless, the result of this scraping process shows that, for some articles from some conference proceedings or journals, some specific metadata is not provided. For example, typical articles in the ACL Anthology do not provide keywords, since, in this context, the papers do not contain them. For such papers, we set "no_keywords" on the paper metadata list. Moreover, for some papers, our Python script failed to get some other metadata not provided by the sources. For example, paper links from ACL Anthology do not provide abstract metadata on the web page on which we applied our

[15] https://selenium-python.readthedocs.io/
[16] https://beautiful-soup-4.readthedocs.io/en/latest/
[17] https://pypi.org/project/metadata-parser/

search procedure, even if the PDF files of the papers contain an abstract. For this case, we extract the abstract directly from the PDF file using the PyMuPDF[18] library. The paper metadata is then saved as a different file for each source. In this research, our Python script still failed to extract particular metadata for a few paper links due to they provide this kind of information using different special page structures. For this case, we manually complete the collection of metadata by opening the PDF file and then manually copying the particular metadata that failed to be scraped into the paper metadata list file.

The next step of our search is focused on the abstract of the papers. For each source paper metadata file, we apply a filter on the abstract to reduce the false positives in our results. In this case, we assume that the relevant papers should contain in their abstract at least one of our Intervention part terms and that all the papers not meeting this feature must be removed from our list. The filtered paper metadata is then saved as a new file.

Finally, we combined all metadata from all sources. However, since there are some papers duplicated between the eight selected sources that must be removed, we defined a procedure for correctly merging and deleting the paper metadata. First of all, since ArXiv is a preprint publisher, we put this resource as the last in the order (8th). As the second last source, we put Scopus, which is an indexer (7th). Lastly, noting how several NLP papers that have already been accepted or even published are sometimes presented in ACL workshops co-located with the main conferences, we decided to put the ACL source in the third last order (6th), so that the newest version of the paper takes precedence when a presentation at some workshops is the cause of duplicates. For the rest of the publishers, the actual order is not relevant. In the end, the final merged order is as follows: (1) ACM, (2) IEEE Xplore, (3) MDPI, (4) Science Direct, (5) SpringerOpen, (6) ACL Anthology, (7) Scopus, (8) ArXiv. After all paper metadata are merged, we run a Python script to drop duplicated papers based on their title and then save the results and get 1,435 papers that are ready for the next paper selection stage.

The next paper selection stage is focused on the title and abstract and exploits the relevant criteria for assessment explained in Table 2. During this process, we still find several duplicated papers between sources. This is because, in the automatic process, we just drop duplicated papers based on the paper title finding several papers which have different typing styles between sources (for the future, we suggest cleaning the title first before we drop duplicate the title). For example, there are some duplicated papers where the paper title contains a dash symbol ("-") in some sources but not in others. Therefore, during this stage, we manually remove the duplicated papers that have a different title writing style. From this step, we obtain 90 papers that are ready for the last paper selection stage.

Finally, the last paper selection stage is a paper quality assessment where we use the criteria that have been explained in Section 3.2. We manually downloaded all the PDF files of the 90 papers. Unfortunately, we cannot download some of the PDF files for the papers that come from Scopus. This is because Scopus is an indexer which also lists papers not published according to an open-access strategy that could not be accessed by us at the time of this writing. For these papers, we decide to remove them from our list, whose size decreased from 90 to 83 items. Out of these 83 papers, after the quality assessment stage, we selected 51 papers that are ready for the final analysis. Table 3 shows the number of selected papers organized in selection stages and source from which they were retrieved.

From Table 3, we can see that the framework we proposed for automatic paper scraping and filtering has been very useful to reduce a huge number of false positives and duplicate papers across different publishers and venues. This result shows that this implementation of the search

[18] https://pymupdf.readthedocs.io/

Table 3. Statistics of Selected Papers for Each Stage

Sources	Statistics for Each Stage							
	Stage 1 (Automatic Process)				Stage 2 (Manual Process)		Stage 3 (Manual Process)	
	Original Paper Links Scraped	Filtered Paper Links	Abstract Filtering	Drop Duplicated Papers Title	Drop Duplicate Papers Title	Relevant Assessment	Papers Download Process	Papers Quality Assessment
ACM Digital Library	71	71	71	71	68	5	5	3
IEEE Xplore	139	137	137	137	137	6	6	3
MDPI	73	73	28	28	28	8	8	6
Science Direct	2176	2176	316	316	316	15	15	8
Springer Open	615	608	127	127	127	2	2	2
ACLAnthology	1829	1110	114	106	106	5	5	4
Scopus	930	719	715	576	467	43	36	20
Arxiv	410	410	92	74	48	6	6	5
Total	**6243**	**5300**	**1600**	**1435**	**1297**	**90**	**83**	**51**

framework could be useful also for other researchers to fasten their paper selection process when conducting a systematic survey regardless of whatever the topic, for achieving reliable results. For the statistics of each stage, we performed a quick analysis of why there are so many false positives in our initial search results. The huge reduction rate we obtained for some of the sources (i.e., Science Direct, ACL Anthology, and Springer Open), depends on the fact that we searched first of all in the metadata associated with the papers. The false positive papers in this case mainly occur because they contain some of our Intervention parts terms, in particular metadata like the conference name, but, actually, the abstract does not contain any term from the Intervention part. In the second stage, we found that many false positives appear because we used several general queries that can have more than one meaning in different contexts, like *"Environment", "Nature", "Green"*, and *"Food"*. In our case, the term that generated the most false positives is "Environment" because this term is very general, see e.g., this term as in "computational environment", "social media environment" and so on. Meanwhile, in the last stage, we removed many papers from the list because their research goals are not relevant to our research goals followed by the paper description.

In our final selection of 51 papers, we found that the source that gives the largest total number of assessed as relevant papers is Scopus. For the year distribution, we found that research on SA on the natural environment topic can be found as early as 2012. In 2012, [34] conduct SA research on food contamination and food safety topics. After a few years' break, SA research on the natural environment topic started again in 2015. Increasing little by little every year, this research topic become trending in 2021 with a significant spike in the number of papers.

5 COMPARATIVE ANALYSIS ALONG THE MAIN DIMENSIONS

In this section, we deeply analyze the SA techniques that have been applied on the natural environment topics by the researchers that authored the 51 papers selected for this survey. These analyses are following the seven dimensions that have been explained in Section 3.3.

5.1 Contribution

In the 51 selected papers, we found several types of contributions. The preliminary analysis for this dimension is focused on the text source type of the datasets used in these papers, and shows that most of the datasets were newly scraped (46 papers) for the research. Among these 46 papers,

24 describe datasets that have been also annotated. However, only one of these datasets has been made available to the public [31], which is focused on organic food. From this finding, it is not surprising that almost all the selected papers in this final analysis involve the scraping of a new dataset for their research, since open datasets for this research topic are uncommon.

The next subdimension that we analyzed for what concerns contribution is what classification has been done by the authors, i.e., whether an SA model is built for the study, an existing library or pre-trained model is applied, or a dataset is created. From the analysis, we found that 27 of 51 papers only used some existing SA library/pre-trained model, 21 papers built an SA model for the purpose of their research, and 2 of them are building their SA model and then combining/comparing it with some existing library/pre-trained model. Meanwhile, the only paper where no sentiment classification is done is Gaspar et al. [28] where the authors only annotated sentiment on their dataset and provided a qualitative analysis of it. In this case, we still consider Gaspar et al. [28] as one of the selected papers since it includes an explanation of the annotation of the dataset.

From the 29 papers that used existing library/pre-trained models (27 papers that only use existing library/pre-trained models plus 2 papers that combine/compare existing models with a newly developed SA model), we found that the models represent several different approaches that, in general, can be categorized into lexicon-based, classic ML-based, DL-based, or a combination of them. Most existing libraries used in these papers are lexicon-based (23 papers purely use lexicon-based libraries only and 1 paper combines it with a pre-trained model). Finding that so many papers use a lexicon-based library is especially interesting because it is contrary to the current mainstream in NLP, where DL-based classifiers are usually employed for addressing SA tasks. It can be hypothesized that this is a consequence of the area being currently interdisciplinary, especially at the intersection with social sciences, and that the interest is focused on the topics addressed rather than the development of advanced NLP techniques or the achievement of state-of-the-art scores and results. This hypothesis is also supported by other findings we can draw from our survey. The first is the novelty of this research area. As discussed in Section 4, we found that this research topic started booming only in 2021. Finally, this hypothesis is also supported by the finding that of 27 papers that only used existing library/pre-trained models to analyze the sentiment of their dataset, no paper provides a valid evaluation of the classifier. Only in Michael and Utama [59], the authors provided an evaluation of the lexicon-based library they used, but the validity of this assessment is severely limited by the very small amount of data on which it is based (six annotated tweets).

Meanwhile, for the 23 papers that build an SA model (21 purely built a model and 2 papers combined/compared a model with some existing library/pre-trained model), most of them train the model on their scraped dataset (17 papers). Nonetheless, not all papers evaluate their model on a natural environment topic dataset. Among 23 papers that build SA models, 4 do not evaluate their model built on the dataset that is related to the topic (1 of them is evaluating the model on a non-topically-related dataset, while 3 of them are not evaluating their model at all).

Besides the lack of datasets and model contribution, we also found that there is no agreed-upon benchmark on this topic. Here by benchmark, we mean one or more datasets commonly used by the community to test their models and, most importantly, to compare their results in a controlled experimental setting, such as a shared task in an evaluation campaign. The only paper that proposed a dataset and a model, and opened them to the public is Hagerer et al. [31]. While the research results are not explicitly proposed as a benchmark, they can be employed as such, also because the dataset is already split into training, validation, and test set to allow reproducibility. For pre-trained language models, we did not find any paper where a novel model for classifying the sentiment on their dataset is proposed. All the selected papers reported in this survey, that build SA models using DL-based approaches, employ existing pre-trained language models. Meanwhile, of the selected papers that build sentiment lexicons for their SA model, actually we found 5 papers

([34, 88, 93, 98, 99]) that build a sentiment lexicon from their own scraped dataset. Unfortunately, none of them is publicly available.

5.2 Topical Focus

As far as the topic is concerned, from the analysis of all the 51 selected papers, we found that SA researchers actually have covered a broad spectrum of subtopics of the whole natural environment field. In this survey, we have mapped the topic and subtopic statistics provided in Table 4.

As we can see in Table 4, the topic that has been explored most in SA research is food (24 papers), followed by environment (15 papers), energy (12 papers), and waste (12 papers). In the selection phase prior to this analysis, we found several NLP papers that discuss climate change topics, but do not properly cite the "sentiment analysis" term. In some cases, SA is cited in the title and abstract but the technique actually applied in the study has more to do with stance classification, e.g., classifying whether a text (post/tweet) expresses a position on the existence or social relevance of climate change. In our survey, we do not include stance classification, which is related to, but different from SA. Especially for the stance on climate change, the correlation between sentiment polarity and stance about climate change itself is not clear: in a few observed instances, climate change believers post content with negative sentiment, e.g., towards specific policies or expressing climate-related anxiety.

The most discussed subtopic is food safety, as shown in Table 4. From our analysis, we found many papers that discuss food safety, and in particular harmful bacteria/elements outbreak cases. For example, Hsieh et al. [34] discuss the DEPH [di(2-ethylhexyl)phthalate] outbreak, while Chung et al. [14] discuss harmful elements produced by a food company outbreak. These findings indicate that we can use SA to detect and analyze outbreak cases in relation to food.

Based on Table 4, we can see that, in general, SA research discusses environmental topics both from the natural science perspective (according to EMBL) and the social science perspective (according to WPI). There are only a few subtopics that have not been addressed in the selected papers, such as food additives (from EMBL) and green technology (from WPI). On the other hand, there are many subtopics found in this survey that are not covered by either EMBL or WPI (e.g., water-related issues, or pollution). A number of subtopics also emerged that are not related to the physical environment. For example, Sluban et al. [83] and Michael and Utama [59] discuss management issues, while Zhang et al. [101] and Srivastava et al. [85] discuss policy issues, showing how the concerns with the health of the environment are intertwined with the scrutiny of governmental management and policies. We believe that these results can be employed to further extend and refine valuable resources such as the natural environment ontology.

5.3 Data Source and Query

For this dimension analysis, we explore several subdimensions regarding the data source and query used by the selected papers including media types, query types, and filtering techniques used in the papers. The statistics for each media type used can be seen in Table 5.

Table 5 shows that the most used data source media type for the experiments in the selected papers is social media. The most used social media is Twitter, not only because it is a highly popular platform, but also because Twitter provides a very open and sophisticated API to collect data.[19] In Dehler-Holland et al. [20], instead of classifying people's perceptions of sentiment like most other papers, SA techniques are employed to analyze news media framing regarding renewable energy policy in Germany. Another interesting example comes from Biehl et al. [7], where the authors

[19] https://developer.twitter.com/en/products/twitter-api. Meanwhile, interesting findings come from papers that have not used social media datasets.

Table 4. Topical Focus Distribution of Selected Papers

Topics	Subtopics	# Papers	Papers
Environment	Environment in general	4	[30, 36, 63, 74]
	Environmental conflict	1	[69]
	Air in general	1	[78]
	Water in general	2	[76, 78]
	Water quality	1	[93]
	Water crisis	1	[96]
	Urban planning	1	[69]
	Urban construction	1	[69]
	Underutilized land	1	[92]
	Livable places	1	[98]
	Urban park	1	[53]
Green	Green park	2	[53, 76]
	Urban green area	1	[76]
	Urban green spaces	1	[75]
	Green tourism	1	[80]
	Green hotel	1	[80]
	Green consumerism	1	[80]
	Green governance	1	[55]
Sustainability	Sustainable Development Goals (SDGs)	1	[74]
	Sustainable urban system	1	[63]
	Sustainable urban mobility	2	[7, 79]
	Sustainable transport	1	[9]
	Sustainable tourism	1	[9]
	Sustainable hotel	1	[80]
	Sustainable agriculture	1	[47]
	Sustainable food consumption	1	[9]
	Sustainable energy consumption	1	[9]
Food	Food quality	1	[93]
	Food safety	7	[14, 29, 30, 34, 56, 84, 99]
	Food contamination	3	[28, 29, 34]
	Food poisoning	2	[14, 29]
	Organic food	4	[9, 30, 31, 82]
	Gluten-free food	1	[71]
	Alternative meat	1	[11]
	Man-made meat	1	[11]
	Plant meat	1	[11]
	Plant cultivated meat	1	[11]
	Plant-based food	1	[90]
	GMO Food	1	[40]
Organism	Genetically Modified Organism (GMO)	2	[30, 40]
Climate Change	Climate change in general	6	[39, 43, 57, 62, 81, 88]
	Climate emergency	1	[77]
	Global warming	3	[44, 58, 77]
Carbon	Carbon in general	1	[83]
	Carbon taxation	1	[101]
Energy	Energy in general	1	[17]
	Fossil fuel	1	[58]
	Wind power energy	1	[19]
	Nuclear energy	1	[83]
	Renewable energy	5	[1, 19, 20, 76, 83]
	Green energy	1	[83]
	Sustainable energy	1	[83]
	Energy saving	1	[76]
Waste	Waste in general	2	[38, 78]
	Menstrual cup	1	[87]
	Plastic	2	[78, 85]
	Food waste	1	[9]
	Sewage	1	[78]
	Sanitation	1	[78]
	Waste collection	1	[76]
	Waste recycling	2	[76, 83]
	Waste management	2	[59, 83]
	Plastic ban policy	1	[85]
Pollution	Pollution in general	1	70
	Air pollution	3	[38, 76, 93]
	Emission	1	[83]
Total	**66**	**102**	

Table 5. Media Type Distribution Among the Selected Papers

Media Category	Media Types	# Papers	Papers
RSS, News Article, and BBS	RSS Feeds	1	[57]
	News Article	6	[19, 20, 30, 34, 39, 88]
	News Article Comment	2	[43, 99]
	Bulleting Board System (Not Specified)	1	[99]
Blog Forum	General Blog Forum and Review (Not Specified)	1	[69, 99]
	Reap Benefit	1	[78]
	IWasPoisoned	1	[29]
	Quora Comment	1	[31]
Social Media	General Social Media Not Specified)	1	[69]
	Twitter	31	[1, 14, 28, 43, 58, 71, 87, 90, 92, 96, 101], [36, 38, 44, 63, 75, 78, 79, 82, 83, 98], [9, 17, 40, 47, 55, 59, 74, 77, 81, 85]
	Instagram	3	[43, 55, 98]
	Weibo	4	[11, 56, 84, 93]
	TieBa	1	[93]
	Facebook	4	[17, 43, 76, 79]
Review Platform	Amazon Food Review	1	[29]
	Airbnb Review	1	[80]
	TripAdvisor Review	1	[79]
	Dianping Review	1	[53]
Other	Transcribed Interview	1	[7]
Total	**19**	**63**	

apply SA to analyze the sentiment of transcribed interviews. These findings show that NLP can be used to explore sentiment not only in user-generated content but also in other sources, such as transcribed interviews, combining quantitative and qualitative research to explore sentiment on natural environment topics.

In this survey, we did not find any paper using data in modalities other than textual, like images, audio, or video recordings. In Pilgun et al. [69], the authors actually also collect a video dataset, but they only analyze the text dataset. Moreover, there are papers that do not specify the blog, forum, or social media used for the data analysis process [69, 99]. This is because they generally scrap the data obtained from search engine results.

As far as the query type subdimension, in general, we can categorize the query types used by the selected papers into two classes, namely *direct query* and *query by the specific target*. They can be respectively defined as a query that is used to scrap the data generally based on some topic, e.g., classical keywords (words or phrases) and hashtags that are related to the main topic explored (direct query) and a query type that is used to scrap the dataset based on some specific entity to be analyzed, like a person (e.g., public figure), place (e.g., park), or organization (e.g., government, company, etc.) (specific target query). We found that, in general, papers only use a single query type for scraping the dataset, while only Chung et al. [14] use both query types. Chung et al. [14] analyze sentiment regarding food contamination by a food company, and they use keywords and hashtags related to the company and the Twitter account of the company itself. The detailed statistics of query types used by the selected papers are shown in Table 6.

In Table 6, we can see that the most commonly used query type are topic-related keywords. This is because most papers that we analyzed focus on SA applied to topics generically related to the natural environment, without considering some particular specific entity. On the other hand, some papers use specific target queries, and, in particular, they collect datasets more suitable for certain research goals. For example, in Hubert et al. [36], the authors collect a dataset on the

Table 6. Query Types used by the Selected Papers

Query Categories	Query Types	# Papers	Papers
Direct	Topic Related Keywords (Not Specified)	8	[1, 17, 38, 40, 43, 78, 93, 99]
	Topic Related Keywords	23	[19, 28, 31, 39, 71, 87, 90, 92], [11, 14, 56, 76, 82–84, 88], [9, 20, 34, 59, 77, 81, 85]
	Topic Related Hashtags (Not Specified)	1	[101]
	Topic Related Hashtags	7	[14, 44, 47, 55, 74, 81, 96]
Specific Target	All Post/Tweet and Comment/Reply Retweet from specific Geolocation	1	[62]
	All Reviews from a Website	1	[80]
	Specific Section/Category Review from a Website	2	[29, 79]
	Company Account Related	1	[14]
	Public Figure Account Related	1	[79]
	Political Party Account Related	1	[58]
	Government Account/Page Related	2	[36, 76]
	Specific Construction Project Related	1	[63]
	Specific Landmark Related	3	[53, 69, 75]
	Interview Respondents	1	[7]
Total	**14**	**53**	

natural environment from Twitter that mentions particular (listed) responsible ministry or secretary names/accounts. This allows authors to explore people's sentiment regarding the performance of those policymakers in handling natural environment issues. Another example comes from Nik Bakht et al. [63], where words/phrases that specifically relate to **light rail transit (LRT)** projects are used in order to explore the public perception of a particular project which allegedly may damage the surrounding environment.

Lastly, as far as the filtering techniques subdimension go, we found that many papers not only simply remove the duplicated data, but also apply some more specific filtering techniques to obtain a dataset more suitable for their analysis needs. In this survey, we divide the filtering techniques used by the selected papers into two categories, i.e., those applied to non-annotated datasets and those applied to annotated datasets. The statistics for each filtering technique used by the selected papers can be seen in Table 7.

From our analysis emerges that the main motivation for filtering non-annotated datasets is to prevent statistical bias. This is because the main spotlight of the selected papers is to analyze the statistics of sentiment polarity labels. However, if we look at Table 7, we can see that most papers that do not annotate their data are also not filtering their datasets at all, which seems contrary to the motivation that has been mentioned. One of the reasons for this apparent mismatch is that authors differentiate original posts, comments/replies, and reposts/retweets for their statistical analysis (e.g., [36]). Another reason is that the datasets are not extracted from social media, which lessens the need for filtering, e.g., to remove duplicates [39]. Meanwhile, for the annotated dataset papers, we see that the main motivation for filtering is to prevent dataset bias, since the presence of duplicated data in a dataset (whether used for training, validation, or testing) can decrease the model's performance and robustness. However, similarly to non-annotated dataset papers, there are several annotated dataset papers that do not filter their dataset. In these cases, the datasets are always very small (less than 1,200 instances), because they are collected from low resource languages [38, 44, 59], non-user generated texts (i.e., transcribed interviews, [7]), or because the topic discussed is very specific [87].

Table 7. Filtering Techniques used by the Selected Papers

Filtering Techniques	Non-annotated Dataset		Annotated Dataset	
	# Papers	Papers	# Papers	Papers
No filtering	13	[30, 36, 39, 58, 69, 71, 96], [34, 55, 57, 78, 79, 81]	6	[7, 38, 44, 59, 87]
Remove duplicate, repost, retweet	9	[19, 80, 82, 88, 92, 101], [9, 20, 40]	10	[11, 28, 53, 56, 63, 75], [17, 47, 74, 83]
Partial duplicate filtering	2	[19, 20]	1	[74]
Remove destroyed/ corrupted/uncompleted data	2	[77]	1	[99]
Remove non-relevant data	3	[40, 82, 88]	3	[31, 53, 63]
Remove spam/news/ commercial data	1	[82]	2	[53, 83]
Text mentioning other user	1	[92]	–	–
Text type filtering (verified types, editorial types)	2	[29, 88]	–	–
Account filtering (news channel, bot)	1	[92]	1	[28]
Text length filtering	1	[101]	1	[14]
Language filtering	3	[40, 77, 80]	1	[14]
Geolocation filtering	2	[1, 101]	1	[98]
Keywords/hashtag filtering	2	[62, 80]	2	[74, 83]
Category/topical focus filtering	2	[76, 77]	2	[31, 90]
Sampling by categories	–	–	1	[90]
Sampling by sentiment polarities	–	–	1	[85]
Random sampling	1	[62]	5	[11, 14, 17, 84, 99]
Total	**45**	–	**38**	–

As shown in Table 7, both non-annotated and annotated dataset papers apply similar filtering techniques. One of the main differences is that, in the case of non-annotated datasets, filtering is based on mentions done by other users (e.g., retweets) and text type (i.e., verified post, and text posted by editorial part), while any application of these forms of filtering is applied in papers which use annotated datasets. On the other hand, annotated dataset papers sometimes perform data sampling, likely to control the size and therefore the cost of manual annotation, while no non-annotated dataset papers do this (except for Moore et al. [62], where the dataset is sampled for the purpose of text classification). Finally, some authors sample the data by category [90] or by sentiment polarity [85], to reduce lexical/label distribution bias, and to obtain a more balanced dataset.

5.4 Annotation

As mentioned in Section 5.1, only 24 of the 51 selected papers annotate a dataset. In this section, we analyze three subdimensions for what concerns annotation, namely annotation aggregation techniques, annotator types, and annotation evaluation. The statistics for annotation aggregation techniques can be seen in Table 8.

Almost all papers in which some datasets are annotated the final labels were not aggregated, because each item of their datasets is annotated by a single annotator. This finding is interesting because it is contrary to the current mainstream in NLP, where usually the annotation process of datasets for SA is conducted by multiple annotators and an aggregation technique is applied after the annotation for deciding the final label to be used as the gold standard. This finding further

Table 8. Annotation Aggregation Techniques used by the Selected Papers

Aggregation Techniques	# Papers	Papers
No aggregation	20	[7, 11, 14, 17, 38, 43, 44, 47, 53, 56, 59, 74, 75, 83–85, 90, 93, 98, 99]
Hard majority voting	3	[28, 31, 63]
Weighted majority voting	1	[87]
Total	**24**	

strengthens the hypothesis previously formulated that most researchers that address as their research topic the natural environment come from the social sciences area and only seldom from computational linguistics.

Meanwhile, in 3 of the 4 papers where an annotation process involving multiple annotators is described [28, 31, 63] the final label is achieved using hard majority voting, while in the remaining one [87] weighted majority voting is used. In Gaspar et al. [28], each instance was annotated by two annotators, while in Hagerer et al. [31] each instance was annotated by 10 annotators. An interesting approach is proposed in Nik Bakht et al. [63], where a gamification strategy is applied, but they do not mention the annotator quorum and the average number of annotators for each instance. In [87], each instance was annotated by three experts consisting of an English professional teacher and the authors themselves, but the authors do not explain how they set the weight of each annotator in applying the weighted majority voting strategy.

As for the type of annotators, among the 24 papers where an annotated dataset is presented, we found that 3 exploited crowdsourcing, 16 used expert annotators, while the rest (5 papers) do not explain who exactly are the annotators [43, 44, 56, 59, 75]. Among the 3 papers that exploited crowdsourcing [31, 63, 90], the annotators' background (e.g., educational level, occupation, etc.) is not detailed. The authors of 14 of the 16 papers that used expert annotators, annotated the data by themselves [7, 11, 17, 28, 38, 47, 74, 83–85, 87, 93, 98, 99]; in 1 they also asked for help from their researchers' colleagues [53], and in 1 they trained their research assistants to perform this task [14]. Moreover, among the 14 papers where the authors describe an annotation process performed by themselves, some also cite the involvement of some external experts, such as a native speaker who is also a professional translator [28] or a professional English teacher [87].

Lastly, for what concerns the annotation evaluation, we only found 5 papers that evaluate the annotation process, all of them using Pairwise Cohen's Kappa score [15] as the evaluation metric. In [28], the authors use two fixed annotators to annotate the whole dataset and then evaluate the quality of the annotation with Cohen's Kappa. In [14, 53, 75], the entire dataset is distributed to be annotated among several annotators (more than two) separately. Then, they pick a sample subset of the data to be annotated together by all annotators, which is evaluated using the Cohen's Kappa. In Troya et al. [90], Cohen's Kappa is not only used to evaluate the annotation of a subset of data, but also for annotator selection. Here, only 10% of the dataset is selected to be annotated by three annotators, then Cohen's Kappa is computed, and finally the annotator with the highest average Cohen's Kappa was selected to annotate the whole dataset. It is interesting how even the works employing more than two annotators use Cohen's Kappa to evaluate the agreement, while in other fields Fleiss' Kappa [26] or Krippendorff's Alpha [46] are typically used in these cases, since these metrics can directly evaluate the annotation process for an arbitrary number of annotators.

5.5 Language

For this dimension, we explore what languages were studied in the selected papers. Similarly to the filtering technique (Section 5.3), we organize the analysis of language addressed by the papers according to the categories of non-annotated and annotated dataset. The statistics for the language addressed in the selected papers can be seen in Table 9.

Table 9. Languages Addressed in the Selected Papers

Languages	Non-annotated Dataset		Annotated Dataset	
	# Papers	Papers	# Papers	Papers
English	20	[1, 30, 39, 57, 58, 71, 80, 92, 96, 101], [9, 29, 40, 55, 62, 77, 79, 81, 82, 88]	11	[7, 14, 31, 63, 74, 75, 83, 85, 87, 90, 98]
German	4	[19, 20, 30, 69]	1	[28]
Spanish	3	[36, 69, 79]	–	–
Catalan	1	[79]	–	–
Russian	1	[69]	–	–
Italian	1	[76]	–	–
Turkish	–	–	1	[44]
Hindi	1	[78]	1	[47]
Chinese	1	[34]	6	[11, 53, 56, 84, 93, 99]
Korean	–	–	1	[43]
Indonesian	–	–	2	[38, 59]
Marrocan Arabic	–	–	1	[17]
Total	**32**	**–**	**24**	**–**

Table 10. Task Types in the Selected Papers

Task Types	# Papers	Papers
Building dataset and analyzing the sentiment manually	1	[28]
Analyzing the sentiment	34	[1, 11, 14, 19, 31, 36, 38, 43, 53, 57, 79, 87, 90, 92, 98, 99], [7, 17, 29, 34, 44, 47, 55, 59, 62, 74, 75, 77, 83–85, 93]
Analyzing the sentiment followed by analyzing the topic	1	[71]
Analyzing the topic followed by analyzing the sentiment	15	[9, 20, 30, 39, 40, 58, 63, 69, 76, 78, 80–82, 96, 101]
Total	**51**	

In the table, as expected, we can see that the mostly addressed language for both the non-annotated and annotated datasets is English. Meanwhile, most non-annotated datasets are from European languages, while most annotated ones include texts in Asian languages. This finding indicates that open SA tools (either in the form of end-user tools, libraries, or pre-trained models) and strategies in building corpora from other languages (e.g., with cross-lingual methods) have not been widely developed and used for Asian languages. On the other hand, the availability of more annotated datasets for Asian languages is promising for the development of NLP in the area. We also notice that there are only 3 papers [30, 69, 79] whose authors use multilingual datasets. However, all of them do not build and annotate their datasets and do not develop an SA model.

5.6 Detail of Task

In this dimension, we investigate three subdimensions considered by the authors of the selected papers in doing their research: task type, sentiment polarity, and sentiment classification levels.

For the detail of task types, the statistics can be seen in Table 10. According to the PICOC queries used, most selected papers are focused on discussing SA on natural environment topics, whether using an existing library/pre-trained model or training the model by themselves. Nevertheless, an interesting finding in this subdimension analysis is that many papers also explore some subtopics discussed in the dataset, as commonly happens in papers that do not build an SA model. In most of the cases, the authors used **Latent Dirichlet Allocation (LDA)** [9, 39, 40, 76, 80–82, 96], k-Means [30, 71, 101], or built topic classifiers by themselves [63, 78], while in a few cases topic-query

Table 11. Sentiment Polarities used by the Selected Papers

Sentiment Polarities	# Papers	Papers
Two polarities (positive, negative)	20	[1, 19, 30, 39, 57, 58, 79, 87, 99, 101], [9, 20, 29, 44, 55, 62, 77, 85, 93]
Three polarities (positive, neutral, negative)	20	[11, 14, 31, 38, 63, 69, 71, 92, 96, 98], [7, 17, 47, 56, 59, 74, 78, 81, 83, 88]
Extending classic sentiment polarities	2	[28, 90]
Emotion polarities	5	[36, 40, 80, 82, 84]
Multiple scenarios	4	[43, 53, 75, 76]
Total	**51**	

matching [58], **structural topic model (STM)** [20], or an instant end-user tool (i.e., just plug and play where the tools do not explain what method they used) [69] are used.

As far as the sentiment polarities used, the statistics are provided in Table 11. The table shows that the most used sentiment polarity classes in the selected papers are based on two labels (positive and negative) or three labels (positive, neutral, and negative). We found however 2 papers extending the classical sentiment classes so that the labels fit their research goals/dataset conditions better. For example, [90] adds the "conflict" class to indicate cases where the positive/neutral/negative polarity labels are all suitable, that is "conflict" was used due to the mixed polarity found in several sentences. We also found 5 papers that use emotion labels rather than sentiment polarity, which we include in this survey because we consider emotion analysis as a particular case of SA. 4 papers use multiple polarity scenarios, i.e., hierarchical emotion and sentiment classification. For example, [53] provides an emotion classification (anger, disgust, fear, sadness, happiness, surprise) in the first layer, and then a mapping of these emotion labels on three sentiment polarities (positive, neutral, negative). We found that in most of the 9 papers studying emotions the labels are based on the variation of Plutchik's model of basic emotions implemented in the NRC [61] emotion lexicon [36, 76, 82] or on Ekman's [24] six basic emotion labels [53, 75, 84]. In one paper, the annotation of emotions is based on Plutchik's [70] eight basic labels [80]. In another paper [40], LIWC [67] emotion labels are exploited, and finally in [43] the authors defined a custom set of labels.

For the sentiment classification level, we found that almost all the selected papers apply the analysis of sentiment on their datasets at the document level. We only found 4 papers that classified the sentiment at the sentence level [30, 31, 57, 90], and we did not find papers that analyzed the sentiment at the phrase or aspect level. However, different papers have different perspectives on the definition of what must be properly considered as 'aspect'. In some papers, the authors declare to perform an aspect-based sentiment classification. However, after the full-text analysis, we found that in these papers each sentence/document is in practice only mapped into one single topic/subtopic and sentiment label, therefore we consider them hierarchical topic modeling/classification followed by classification of the sentiment, rather than a direct aspect-level classification.

5.7 Technology/Algorithm

For the analysis of this last dimension, we investigated several subdimensions including approach, library/algorithm, feature extraction, and evaluation metric used by the selected papers. The statistics on the approaches used in the selected papers are provided in Table 12, while Figure 2 shows their distribution over time.

Among the papers that build their own SA model, we found that most of them exploit classic ML-based algorithms. This is an interesting finding since it is contrary to the current mainstream in NLP, where DL-based approaches currently are the ones most frequently applied. This finding

Table 12. Approaches used by the Selected Papers

Approaches	# Papers	Papers
No sentiment classification (only build sentiment dataset and analyze them manually)	1	[28]
Only using existing library/pre-trained model	27	[19, 30, 36, 39, 53, 57, 58, 69, 71, 76, 80, 92, 96, 101], [9, 20, 29, 40, 55, 59, 62, 75, 77–79, 82, 88]
Build lexicon-based model	2	[1, 34]
Build classic ML-based model	15	[7, 17, 38, 47, 56, 56, 63, 74, 83–85, 87, 93, 98]
Build DL-based model	3	[31, 81, 90]
Build and/or compare various approaches	3	[11, 14, 43, 99]
Total	**51**	

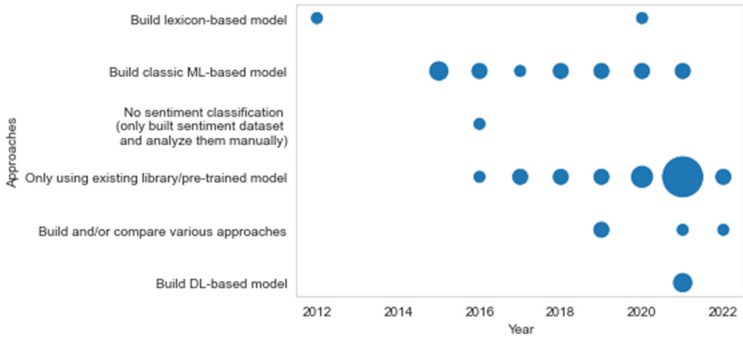

Fig. 2. Distribution of approaches used by selected papers over the years. Larger circles correspond to a higher number of papers using the approach in the same row. Note that papers may propose multiple approaches.

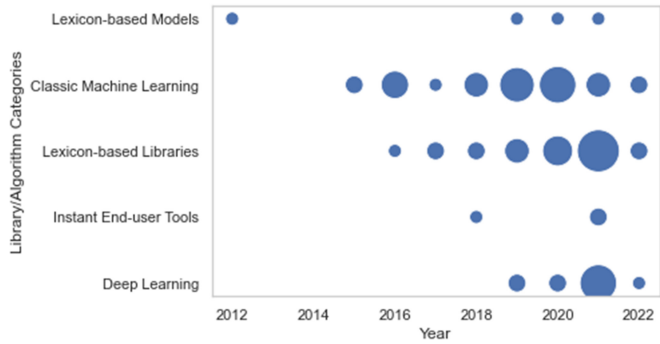

Fig. 3. Distribution of libraries/algorithms found in the selected papers over the years. Larger circles correspond to a higher number of papers using the approach in the same row. Note that papers may propose multiple libraries/algorithms.

indicates that the development of SA tools and models for natural environment topics is still quite lagging behind compared to other topical foci.

For what concerns the library/algorithm, we found that they are quite diverse in different papers. In general, from the algorithm basis, we can categorize them into five categories including instant end-user tools, lexicon-based libraries, lexicon-based models, classic machine learning, and deep learning. The library/algorithm trend over time can be seen in Figure 3, while the detailed statistics of each library/algorithm name is in Table 13.

Table 13. Libraries/Algorithms used by the Selected Papers

Library/Algorithm Categories	Libraries/Algorithms	# Papers	Papers
Instant End-user Tools	TextAnalyst[20]	1	[69]
	MonkeyLearn[21]	1	[74]
	Azzure Machine Learning[22]	1	[88]
Lexicon-based Libraries	VADER [37]	5	[62, 71, 78, 92, 96]
	TextBlob[23]	5	[1, 30, 59, 77, 101]
	NRC Emotion Lexicon [61]	6	[9, 36, 76, 80, 82, 98]
	Extended ANEW Lexicon [95]	2	[53, 75]
	GATE [16]	1	[58]
	SentiWS [72]	2	[19, 20]
	SentiWordNet [4, 25]	2	[39, 79]
	LIWC [66, 67]	3	[14, 40, 62]
	Harvard General Inquirer [42]	1	[29]
	AFINN [103]	1	[29]
	Other previous works [35]	1	[55]
Lexicon-based Models	Statistical score defined by the authors	3	[14, 34, 99]
	Fuzzy Dempster-Shafer	1	[1]
Classic Machine Learning	NB	7	[11, 17, 43, 44, 47, 63, 98]
	SVM	13	[11, 17, 43, 44, 47, 56, 63, 83–85, 93, 98, 99]
	kNN	6	[17, 38, 44, 47, 63, 87]
	CRF [48]	1	[85]
	Tree-based	6	[17, 47, 57, 63, 87, 98]
	Lasso	1	[7]
	Elastic-net	1	[7]
Deep Learning	Classic MLP/DNN	2	[47, 99]
	Basic Attention Model [5]	1	[31]
	LSTM [33]	2	[90, 99]
	Bi-LSTM	1	[43]
	GRU [12]	1	[90]
	CNN [50]	2	[43, 76]
	LTNet	1	[31]
	BERT [21]	3	[11, 90, 99]
	RoBERTa [54]	1	[81]
Total	**33**	**85**	

As can be seen in Figure 3, as also already mentioned before, most selected papers in this survey are using lexicon-based libraries and classic machine learning algorithms. From Table 13, it can be seen that the most used lexicon libraries are NRC Emotion Lexicon followed by VADER and TextBlob. Meanwhile, for classic machine learning algorithms, the most used algorithm is **Support Vector Machine (SVM)** followed by **Naive Bayes (NB)**, **k-Nearest Neighbors (kNN)**, and other tree-based algorithms. Of 6 papers that used tree-based algorithms, 3 use decision tree algorithms [17, 47, 63], while the rest use maximum entropy [57, 87] and J48 [98]. The authors of most of the papers that use deep learning algorithms selected one of the most known state-of-the-art deep learning models, i.e., BERT. The **Bidirectional Encoder Representations from Transformers (BERT)** architecture [21] is composed of a stack of transformer modules [91] using a bidirectional

[20] https://textanalyst.software.informer.com/2.3/
[21] https://monkeylearn.com/
[22] https://azure.microsoft.com/en-us/services/machine-learning/
[23] https://textblob.readthedocs.io/en/dev/api_reference.html#module-textblob.en.sentiments

Table 14. Distribution of Feature Extraction Techniques Found in the Selected Papers Over the Years

Feature Categories	Feature Name	# Papers	Papers
Hand Crafted	word *n*-grams	3	[44, 56, 63]
	$TF * IDF$	6	[11, 17, 38, 83, 87, 93]
	Lexicon Dictionary	2	[93, 98]
	Token Class Tagging	2	[63, 90]
	Text Element	2	[56, 63]
	User information and other metadata	1	[56]
Pre-trained Word Embedding	Word2Vec [60]	1	[99]
	GLoVE [68]	1	[31]
	BERT-based [21, 54, 89]	4	[11, 81, 90, 99]
Total	**9**	**23**	

Larger circles correspond to a higher number of papers using the approach in the same row. Note that papers may propose multiple feature extraction techniques.

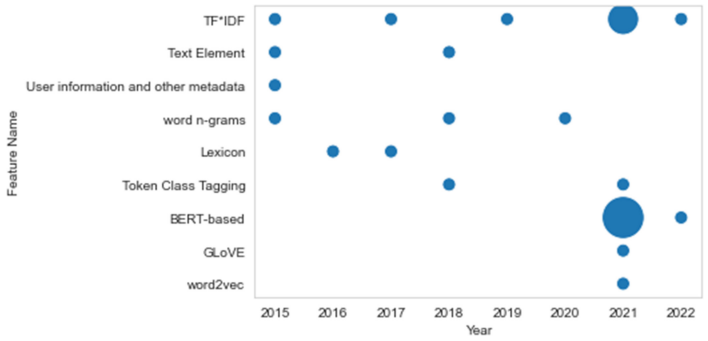

Fig. 4. Feature extraction techniques trend used by the selected papers.

approach so that in general BERT can learn the features of the text better than several classical deep learning approaches and solve several different NLP downstream tasks. In this survey, we found 4 papers that build their model using BERT, 3 using the original BERT architecture and the other one using the RoBERTa [54] architecture.

As far as feature extraction goes, we found only a few papers explaining the process used for building the SA model. In general, two categories, namely handcrafted and pre-trained word embedding features, are the techniques used in the selected papers. The statistics for feature extraction techniques used by the selected papers are provided in Table 14, while the trend distribution through the years can be seen in Figure 4. From Table 14, in particular it can be seen that where the feature extraction is explained, in most of the cases the hand-crafted feature used is Term Frequency × Inverse Document Frequency ($TF * IDF$). The next hand-crafted feature that is most used by the selected papers is the word *n*-grams. In this case, most of the papers use $n = 1$ (word unigram) [44, 56, 63]. While most of the selected papers are using hand-crafted features, from Figure 4, we found that the use of pre-trained word embedding features become a trend in the last three years. This finding shows a positive trend where this topic is following the state-of-the-art NLP trend in the feature extraction techniques side. However, for the pre-trained language model features, we found that there are no papers that build pre-trained language models, i.e., all of them are using existing pre-trained language models.

Lastly, for the metric evaluation subdimension, as mentioned in Section 5.1, not all selected papers in this survey provide an evaluation of the SA model. The statistics of the metrics used in the papers that provide some evaluation can be seen in Table 15. In most cases, the employed metric

Table 15. Evaluation Metrics used by the Selected Papers

Evaluation Metric	# Papers	Papers
Accuracy	13	[11, 14, 17, 31, 43, 47, 59, 63, 85, 87, 90, 93, 98]
True Positive Rate	1	[44]
False Positive Rate Rate	1	[44]
Precision	7	[44, 47, 56, 84, 85, 90, 99]
Recall	6	[44, 47, 81, 85, 90, 99]
F1-Score	9	[11, 14, 31, 44, 83, 85, 90, 98, 99]
Total	**37**	

is accuracy, which is also the only one applied in some cases. It is interesting to observe that many papers that have an imbalanced dataset are evaluating their model used only using accuracy, while this metric may not represent the performance correctly in this scenario, being biased towards the most frequent labels. This finding further strengthens our hypothesis that most researchers that are addressing this research topic come from the social sciences area rather than computational linguistics, where classification models are typically evaluated by precision and recall rather than, or in addition to, accuracy.

6 FINAL RESULTS AND DISCUSSIONS

In this section, we summarized what trends have been observed in this survey about the application of SA to the natural environment topics, based on seven major dimensions that we have analyzed in Section 5. After that, we discuss the challenges to be addressed for advancing this research field.

6.1 Tasks, Techniques, and Benchmarks for Sentiment Analysis on Natural Environment Topics

In this subsection, we not only discuss the trends of SA applied to the natural environment topics we observed during our survey, i.e., the tasks, techniques, and benchmarks that have been provided and discussed by previous works but also what has not been done until now about this topic compared to SA on general/other topics. For the tasks and techniques (including classifiers and features), we created a taxonomy to summarize what tasks and techniques have been done (colored green), partially done (colored yellow), and have not been done (colored red) by the previous works. This taxonomy is the result of a comparison of our findings from the previous works about the application of SA on the environment topics with what has been done applying SA on general topics [8, 51, 94].

6.1.1 Tasks. Based on the analysis of seven major dimensions (Section 5), we organized the tasks into a taxonomical structure, that, using colors cited above, helps to visualize at a glance what tasks have been addressed, only partially addressed and never addressed by researchers applying SA on natural environment topics. In general, the tasks in this taxonomy, shown in Figure 5, can be divided into *main task*, i.e., sentiment classification, and *support tasks*, that include filtering, topic/category detection, and implicit language detection. While the main task can be defined as the goal task of a research work, support tasks are all those activities, which are not strictly required, but are part of the processing pipeline, e.g., useful to obtain more valid and robust sentiment classification results.

According to the results of our survey, many subtasks have never been addressed. For instance, at the classification level, we found no works where sentiment classifiers work at the phrase level or for specific aspects, which is surprising considering the popularity of aspect-based SA applied

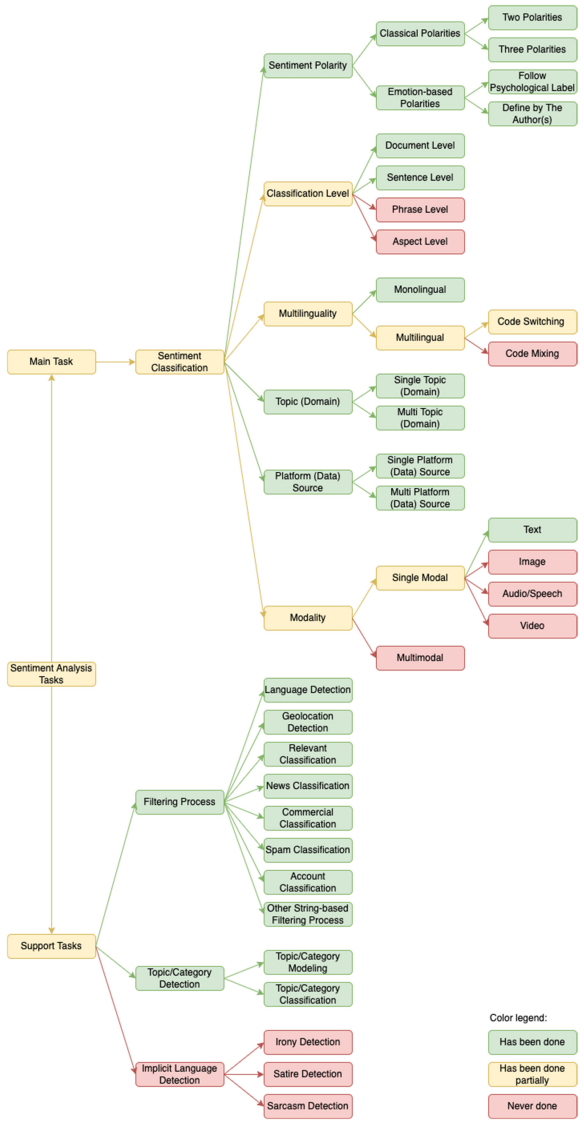

Fig. 5. Task taxonomy for SA on the natural environment.

to other topics. Moreover, aspect- or phrase-level SA may be particularly useful for support to policy-making on the environment. As far as multilingualism, we found that almost all selected papers are investigating SA in natural environment topics in a monolingual setting only. We only found 3 papers that include more than one language in their experiments, but all of them are simply combining datasets from several languages (code-switching) and use existing SA libraries to analyze the sentiment without evaluating the performance. Multilingual SA can be crucial to catching the opinion of people, since, especially in countries whose population use English as a second national language, people may use more than one language in a single message also. [8, 94]. Besides code-switching, where more than one language is used by the same population, there is also code-mixing, where languages are used in the same sentence. Both phenomena need to be

addressed in order to develop more effective SA techniques for these topics. Lastly, as far as the modality of data go, we found that all researchers that have investigated SA in natural environment topics are only using textual datasets. This is a further case of research on this specific topic lagging behind the NLP mainstream in terms of techniques and tasks.

For the support tasks, we found that there are no papers that discuss implicit language detection on the SA model they used or built, even though support tasks such as irony and sarcasm detection were found to improve the robustness of SA models [8, 51, 94]. In general, all the other support subtasks have been performed in the surveyed papers, but there are venues for improvement also under this respect. For example, on the account filtering through account classification, [28] applies a manual process for filtering news channel accounts, which cannot be considered robust nor scalable processing.

6.1.2 Techniques. As for the tasks, we also created a taxonomical structure for organizing and visualizing the techniques which have been developed, fully or partially, or never used by the researchers surveyed in this paper (using colors). Since all the selected papers in this survey only use text datasets for the experiments and analysis, the taxonomies only include techniques proposed in the literature for this kind of data. At the highest level, the techniques are subdivided into classifier, feature extraction, and evaluation metric. For the classifiers, the mind map can be seen in Figure 6.

Most of the classifiers used for SA in natural environment topics are based on lexicons and classic machine learning models. There is quite a variety of classic ML-based classifiers to be found among the surveyed works. Many classifiers that we found in this survey, employed as the main model, are often used as the baseline for SA in the general domain, e.g., **logistic regression (LR)**. At the same time, only a few papers employ DL-based classifiers, and, in particular, they almost exclusively employ BERT. Variants of BERT like ALBERT were not found [49], nor other transformer-based architectures like XLNet [97]. In this survey, we do not find works that use hybrid models, which are well known in the literature for their positive impact on classification [51].

We organized in taxonomy also the features used by the selected papers, taking into account two main categories, namely handcrafted and pre-trained word embedding features, as shown in Figure 7. Among the several handcrafted features that have not been explored by the previous works in SA applied on natural environment topics, we find character n-grams, emoji/emoticon lexicons and orthography-based features. Character n-grams have been used in the SA literature, especially on short-text datasets where many **out-of-vocabulary (OOV)** cases may occur because of informal writing styles, such as microblogs, since these features can extract sub-words also [32]. The same can be said for the investigation of emoji/emoticon lexicons, since these devices are often used to convey specific tones of emotion and sentiment [94]. Orthographic features, such as punctuation (e.g., exclamation marks), are used by social media users to highlight the sentiment of their message [94]. For weighting handcrafted features, term frequency and IDF-based schemes were widely used in the surveyed papers, matching the SA literature for SA in the general domain. However, we did not find papers in this survey that encode weights in a one-hot fashion, as done for instance by Birjali et al. [8]. Finally, we did not find any paper that applies some form of feature selection, even though this process is quite widespread in NLP methods using handcrafted features to improve the classification performance and efficiency by removing the irrelevant and redundant features [8].

A limited variety of pre-trained word embeddings has been used by the papers we surveyed. As non-contextual word embeddings, most works use Word2Vec [60] and GloVe [68]. However, several SA studies outside this survey employ other embedding models, such as FastText and Doc2Vec. FastText can be especially useful for addressing OOV words, often occurring in social media texts, because it accounts for sub-word level information. As for contextual pre-trained word

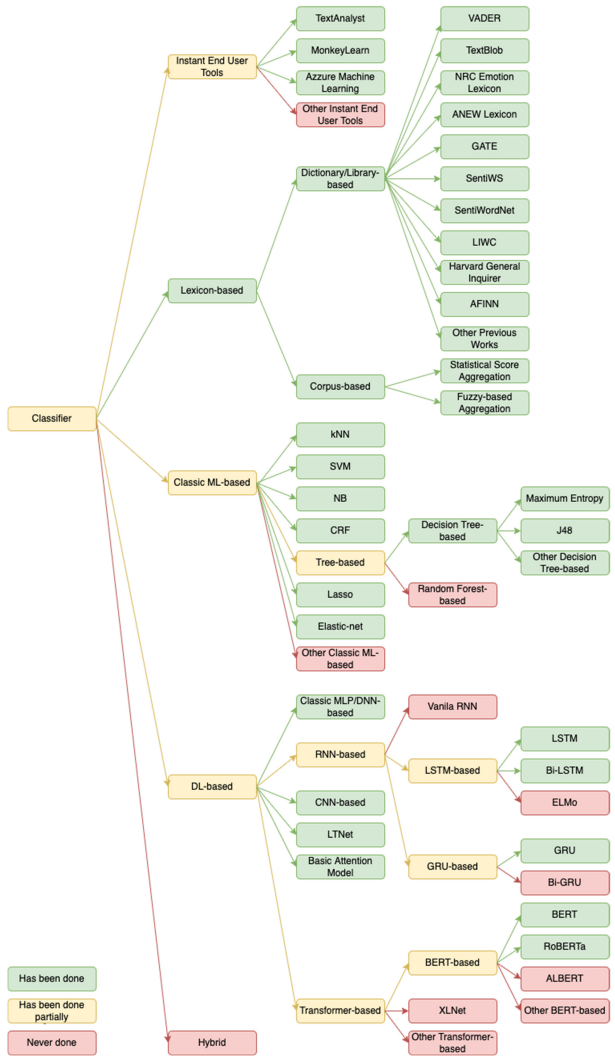

Fig. 6. Classifier taxonomy for SA on the natural environment.

embeddings, we only found papers that use BERT-based models to compute them, while other architectures of contextual word embedding, like the LSTM-based model namely **Embeddings from Language Model (ELMo)**, or transformer-based models, were not found in the surveyed papers.

6.1.3 Benchmarks. As discussed in Section 5.1, we do not find any model and dataset benchmark for SA in natural environment topics, while several mature benchmarks are currently available for SA in other domains. Publicly available benchmarks are crucial for advancing the state of the art in any NLP area, by fostering data circulation, reproducibility, and a quicker growth of the research, since the development focus can shift from data curation to model building. From our survey findings, we can also see that SA datasets open for the research community are very few for what concerns the natural environment topics. The only paper that provides a similar resource is Hagerer et al. [31] where a dataset on organic food topics is presented. Their dataset can be exploited as a benchmark since it is open to the public and associated with a clear explanation

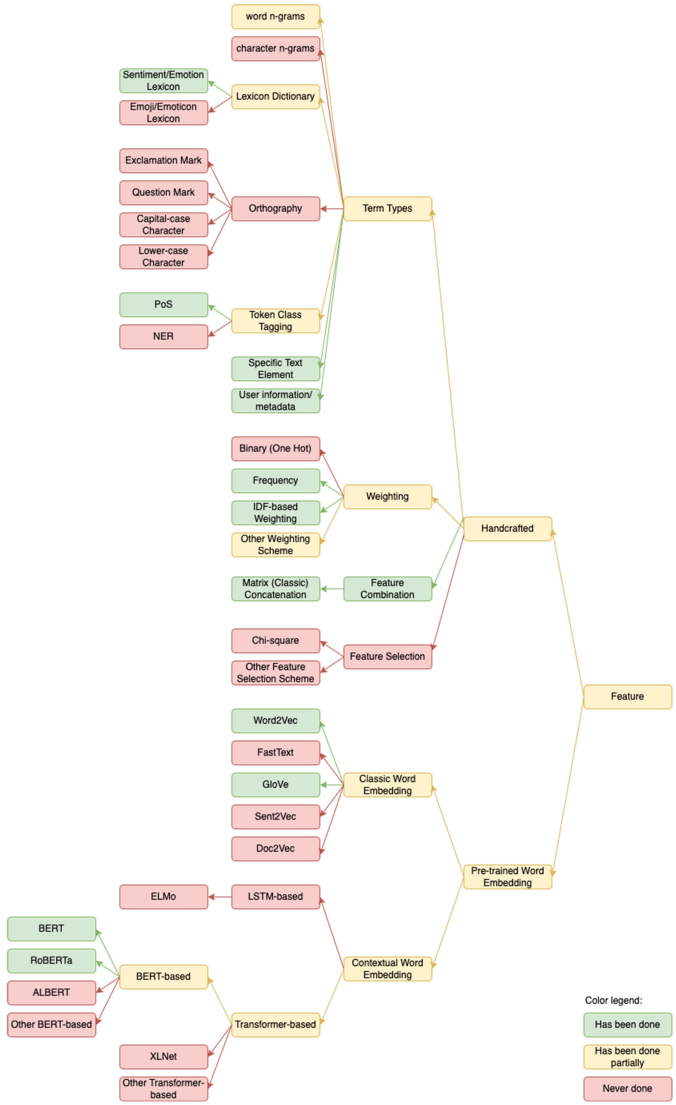

Fig. 7. Feature taxonomy for SA on the natural environment.

of how it has been annotated. Moreover, they also provide a train-validation-test split, to make comparison easier.

6.2 Challenges of Sentiment Analysis on Natural Environment Topics

In this section, we summarize the challenges that need to be tackled in future work, not only discussing them but also providing some hints about how SA methodologies applied for addressing general topics can be adapted to environmental topics. In general, we divide the challenges into three categories: classification level and structured SA, data availability, and methodology.

6.2.1 Classification Level and Structured Sentiment Analysis. As discussed in Section 6.1.1, we found that SA in natural environment research has been applied at the document and sentence

Fig. 8. An example of social media message containing multiple aspects with different sentiments.

level only, while phrase and aspect levels have not been explored. Structured SA is recently gaining traction in the NLP research community. While in unstructured SA the output of a model is typically a sentiment label (whether at the document, sentence, or phrase level) and perhaps the aspects in aspect-level SA, a finer-grained approach could extract information such as the opinion holder, the text span that specifically expresses the sentiment, and the target of the opinionated utterance. For the natural environment topics, structured SA could therefore be very useful, especially for the stakeholders and policymakers interested in extracting more precise information, e.g., about what is more criticized by the citizens.

The lack of previous work on structured SA applied to environmental topics makes the development of this kind of technique especially challenging. One of the challenges is to select the relevant aspects. For aspect-level SA in natural environment topics, taking inspiration from the results of this survey, good candidates for aspects may be ecological or managerial aspects. More in detail, four ecological aspects could be considered, namely the hydrosphere (i.e., water including groundwater, river, lake, sea, and ocean), atmosphere (i.e., gas), lithosphere (i.e., solid including surface ground and underground), and biosphere (i.e., organisms including human, plants, and animal life). On the other hand, the managerial aspects could include habits, non-governmental organizations (e.g., organizations working to protect the environment, like GreenPeace or the World Wildlife Fund) and governmental organizations and initiatives (e.g., central government sub-divisions that need to tackle environmental issues). For example, the message drawn from Twitter shown in Figure 8 includes several aspects and a different sentiment can be detected with respect to each of them: *ecological-atmosphere: positive* and *managerial-people's habits: negative*. Some of the topical focus categories shown in Table 4 can also be employed as ecological and managerial aspects.

Our findings reported above show that in the observed research area, there is a certain interest in hierarchical topic modeling. However, this approach assumes that a single sentiment and topical focus can be found in each document. Meanwhile, the literature on structured SA is relatively limited, even in the general domain, since this research field is still recent. In Barnes et al. [6], the authors propose a structured form of SA that aims at extracting the holder, target, and expression including its polarity and intensity level. For structured SA in natural environment topics, this structure can be modified, e.g., by categorizing the holder and target for providing the stakeholders' information about who is giving the opinion (e.g., citizen or news channel). The topical focuses shown in Table 4 can be employed as labels to link to the sentiment expressions.

6.2.2 Data Availability. The development of datasets is one of the biggest challenges in SA, since the lack of data organized as linguistic resources may be a bottleneck for the advancement of artificial intelligence models. On the one hand, resources allow us to make explicit the phenomena to be detected by SA techniques. On the other hand, they are very helpful for training models and

critical for their assessment. Also the performance of few- or zero-shot text classification models is indeed typically evaluated against a dataset. For SA applied to natural environment topics, we found five major challenges in building a dataset, which includes dataset public availability and validity, language variety, topical focus, data types, and cross-source.

(1) **Dataset public availability and validity**
We found that almost all the authors that annotated their dataset did not make their data open to the public research community, with the only exception of Hagerer et al. [31]. We think that this issue should be addressed in the future, but we also notice that the development of an annotated corpus for SA is not an easy task. This depends on the subjectivity involved in the evaluation of sentiment and on the subsequent bias to which the annotators can be subject in judging the sentiment. Among the solutions applied for addressing this issue in SA applied on general/other topics, we can in particular take into account the involvement of multiple expert annotators or crowdsourcing users, by considering their background variety to reduce the bias and monitoring the annotators, but also the distribution of detailed guidelines. However, a careful and reliable annotation process for the development of a resource can be very time-consuming.

(2) **Language variety**
We found that the datasets for SA in natural environment topics are still mostly focused on English. Nevertheless, since the natural environment comprises a wide set of global issues, the availability of resources for more languages and models built on top of them is gradually becoming a pressing need, at least as test sets.

(3) **Topical focus**
As observed in Section 5.2, in this survey we have found papers discussing a broad variety of natural environment topical focuses, but most of them are not annotating and opening their datasets to the public. There is therefore the need for openly accessible datasets with a wider variety of natural environment topical focuses, that allow us to evaluate SA models also in a cross-topical focus perspective.

(4) **Data types**
As discussed in Section 6.1.1, all of the papers considered in this survey use only textual datasets. Multimodal datasets may help the research in this field to discover different facets of sentiment [8, 94] on the topics of the natural environment, and open up venues for the analysis of social media and news data in a more complete setting.

(5) **Cross-source**
Datasets described in the surveyed papers are drawn from different sources, including social media, news articles, and others, but the only publicly available annotated dataset is collected from Twitter. The availability of datasets from different sources allows us to evaluate models in a cross-source perspective, considering the different domains that are discussed in different sources and the features they comprise.

6.2.3 Methodology. Building a robust model for SA on natural environment topics means facing even more difficult challenges than for other topics. This is indeed a fairly new and complex topic and has not been widely explored by previous research works. In particular, for SA in natural environment topics, we found the five major challenges listed below. Given the unavoidable relationship between resources and language models, these challenges are closely linked to those observed from a data development perspective in Section 6.2.2.

(1) **Limited data training**
The limited availability of data already observed in Section 6.2.2 makes especially challenging the development of data-driven models. This challenge may be addressed by developing

models that can learn from limited or even with no training data. In recent years, many researchers have developed several approaches for few-shot and zero-shot learning [27, 65] for SA in general topics and other NLP tasks that can be adapted to this topic.

(2) **Cross-topical focus and cross-source**
Building SA models for natural environment topics also consists of dealing with different topical focuses, that is also a specialized language with a different vocabulary distributions with respect to general topics. Moreover, a topical focus that is discussed on different sources can also have a different writing style. Despite the current state-of-the-art pre-trained language models, giving good results on many NLP tasks, their performance on cross-domain tasks is still not comparable to in-domain, especially when the language models are trained on the general text, such as Wikipedia [22]. Domain knowledge injection [18, 22] may alleviate this issue, as could re-training the language model by adding a new relevant topical focus dataset [10].

(3) **Multilinguality**
Considering that natural environment issues are faced and discussed by users in all countries in the world, multilingualism seems an unavoidable requirement for SA models to be applied to natural environment topics. This challenge can be addressed by adopting multilingual approaches, such as translation-based approaches or multilingual pre-trained models [3].

(4) **Multimodality**
Following the trend that currently people often express their opinion using modalities other than text, like images, audio, or video, future works for building models for SA in natural environment topics should take into account also multimodality. Several deep learning approaches which have been applied to SA in general topics, as summarized in the survey [2], could be extended to deal with natural environment topics.

(5) **SA subtasks**
Building a robust SA model for natural environment topics is not only about building a sentiment classifier. As presented in Figure 5, we can see that there is no previous research on SA in natural environment topics that tackle implicit or figurative language (i.e., irony, satire, and sarcasm) detection. However, implicit language often induces confusion in the classifiers [8, 51, 94]. To address this issue, we can adapt several deep learning approaches such as those explored by Ortega-Bueno et al. [64], Ren et al. [73]. Topic modeling and categorization are also important to make sure that the analyzed data is relevant to natural environment topics. Data filtering is another useful support task, in particular, to counter topic bias in the data. Some techniques in this direction have been proposed in the papers object of this survey (see Table 7). In some cases, these approaches could be improved, e.g., the manual filtering in Gaspar et al. [28] which could be partially or fully automated.

7 CONCLUSIONS AND FUTURE WORKS

This survey provides an overview of SA applied to natural environment topics. In exploring what has been done and what is still needed for SA in this area, we have identified seven major dimensions and then compared them with what has been done on SA in the general domain. In general, our findings show that this is a quite new research spotlight. By organizing concepts in taxonomies and visualizing them, we have shown that there are still gaps that need to be filled in order to obtain a good framework that can robustly classify the sentiment and give unbiased statistical SA on environmental topics.

We also found that SA on natural environment topics is still mostly focused on the English language and Twitter datasets, and that most of the approaches are based on lexicon libraries. The majority of the authors that build SA models employ classic machine learning approaches. Of

the annotated datasets created for training and testing models, only one is open to the research community. All of these findings indicate that this research topic development is still far behind with respect to general-domain SA and other NLP tasks.

Finally, some lessons can be learned for future work. An open and valid dataset for SA on natural environment topics is still missing. In particular, the community needs a dataset that covers wide topical focuses and languages that come from various sources so that the next researchers can focus on building a robust model. Besides building a dataset, there is also the need to define the appropriate structure for structured SA on natural environment topics. In this perspective, building a dataset annotated with sentiment holder, target and aspect could lead to more advanced SA models, more useful for all the natural environment stakeholders.

ACKNOWLEDGMENTS

Muhammad Okky Ibrohim thanks FSE REACT-EU for funding his PhD Research Projects dedicated to GREEN topics.

REFERENCES

[1] M. Abdar, M. E. Basiri, J. Yin, M. Habibnezhad, G. Chi, S. Nemati, and S. Asadi. 2020. Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets. *Renewable and Sustainable Energy Reviews* 124 (2020).

[2] Sarah A. Abdu, Ahmed H. Yousef, and Ashraf Salem. 2021. Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion* 76 (2021), 204–226. https://doi.org/10.1016/j.inffus.2021.06.003

[3] Marvin M. Agüero-Torales, José I. Abreu Salas, and Antonio G. López-Herrera. 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing* 107 (2021), 107373. https://doi.org/10.1016/j.asoc.2021.107373

[4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[5] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015* ; Conference date: 07-05-2015 through 09-05-2015.

[6] Jeremy Barnes, Laura Ana Maria Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval-2022 Task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, Seattle.

[7] Alec Biehl, Ying Chen, Karla Sanabria-Véaz, David Uttal, and Amanda Stathopoulos. 2019. Where does active travel fit within local community narratives of mobility space and place? *Transportation Research Part A: Policy and Practice* (May 2019), 269–287. https://doi.org/10.1016/j.tra.2018.10.023

[8] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134. https://doi.org/10.1016/j.knosys.2021.107134

[9] Paweł Brzustewicz and Anupam Singh. 2021. Sustainable consumption in consumer behavior in the time of COVID-19: Topic modeling on Twitter data using LDA. *Energies* 14, 18 (2021). https://doi.org/10.3390/en14185787

[10] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, Online, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3

[11] Y. Chen and Z. Zhang. 2022. Exploring public perceptions on alternative meat in China from social media data using transfer learning method. *Food Quality and Preference* 98 (2022).

[12] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[13] Jaewoong Choi, Janghyeok Yoon, Jaemin Chung, Byoung-Youl Coh, and Jae-Min Lee. 2020. Social media analytics and business intelligence research: A systematic review. *Information Processing & Management* 57, 6 (2020), 102279. https://doi.org/10.1016/j.ipm.2020.102279

[14] S. Chung, M. Chong, J. S. Chua, and J. C. Na. 2019. Evolution of corporate reputation during an evolving controversy. *Journal of Communication Management* 23, 1 (2019), 52–71.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104 arXiv:https://doi.org/10.1177/001316446002000104

[16] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLOS Computational Biology* 9, 2 (02 2013), 1–16. https://doi.org/10.1371/journal.pcbi.1002854

[17] Monir Dahbi, Rachid Saadane, and Samir Mbarki. 2019. Social media sentiment monitoring in smart cities: An application to Moroccan dialects. In *Proceedings of the 4th International Conference on Smart City Applications* (Casablanca, Morocco) *(SCA'19)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3368756.3368997

[18] Luna De Bruyne, Orphee De Clercq, and Veronique Hoste. 2021. Emotional RobBERT and insensitive BERTje: Combining transformers and affect lexica for Dutch emotion detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Online, 257–263. https://aclanthology.org/2021.wassa-1.27

[19] J. Dehler-Holland, M. Okoh, and D. Keles. 2022. Assessing technology legitimacy with topic models and sentiment analysis – The case of wind power in Germany. *Technological Forecasting and Social Change* 175 (2022).

[20] Joris Dehler-Holland, Kira Schumacher, and Wolf Fichtner. 2021. Topic modeling uncovers shifts in media framing of the German Renewable Energy Act. *Patterns* 2, 1 (2021), 100169. https://doi.org/10.1016/j.patter.2020.100169

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[22] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4019–4028. https://doi.org/10.18653/v1/2020.acl-main.370

[23] Xu Du, Matthew Kowalski, Aparna S. Varde, Gerard de Melo, and Robert W. Taylor. 2019. Public opinion matters: Mining social media text for environmental management. *SIGWEB Newsl.* Autumn, Article 5 (Feb. 2019), 15 pages. https://doi.org/10.1145/3352683.3352688

[24] Paul Ekman. 1999. *Basic Emotions*. John Wiley & Sons, Ltd, Chapter 3, 45–60. https://doi.org/10.1002/0470013494.ch3 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013494.ch3

[25] Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf

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

[27] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

[28] Rui Gaspar, Cláudia Pedro, Panos Panagiotopoulos, and Beate Seibt. 2016. Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior* 56 (2016), 179–191. https://doi.org/10.1016/j.chb.2015.11.040

[29] D. M. Goldberg, S. Khan, N. Zaman, R. J. Gruss, and A. S. Abrahams. 2020. Text mining approaches for postmarket food safety surveillance using online media. *Risk Analysis* (2020).

[30] Gerhard Hagerer, Wing Sheung Leung, Qiaoxi Liu, Hannah Danner, and Georg Groh. 2021. A case study and qualitative analysis of simple cross-lingual opinion mining. *CoRR* abs/2111.02259 (2021). arXiv:2111.02259 https://arxiv.org/abs/2111.02259

[31] Gerhard Hagerer, David Szabo, Andreas Koch, Maria Luisa Ripoll Dominguez, Christian Widmer, Maximilian Wich, Hannah Danner, and Georg Groh. 2021. End-to-end annotator bias approximation on crowdsourced single-label sentiment analysis. In *Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. Association for Computational Linguistics, Trento, Italy, 1–10. https://aclanthology.org/2021.icnlsp-1.1

[32] Qi Han, Junfei Guo, and Hinrich Schuetze. 2013. CodeX: Combining an SVM classifier and character n-gram language models for sentiment analysis on Twitter text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, 520–524. https://aclanthology.org/S13-2086

[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[34] Wen-Tai Hsieh, Chen-Ming Wu, Tsun Ku, and Seng-cho T. Chou. 2012. Social event radar: A bilingual context mining and sentiment analysis summarization system. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, Jeju Island, Korea, 163–168. https://aclanthology.org/P12-3028

[35] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA) *(KDD'04)*. Association for Computing Machinery, New York, NY, USA, 168–177. https://doi.org/10.1145/1014052.1014073

[36] Rocío B. Hubert, Elsa Estevez, Ana Maguitman, and Tomasz Janowski. 2018. Examining government-citizen interactions on Twitter using visual and sentiment analysis *(dg.o'18)*. Association for Computing Machinery, New York, NY, USA, Article 55, 10 pages. https://doi.org/10.1145/3209281.3209356

[37] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H. Oh (Eds.). The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

[38] Z. Indra, A. Setiawan, and Y. Jusman. 2021. Implementation of machine learning for sentiment analysis of social and political orientation in Pekanbaru city. In *Journal of Physics: Conference Series*, Vol. 1803.

[39] Ye Jiang, Xingyi Song, Jackie Harrison, Shaun Quegan, and Diana Maynard. 2017. Comparing attitudes to climate change in the media using sentiment analysis based on Latent Dirichlet Allocation. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Association for Computational Linguistics, Copenhagen, Denmark, 25–30. https://doi.org/10.18653/v1/W17-4205

[40] I. Jun, Y. Zhao, X. He, R. Gollakner, C. Court, O. Munoz, J. Bian, I. Capua, and M. Prosperi. 2020. Understanding perceptions and attitudes toward genetically modified organisms on Twitter. In *ACM International Conference Proceeding Series*. 291–298.

[41] Zenun Kastrati, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. *Applied Sciences* 11, 9 (2021). https://doi.org/10.3390/app11093986

[42] Edward F. Kelly and Philip J. Stone. 1975. *Computer Recognition of English Word Senses*. North-Holland Pub., Amsterdam.

[43] D. Kim, S. Kang, and S. Park. 2019. Bi-LSTM sentiment classifier for climate change issues in South Korea. *International Journal of Recent Technology and Engineering* 8, 2 Special Issue 6 (2019), 295–299.

[44] Y. Kirelli and S. Arslankaya. 2020. Sentiment analysis of shared tweets on global warming on Twitter with data mining methods: A case study on Turkish language. *Computational Intelligence and Neuroscience* 2020 (2020).

[45] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Keele University. Technical Report TR/SE-0401. Department of Computer Science, Keele University, UK.

[46] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30, 3 (01 2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x arXiv:https://academic.oup.com/hcr/article-pdf/30/3/411/22338169/jhumcom0411.pdf

[47] Akshi Kumar and Abhilasha Sharma. 2020. Socio-sentic framework for sustainable agricultural governance. *Sustainable Computing: Informatics and Systems* 28 (2020), 100274. https://doi.org/10.1016/j.suscom.2018.08.006

[48] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[49] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=H1eA7AEtvS

[50] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[51] Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review* 54 (2021), 4997–5053.

[52] Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

[53] Ruixue Liu and Jing Xiao. 2021. Factors affecting users' satisfaction with urban parks through online comments data: Evidence from Shenzhen, China. *International Journal of Environmental Research and Public Health* 18, 1 (2021). https://doi.org/10.3390/ijerph18010253

[54] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[55] F. Loia, N. Capobianco, and R. Vona. 2022. Towards a resilient perspective for the future of offshore platforms. Insights from a data driven approach. *Transforming Government: People, Process and Policy* 16, 2 (2022), 218–230.
[56] Yujie Lu, Jinlong Guo, Kotaro Sakamoto, Hideyuki Shibuki, and Tatsunori Mori. 2015. Predicting sector index movement with microblogging public mood time series on social issues. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China, 563–571. https://aclanthology.org/Y15-1065
[57] Yafeng Lu, Michael Steptoe, Sarah Burke, Hong Wang, Jiun-Yi Tsai, Hasan Davulcu, Douglas Montgomery, Steven R. Corman, and Ross Maciejewski. 2016. Exploring evolving media discourse through event cueing. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 220–229. https://doi.org/10.1109/TVCG.2015.2467991
[58] Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, and Kalina Bontcheva. 2017. A framework for real-time semantic social media analysis. *Journal of Web Semantics* 44 (2017), 75–88. https://doi.org/10.1016/j.websem.2017.05.002 Industry and In-use Applications of Semantic Technologies.
[59] C. Michael and D. N. Utama. 2021. Social media based decision support model to solve Indonesian waste management problem: An improved version. *International Journal of Emerging Technology and Advanced Engineering* 11, 10 (2021), 1–12.
[60] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
[61] Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Los Angeles, CA, 26–34. https://aclanthology.org/W10-0204
[62] F. C. Moore, N. Obradovich, F. Lehner, and P. Baylis. 2019. Rapidly declining remarkability of temperature anomalies may obscure public perception of climate change. *Proceedings of the National Academy of Sciences of the United States of America* 116, 11 (2019), 4905–4910.
[63] Mazdak Nik Bakht, Tamer E. El-Diraby, and Moein Hosseini. 2018. Game-based crowdsourcing to support collaborative customization of the definition of sustainability. *Advanced Engineering Informatics* 38 (2018), 501–513. https://doi.org/10.1016/j.aei.2018.08.019
[64] Reynier Ortega-Bueno, Paolo Rosso, and José E. Medina Pagola. 2022. Multi-view informed attention-based model for Irony and Satire detection in Spanish variants. *Knowledge-Based Systems* 235 (2022), 107597. https://doi.org/10.1016/j.knosys.2021.107597
[65] Ramakanth Pasunuru, Veselin Stoyanov, and Mohit Bansal. 2021. Continual few-shot learning for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5688–5702. https://doi.org/10.18653/v1/2021.emnlp-main.460
[66] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate G. Blackburn. 2007. *The Development and Psychometric Properties of LIWC2007*. Technical Report.
[67] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate G. Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. Technical Report.
[68] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162
[69] M. Pilgun, A. Rashodchikov, and O. Koreneva Antonova. 2021. Environmental digital conflicts: Spanish-, German-, and Russian-speaking actors. *Revista Latina de Comunicacion Social* 79 (2021), 303–332.
[70] Robert Plutchik. 1994. *The Psychology and Biology of Emotion*. HarperCollins College Publishers.
[71] Martín Pérez-Pérez, Gilberto Igrejas, Florentino Fdez-Riverola, and Anália Lourenço. 2021. A framework to extract biomedical knowledge from gluten-related tweets: The case of dietary concerns in digital era. *Artificial Intelligence in Medicine* 118 (2021), 102131. https://doi.org/10.1016/j.artmed.2021.102131
[72] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - A publicly available German-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf
[73] Lu Ren, Bo Xu, Hongfei Lin, Xikai Liu, and Liang Yang. 2020. Sarcasm detection with sentiment semantics enhanced multi-level memory network. *Neurocomputing* 401 (2020), 320–326. https://doi.org/10.1016/j.neucom.2020.03.081
[74] Ana Reyes-Menendez, José Ramón Saura, and Cesar Alvarez-Alonso. 2018. Understanding #WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. *International Journal of Environmental Research and Public Health* 15, 11 (2018). https://doi.org/10.3390/ijerph15112537

[75] H. Roberts, B. Resch, J. Sadler, L. Chapman, A. Petutschnig, and S. Zimmer. 2018. Investigating the emotional responses of individuals to urban green space using Twitter data: A critical comparison of three different methods of sentiment analysis. *Urban Planning* 3, 1 (2018), 21–33.

[76] L. Rocca, D. Giacomini, and P. Zola. 2020. Environmental disclosure and sentiment analysis: State of the art and opportunities for public-sector organisations. *Meditari Accountancy Research* 29, 3 (2020), 617–646.

[77] Arman Sarjou. 2021. The power of language: Understanding sentiment towards the climate emergency using Twitter data. *CoRR* abs/2101.10376 (2021). arXiv:2101.10376 https://arxiv.org/abs/2101.10376

[78] A. Satish, S. B.++ Shankar, and K. N. Kavitha. 2021. Naagarik: A machine learning framework for intelligent analysis of civic issues. In *2021 Asian Conference on Innovation in Technology, ASIANCON 2021*.

[79] Ainhoa Serna, Tomas Ruiz, Jon Kepa Gerrikagoitia, and Rosa Arroyo. 2019. Identification of enablers and barriers for public bike share system adoption using social media and statistical models. *Sustainability* 11, 22 (2019). https://doi.org/10.3390/su11226259

[80] L. Serrano, A. Ariza-Montes, M. Nader, A. Sianes, and R. Law. 2021. Exploring preferences and sustainable attitudes of Airbnb green users in the review comments and ratings: A text mining approach. *Journal of Sustainable Tourism* 29, 7 (2021), 1134–1152.

[81] Zhongkai Shangguan, Zihe Zheng, and Lei Lin. 2021. Trend and thoughts: Understanding climate change concern using machine learning and social media data. *CoRR* abs/2111.14929 (2021). arXiv:2111.14929 https://arxiv.org/abs/2111.14929

[82] A. Singh and A. Glińska-Neweś. 2022. Modeling the public attitude towards organic foods: A big data and text mining approach. *Journal of Big Data* 9, 1 (2022).

[83] B. Sluban, J. Smailović, S. Battiston, and I. Mozetič. 2015. Sentiment leaning of influential communities in social networks. *Computational Social Networks* 2, 1 (2015).

[84] Yunya Song, Xin-Yu Dai, and Jia Wang. 2016. Not all emotions are created equal: Expressive behavior of the networked public on China's social media site. *Computers in Human Behavior* 60 (2016), 525–533. https://doi.org/10.1016/j.chb.2016.02.086

[85] Swati Srivastava, Juginder Pal Singh, and Deepak Mangal. 2020. Time and domain specific Twitter data mining for plastic ban based on public opinion. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. 755–761. https://doi.org/10.1109/ICIMIA48430.2020.9074935

[86] Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online, 8–18. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.2

[87] E. Sugiharti and D. Fauziah. 2021. Comparative study between KNN and maximum entropy classification in sentiment analysis of menstrual cup. In *Journal of Physics: Conference Series*, Vol. 1918.

[88] T. E. Taufek, N. F. M. Nor, A. Jaludin, S. Tiun, and L. K. Choy. 2021. Public perceptions on climate change: A sentiment analysis approach. *GEMA Online Journal of Language Studies* 21, 4 (2021), 209–233.

[89] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4067–4076. https://www.aclweb.org/anthology/2020.acl-main.374

[90] Anina Troya, Reshmi Gopalakrishna Pillai, Dr. Cristian Rodriguez Rivero, Dr. Zulkuf Genc, Dr. Subhradeep Kayal, and Dogu Araci. 2021. Aspect-based sentiment analysis of social media data with pre-trained language models *(NLPIR 2021)*. Association for Computing Machinery, New York, NY, USA, 8–17. https://doi.org/10.1145/3508230.3508232

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[92] Yan Wang, Shangde Gao, Nan Li, and Siyu Yu. 2021. Crowdsourcing the perceived urban built environment via social media: The case of underutilized land. *Advanced Engineering Informatics* 50 (2021), 101371. https://doi.org/10.1016/j.aei.2021.101371

[93] Zhibo Wang, Lei Ke, Xiaohui Cui, Qi Yin, Longfei Liao, Lu Gao, and Zhenyu Wang. 2017. Monitoring environmental quality by sniffing social media. *Sustainability* 9, 2 (2017). https://doi.org/10.3390/su9020085

[94] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* (2022). https://doi.org/10.1007/s10462-022-10144-1

[95] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45, 4 (Dec. 2013), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

[96] Jiangmei Xiong, Yulin Hswen, and John A. Naslund. 2020. Digital surveillance for monitoring environmental health threats: A case study capturing public opinion from Twitter about the 2019 Chennai water crisis. *International Journal of Environmental Research and Public Health* 17, 14 (2020). https://doi.org/10.3390/ijerph17145077

[97] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[98] Linlin You and Bige Tunçer. 2016. Exploring public sentiments for livable places based on a crowd-calibrated sentiment analysis mechanism. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 693–700. https://doi.org/10.1109/ASONAM.2016.7752312

[99] Dachuan Zhang, Haoyang Zhang, Zhisheng Wei, Yan Li, Zhiheng Mao, Chunmeng He, Haorui Ma, Xin Zeng, Xiaoling Xie, Xingran Kou, and Bingwen Zhang. 2021. IFoodCloud: A platform for real-time sentiment analysis of public opinion about food safety in China. *CoRR* abs/2102.11033 (2021). arXiv:2102.11033 https://arxiv.org/abs/2102.11033

[100] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery* 8, 4 (2018), e1253. https://doi.org/10.1002/widm.1253 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1253

[101] Y. Zhang, M. Abbas, and W. Iqbal. 2021. Analyzing sentiments and attitudes toward carbon taxation in Europe, USA, South Africa, Canada and Australia. *Sustainable Production and Consumption* 28 (2021), 241–253.

[102] Anastazia Zunic, Padraig Corcoran, and Irena Spasic. 2020. Sentiment analysis in health and well-being: Systematic review. *JMIR Med. Inform.* 8, 1 (28 Jan. 2020), e16023. https://doi.org/10.2196/16023

[103] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv:1103.2903 [cs.IR].

Received 22 July 2022; revised 29 April 2023; accepted 31 May 2023