Environmental Modelling and Software 187 (2025) 106382

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Position Paper

SeqIA: A Python framework for extracting drought impacts from news archives

Miguel López-Otal [a,*], Fernando Domínguez-Castro [b,d], Borja Latorre [c], Javier Vela-Tambo [d], Jorge Gracia [a]

[a] Aragon Institute of Engineering Research, Universidad de Zaragoza, Mariano Esquillor s/n, Zaragoza, 50018, Spain
[b] Departamento de Geografía y Ordenación del Territorio, Facultad de Filosofía y Letras, Universidad de Zaragoza, C. San Juan Bosco, 7, Zaragoza, 50009, Spain
[c] Estación Experimental Aula Dei, CSIC (Spanish National Research Council), Av. Montañana, 1005, Zaragoza, 50059, Spain
[d] Pyrenean Institute of Ecology, Av. Montañana, 1005, Zaragoza, 50059, Spain

ARTICLE INFO

Keywords:
Drought
Impacts
Newspaper
Spain
Classification
Python

ABSTRACT

Drought is a hazard that causes great economic, ecological, and human loss. With an ever-growing risk of climate change, their frequency and magnitude are expected to increase. While there are many indices and metrics available for the analysis of droughts, assessing their impacts represents one of the best ways to understand their magnitude and extent. However, there are no systematic records outlining these impacts.

To help in their ongoing creation, we present a software framework that leverages raw newspaper articles, identifies any drought-related ones, and automatically classifies them according to a set of socioeconomic impacts. The information is provided to the user in a structured format, including geographical coordinates and their date of reporting. Our approach employs state-of-the-art Transformer-based Natural Language Processing (NLP) techniques, which achieve great accuracy. We currently support newspaper articles in the Spanish language within Spain, but our framework can be expanded to other countries and languages.

Software and data availability

- Name of software: SeqIA
- Year first available: 2024
- Developers: Miguel López-Otal
- License: Open source under BSD 3 clause license
- Hardware and software requirements: Windows or Linux-based machine, CUDA-compatible GPU with 16 GB VRAM or more (can optionally run without GPU, but its use is highly recommended).
- Programming language: Python.
- Source code access: github.com/sid-unizar/seqia
- Training datasets and model weights access: https://doi.org/10.20350/digitalCSIC/16540

The raw newspaper articles used for building the training dataset for drought news detector are not available due to copyright issues, although we do provide the articles' URLs for reproducibility purposes. However, the datasets for drought impact identification modules are available from the repository listed above. We also provide open access to the weights of the trained models — necessary for the framework to function.

The files that comprise the coordinates database module of the system are available from the Github repository, and instructions are provided on how to download them from their original sources.

Open access is also available to the end-to-end evaluation newspaper dataset for reproducibility purposes.

1. Introduction

Drought is among the natural hazards that cause the most losses and fatalities. For example, a drought that began in Ethiopia and Sudan, in 1983, resulted in 450,000 victims. Economic loss is also significant. For instance, a 1994 drought in China caused loss estimated at 23.72 US$ billion, while a 2012 drought in the United States amounted to 21.79 US$ billion (WMO, 2021). The impacts arising from droughts are complex and affect various systems, including agricultural production (Garcia-Garizabal et al., 2014; Peña-Gallardo et al., 2019),

* Corresponding author.
E-mail addresses: mlopezotal@unizar.es (M. López-Otal), fdominguez@unizar.es (F. Domínguez-Castro), borja.latorre@csic.es (B. Latorre), jvela@ipe.csic.es (J. Vela-Tambo), jogracia@unizar.es (J. Gracia).

https://doi.org/10.1016/j.envsoft.2025.106382
Received 1 October 2024; Received in revised form 7 February 2025; Accepted 17 February 2025
Available online 24 February 2025

hydroelectric production (van Vliet et al., 2016; Zhao et al., 2023), and human supplies (Nobre et al., 2016), among others. Moreover, drought also has a notable environmental impact (Vicente-Serrano et al., 2020).

Agricultural and ecological droughts have been increasing in severity over recent decades, particularly in arid regions. This is mainly due to an increase in atmospheric evaporative demand (AED) caused by rising temperatures (Vicente-Serrano et al., 2020). This process appears likely to intensify in future scenarios, potentially leading to even more severe droughts. Moreover, certain regions may experience decreased precipitation according to models (Vicente-Serrano et al., 2022) – i.e. North and Central America, the northern part of South America and the Amazon basin, southwestern America, the Mediterranean region, western and southern Africa and South Australia.

High losses and the likely increase in the severity of future droughts make them a fundamental field of research concerning adaptation to climate change. Unfortunately, there is no single physical variable to quantify droughts, making it significantly challenging to measure when and where a drought begins, as well as its magnitude or severity (Wilhite, 2000). Consequently, numerous indices and indicators have been developed to "measure" the development of a drought (Svoboda et al., 2016). However, how we truly perceive a drought is through its impacts. Quantifying these impacts is the best way to measure the severity, magnitude, duration, and extent of a drought. They also represent a valuable source of information to know how vulnerable a specific region or hydrological systems are to drought. Nevertheless, there are few databases for drought impacts, and those that do exist have significant limitations. For example, the University of Freiburg maintains a database of impacts mainly concerning the review of socioeconomic reports (Stahl et al., 2016). Similarly, the European Union has presented another project in the same line: the European Drought Impacts Database (Stahl et al., 2023). While these are both commendable initiatives, the mentioned databases have notable gaps, and their level of usefulness varies across countries. Hence, there is an urgent need to generate detailed and high-quality drought impact databases at local, regional, and global levels. One challenge in creating these databases is that impact information is highly disparate and recorded heterogeneously in reports. One documentary source that has been recording drought impacts since the late 17th century is newspapers. They document drought impacts when they affect human activities or ecosystems of interest.

Using newspaper data to extract information on drought impacts has been shown to be efficient across numerous research studies, including those conducted in Ireland O'Connor et al. (2023), Eva et al. (2022), the United States (Dow, 2010), Australia (Bell, 2009; Hurlimann and Dolnicar, 2012; Rutledge-Prior and Beggs, 2021), the United Kingdom (Dayrell et al., 2022), and Spain Llasat et al. (2009), Ruiz Sinoga and León Gross (2013). However, these studies have typically been confined to short time frames or limited geographical areas due to the manual nature of the drought impacts extraction. Such manual processes represent a significant drain on resources and constrain the scope of research objectives. To address this challenge, we propose using a software framework that has been designed to automatically derive drought impacts from newspaper articles. This framework promises to streamline research efforts, saving both time and resources, while also enabling broader and more comprehensive exploration of drought impacts.

To automate the generation of drought impact databases from newspapers we have designed the 'SeqIA' software framework. This framework is capable of working with raw, unstructured texts as provided by newspaper articles, classify them depending on their detected drought impacts, and ultimately derive a series of structured database entries with specific drought impact events, organized in chronological order and geographical location. To achieve this, this project has built on Natural Language Processing (NLP) technologies, most specifically current state-of-the-art Language Models based on the Transformer architecture (Vaswani et al., 2017). The architecture is provided in open access as a simple-to-use Python package, available from the following Github repository: github.com/sid-unizar/seqia.

Our framework provides a custom architecture that is well suited for the detection of drought reports in the media, which have unique characteristics compared to other natural hazards –e.g. heat or cold waves, floods, hailstorms–, such as their slow development, long duration and the effect they have on a wide range of socioeconomic sectors. These features of drought render existing information extraction systems unsuitable for this task, whereas our framework has been designed from the ground-up to account these characteristics of drought.

While the package benefits from powerful GPU resources to run its Transformer-based models, it can also function on CPU-only machines. GPUs are recommended for optimal performance, although not strictly required. Our framework software has been tested in Spain since it is a region where droughts are recurrent and produce a wide range of impacts on socioeconomic sectors and the environment (Gonzalez-Hidalgo et al., 2018). Due to the diversity of atmospheric mechanisms implicated in the occurrence of droughts and the topographical heterogeneity of the region, the variations concerning duration, severity and magnitude of droughts are extensive in both time and space (Domínguez-Castro et al., 2019). The framework, however, can be adapted for use with other countries as well, given some manual work.[1]

This paper is organized as follows. First, we are going to describe the related works that exist for the automatic detection of extreme weather impacts. We are then going to introduce our framework and its individual components, detailing their internal strategies and functionality, followed by a description of the training datasets that were built to create these systems. This will be followed by a thorough explanation of our used evaluation strategies and obtained results, which aim to measure the performance of both the individual modules of our system and the framework as a whole, in its intended final task. We are then also going to expose a small practical examination of a series of drought impacts detected by our framework in Spain in a collection of newspapers articles ranging from 1991 to 1995, a period during which that country underwent a major drought event (Trullenque-Blanco et al., 2024). Next, we are going to discuss the advantages and disadvantages of our approach. Finally, concluding remarks and possible lines of future work will be outlined.

2. Related work

Machine-learning based solutions have already been used in the past to detect many types of climate extremes from newspaper sources, not only droughts but also other phenomena such as floods. Among these we can mention the work done by Yzaguirre et al. (2015), focusing on floods; Liu et al. (2018), which uses data mining techniques to extract information on several events such as hailstorms or floods in China; Domala et al. (2020), which introduces a software framework designed to automatically provide crisis management websites, in real time, with newspaper articles related to various types of climate hazards; Akbari et al. (2022), which relies on an ontology-based approach to extract information from a source text regarding several types of natural disasters. There is also Gopal et al. (2023), which aims to recover, from newspaper articles, the impacts of different types of natural hazards in India. One further example is that of Lai et al. (2022), which focuses on flood events and relies on a Named Entity Recognition (NER) model implementation. We can also mention Otudi et al. (2024) for the detection of weather extremes from social media, information which is

[1] Thanks to our framework's main language being Spanish, this means we can also target other Spanish-speaking countries, such as Mexico or Colombia, for the detection of drought impacts. This functionality would be straightforward to implement, as it would only require the creation of a suitable database of geographical database for each country.

used as support data for monitoring these events when weather sensors are found to fail during severe weather conditions. Recently, we also have the work of Madruga de Brito et al. (2025), which analyze flood impacts by relying on text mining techniques, as well as that of Zou et al. (2024), focusing on typhoons, and relying on several Natural Language Processing-based technologies. Tounsi and Temimi (2023) provides a good overview of several additional NLP-based solutions.

Droughts have unique characteristics compared to other natural hazards, such as their slow development, long duration and the effect they have on a wide range of socioeconomic sectors. As a result, the detection of drought impacts can present some difficulties, which require the implementation of more elaborate software solutions to solve it, rendering solutions such as the ones presented by publications such as Liu et al. (2018) or Lai et al. (2022) –among others–, used to detect other types of weather phenomena, unsuitable for drought, owing to its more complex nature.

Regarding the reporting of drought-related events, Zhang et al. (2022) presents TweetDrought, which leverages posts from the social platform formerly known as Twitter (now X) to detect a series of drought impacts on ongoing Californian droughts as reported by users, using Transformer-based models. A similar system, proposed by Musaev et al. (2018), which also focuses on Californian droughts, uses posts from this platform as well, although this research focuses on multimodal input (text and images) and, among other techniques, creates a series of word clouds containing the most important keywords found in the overall reported posts. Other research carried out that uses Twitter/X data is that of Mukherjee et al. (2022). However, in this case the use of newspaper data is different: this framework is based on the analysis of drought indices, and only leverages data from social media as a way of deepening their analysis on drought. The aforementioned social media-based papers, while dealing with the detection of drought events from a contemporary information source, have one main limiting factor, which is in fact their reliance on social media posts. This content, due to it being open access user-generated media, is prone to bias and subjectivity.

On top of that, when using social media posts, researchers can find an added difficulty, which is determining the relevance of each post. The reason is that any given post may have been retrieved for a study by unreliable factors such as a user's prior popularity on a platform.

Sodoge et al. (2023) suggests an R-based framework to derive drought impact information from German newspaper sources. Their software framework in this case is presented only as a means of reproducing their reported results, and displays some difficulties in its use as a real-time system.[2] Another similar research study is that of Pita Costa et al. (2024), which presents a framework to detect newspaper articles, not only on drought but on other hydrological events as well –e.g. floods. It also has support enabling the use of multiple languages and leverages weather metrics as a supporting source of information. They rely on a custom method by which each article is processed by a tool called 'Wikifier' –which performs entity linking between a text and a series of concepts found in Wikipedia pages across several languages, creating a language-agnostic representation in the process — and then passes it through a fastText-based classifier. Duarte et al. (2024) is another solution that uses NLP-based tools to identify any water resource problems found in news media – not necessarily drought – in a specific area of Colombia. It then correlates that newspaper data with actual weather reports.

Our framework, compared to the aforementioned proposals, represents the first solution for the detection of drought impacts in Spain, a Mediterranean country prone to recurring drought events. Other than this, our framework also presents a series of technical advantages – e.g. the use of Transformer-based classifiers – which help it derive this type of information automatically, with little to no human supervision, whereas the existing solutions require some manual work to obtain a series of results — see Section 8.

3. Software framework overview

We have developed a Python package called 'seqIA'. This package offers a series of API functions that allow a straightforward classification process of newspaper articles. The main package functionality is implemented in a class called 'DroughtClassifier'.

Internally, the package is implemented as a pipeline (Fig. 1):

1. Article loading and preprocessing: including removal of duplicates, correction of text codification to Unicode (UTF-8) and text normalization.
2. Keyword-based filtering (optional): detects whether an article contains a series of keywords that are commonly associated with drought.
3. Drought news detector: detects if an article is truly related to droughts (whether meteorological, agricultural, etc.). Makes use of a Transformer-based binary classifier. This step can filter out false positives –i.e. articles where the use of the word 'drought' is metaphorical and not on an actual drought phenomenon.
4. Drought impacts identification: if an article is drought-related, it is passed through a series of classification modules that detect what sorts of impacts the article details.
5. Drought impact locator: detection and cross-reference with a static and curated list of geographical coordinates. Allows extraction of specific locations for detected drought impacts.

Our package relies on two popular NLP libraries for their general operations: 'spaCy' (Honnibal et al., 2020) (which is used to separate an input text into a list of sentences – a technique used in some modules – as well as obtain the text dependency parse trees) and HuggingFace's 'transformers' package (Wolf et al., 2020) – used for the classification tasks.

The framework output is that of a Python list with a series of dictionaries, one for each of the analyzed articles. For each dictionary there is a boolean value that tells whether the article is drought-related or not, a list with all the detected drought impacts in a string format, and a list of all detected geographical location names with their corresponding geographical coordinates data. The latter one is explained in more detail in Section 3.5.

This output format[3] can be serialized to JSON for further manual analysis, although our library also offers the option to adapt it to a more human-readable format – e.g. CSV – with a list of exported drought impact names and their locations.

3.1. Article loading and preprocessing

As a first step, the framework loads a set of articles to be analyzed from disk. Our implementation requires that each individual article be stored in a separate JSON file, conforming to the 'NewsArticle' Schema.org specification (https://schema.org/NewsArticle).

The framework detects the encoding format of each of the articles. Since our framework is designed to preferably work with Unicode encoding (UTF-8), it will automatically convert any non-Unicode article files to this encoding. The framework also checks for duplicate articles and removes them accordingly.

As a last step, our framework also runs each article through a text normalization function, which standardizes variations of some similar orthographic expressions, such as quote characters. It also removes some characters that may contain non-relevant information, such as HTML tags.

[2] Namely, the need to manually revise each new corpus of analyzed news to discard false positives, as we will explain in more detail in Section 8.

[3] A comprehensive description of the complete, raw output format can be found in our Github repository, inside the 'OUTPUT_FORMAT.MD' file.

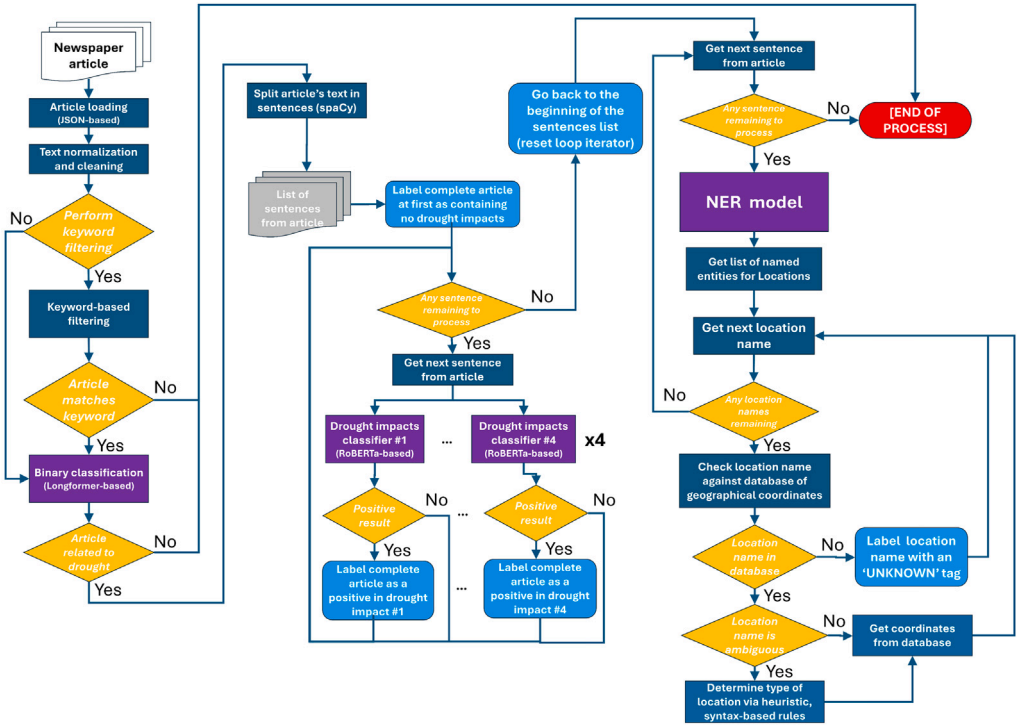

Fig. 1. An overview of the package and its internal components.

3.2. Keyword-based filtering (optional)

Once the newspaper articles are loaded and cleaned, an optional first step is to individually pass them through a keyword-based filter, which selects a series of articles based on the presence of either of the following keywords: 'sequía' ('drought' in Spanish), 'agua' ('water') 'o' or 'lluvia' ('rain'). Overall, this step serves as the first way of obtaining a large number of drought-related articles, while discarding most non-related articles. This is especially useful when handling large amounts of newspaper corpora, as it can alleviate the computing time that the rest of the framework would require to classify many unrelated articles.

3.3. Drought news detector

Each article is then passed through a Transformer-based binary classifier. The purpose of this classifier is to confirm whether each analyzed article deals with the reporting of a drought event or not on a general basis. This classifier serves as a barrier, ensuring that non-related articles are not being propagated to the rest of the system. A critical step that is also undertaken by this module is that of detecting articles that might be false positives, where either the keyword 'drought' is used in a metaphorical sense[4] or where a reported meteorological event is anything but a drought (e.g. a hailstorm). This step allows us to detect these kinds of false positives and only process those related to drought.

In addition, this model has been trained not only to distinguish drought-related articles from false positives, but also to completely filter out unrelated articles. This makes it possible to use this module for news classification and not have to rely whatsoever on the use of the keyword filtering module from Section 3.2. The use of the latter module, however, can provide reduced computing time given large amounts of analyzed newspaper articles.

The chosen Transformer-based architecture is a fine-tuned Longformer (Beltagy et al., 2020) model, an architecture that contains a variation of the self-attention mechanism proposed by Vaswani et al. (2017), making it capable of processing longer sequences of text compared to regular Transformer models. This is essential to ensure the correct handling of newspaper articles, whose length usually exceed the maximum size accepted by many Transformer-based architectures — such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) (512 BPE – Byte-Pair Encoding – tokens). With a Longformer-based model, which can handle up to 4,096 BPE tokens, the entirety of an article can, in most cases, be read in one setting, which allows the model to see the entire context of each article to be analyzed.

The model was fine-tuned from a pretrained Longformer model in Spanish,[5] provided by Fandiño et al. (2022). More details on the training parameters used for this model can be found in Appendix A.1.

3.4. Drought impacts identification

Our system currently supports the detection of impacts in four areas: (i) agriculture, (ii) livestock, (iii) hydrological resources, and (iv) energy.

Internally, this step consists of two interconnected sub-modules (Fig. 2). The first one is a software component, based on the use of the Python NLP library 'spaCy' (Honnibal et al., 2020), that separates the text from each article into its set of individual sentences. Alongside the textual content of these sentences, we also keep the dependency parse trees for each of the sentences. These are also obtainable via the spaCy library, and will be used in the drought impacts locator step — in order to help choose the correct form for some ambiguous location names that happen to be shared by a river or a city or town (see Section 3.5).

In the following step, each sentence is analyzed in search of references of drought impacts within them. Each sentence serves as the input for the four Transformer-based binary classifiers, one for each drought

[4] As in sports articles; e.g. "La **sequía *goleadora*** de [Iván] Zamorano empieza a ser alarmante. Desde el cinco de diciembre el chileno no marca un gol en partido de Liga".

[5] Available at https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es.

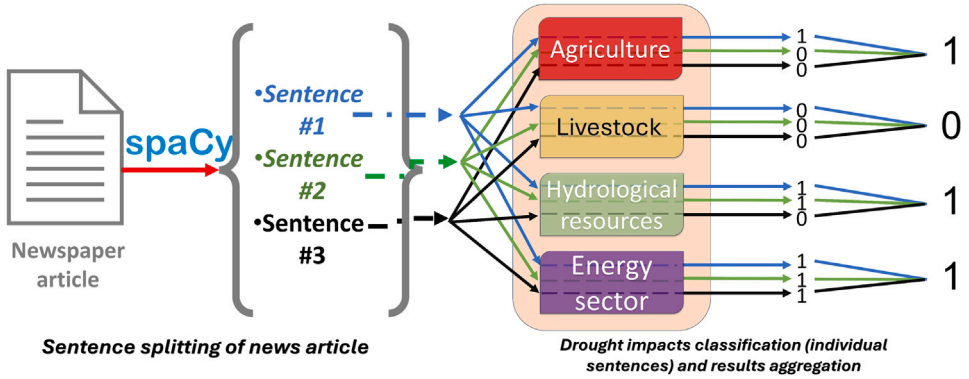

Fig. 2. An overview of the drought impacts identification step of the pipeline.

impact. Each classifier has been trained to detect whether an individual sentence deals with a specific drought impact or not. We rely on a series of fine-tuned Spanish-based RoBERTa (Liu et al., 2019) models.[6]

More details on the training parameters used for these models can be found in Appendix A.2. The training datasets used to build this system are detailed in Section 4.2.

If at least one of the sentences from an accepted article is detected to be related to one drought impact, such impact is assigned to the complete newspaper article.

By carrying out the classification of newspaper articles using its set of individual sentences – instead of the raw, complete article – we are able to vastly simplify the architecture of the module, as we can skip the input limits of many Transformer-based models, such as RoBERTa (Liu et al., 2019) – which, as stated in Section 3.3, only accepts up to 512 BPE tokens, usually less than that of a regular newspaper article.[7]

3.5. Drought impact locator

This module examines each article, one sentence at a time, in order to find named locations where drought impacts may have occurred. These locations range from towns, cities, rivers, basins, dams, reservoirs to intra-national regions (such as provinces).

The module first performs an automatic recognition process of named locations with a RoBERTa-based model, fine-tuned for Named Entity Recognition (NER).[8] This specific model (Fandiño et al., 2022) is designed to work as a token classification task that labels each token from a text according to their named entity type: person, organization, location, etc. For our purposes, we only leverage the named entities detected as locations.

Our module matches each identified location with a curated geo-referenced database of Spanish named locations (see Section 4.4 for details). Our model gives each location name its corresponding set of geographical coordinates as stored in the database.[9] If a reference for a location name is not found within the database, the model gives the location a generic 'UNK' ('UNKNOWN') tag instead.

A key challenge addressed in this module step is location name disambiguation. Our database contains entities of various types, like towns and rivers, that can share the same name. This step ensures accurate identification of the intended location. This part of the module first detects the cases in which an accepted name is shared across multiple toponym types. In that situation, the module runs a disambiguation function that is based on the use of some heuristic syntactic rules, which rely on a set of dependency parse trees obtained for each of the sentences in the previous step (see Section 3.4 and Appendix E).

4. Data

We are going to we introduce a series of training and evaluation datasets that were built to train various modules in our system, more specifically, the drought news detector (Section 4.1) and the drought impact identification modules (Section 4.2). We are also going to present a geographical coordinate database built for use in the drought impact locator module (Section 3.5).

Additionally, we also decided to run a series of experiments on our framework end to end in order to test its effectiveness on a real case scenario. For this purpose, we built a small-scale dataset that imitates the input of our system, so as to test its performance as a whole and not only that of its individual modules. This experiment is explained in more detail in Section 5.

All these data sources are openly available for allowing easy replication of the obtained results and training of the models[10] –except for the data for the drought impact locator module, which consist of a series of geographical files stemming from a series of preexisting databases (see Section 4.4).

4.1. Drought news detector training dataset

The data for this training set originates from a large pool of news taken from the Spanish national newspaper El País, with articles ranging from 1976 to 2023. To collect these articles, we first performed a keyword-based search for pieces containing the keywords 'sequía' ('drought') and 'agua' ('water'), and then leveraged some additional articles that were found to contain the tag 'Sequía' in the El País newspaper website.[11] These two references formed a raw source of potentially positive articles. This was expanded with a series of unrelated articles retrieved from random newspaper categories, to serve as a source of potentially negative articles.

[6] The original pretrained model checkpoint can be retrieved at https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne.

[7] We had already made use of a non-standard architecture to solve the input length limit problem for our binary classifier in Section 3.3, via a fine-tuned Longformer (Beltagy et al., 2020) model, but we decided against reusing it for the drought impacts identification modules; one of the reasons lied in that it is a memory-intensive architecture that would be too costly to be used as a set of multiple binary classifiers.

[8] Available at https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner.

[9] Coordinates are provided as latitude and longitude points for towns and cities, and as PolyLines for the other types of geographical entities –e.g. rivers and provinces.

[10] Located at https://doi.org/10.20350/digitalCSIC/16540.

[11] We could not assume this tag represented a gold standard for drought articles, however, since we are unsure if the tagging process was performed manually or semiautomatically by the newspaper maintainers.

Table 1
Number of document instances used for building the training and test datasets of the drought news detector.

Dataset	Positives	Negatives	Total articles
Train	902	643	1545
Test	366	288	654
Complete	1268	931	2199

In order to balance[12] this collection of articles, we then examined the newspaper website tags for each article, and manually determined two categories of tags that would aid in the identification of drought related, as well as unrelated, articles. The first one, which we called *negative* tags, dealt with topics completely unrelated to drought. Examples include the following tags in Spanish: 'deportes' ('sports'), 'sucesos' ('social events'), 'viajes' ('trips'), 'cine' ('cinema') or 'música' ('music'). The other one, *related* tags, included some tags found in drought articles as well as in other articles that, while not reporting on drought, dealt with similar topics (e.g. storms, floods...). Examples include 'lluvia' ('rain'), 'meteorología' ('weather forecast'), 'embalses' ('dams'), 'inundaciones' ('floods')...

These two categories allowed us to gear our data collection strategy towards the following collection of documents[13]:

- 2000 articles from the positive corpus that did not belong to the set of negative tags
- 1000 articles from the negative corpus that belonged to the set of negative tags
- 480 articles from the positive corpus that, coincidentally, belonged to the set of negative tags
- 1000 articles from the negative corpus that belonged to the set of related tags — hence potentially being a list of non-drought articles that discussed similar topics to positive articles

We then randomly selected a set of 2199 news articles from this raw corpus of 4480 articles — roughly 50% of the overall corpus. The total number of chosen articles was motivated by the need to reduce the workload required by manual annotation, while still dealing with enough articles to successfully train our system.

From these articles, 1268 were eventually labeled as drought-related, while the remaining 931 were labeled as negatives. Two human annotators performed the labeling task on the same set of documents, and later reached a global consensus if any differences arose with the annotation. This was performed according to the guidelines outlined in Appendix B. The inter annotator agreement rate, measured in Cohen's kappa coefficient, was 0.833 ± 0.011, which according to Landis and Koch (1977) means an "almost perfect" agreement rate.

Finally, this set of articles was further divided into two subsets: 70% for a training set (1545 articles) and 30% for an evaluation set (654 articles). We can see the total distribution of training instances detailed in Table 1.

4.2. Drought impact identification training datasets

We built four training datasets for each drought impact identification module — i.e., agriculture, livestock, hydrological resources and energy. Each dataset consists of a series of labeled sentences related to each impact.

Table 2
Number of sentences used in each of the four training datasets for the drought impact identification modules.

Dataset	Subset	Positives	Negatives	Total
Agriculture	Train	228	622	850
	Test	93	274	367
	Combined	321	896	1217
Livestock	Train	97	757	854
	Test	36	333	369
	Combined	133	1090	1223
Hydrological	Train	113	752	865
	Test	51	322	373
	Combined	164	1074	1238
Energy	Train	104	761	865
	Test	42	331	373
	Combined	146	1092	1238

As a means of building this corpus, we leveraged a small collection of articles from El País's corpus, while at the same time also relying on a different collection of articles from those in Section 4.1, so as to gather some more diverse training data. For the latter purpose, we also retrieved a series of articles from different regional newspapers belonging to Spanish news conglomerate Prensa Ibérica ("La Nueva España", "La Opinión de A Coruña", "La Provincia", "Diario de Ibiza", "La Opinión de Málaga", "Faro de Vigo", "La Opinión de Zamora", "Diario de Mallorca", "Información", "El Periódico Mediterráneo", "Diario de Córdoba", "El Periódico de Extremadura", "El Periódico de Aragón" and "Levante-EMV") via a simple keyword-based search using the words 'sequía' and 'agua'. We then manually compiled a list of keywords commonly associated with each type of drought impact we sought to recover.[14] This collection of keywords – which are outlined in Appendix C – was used as a guideline to select a set of 353 articles, divided into a preliminary set of four categories corresponding to the different drought impacts chosen to annotate.

With this content-based division, ten human annotators were given the task of manually extracting and labeling from each section a series of sentences from these articles that they identified as being representatives of each specific drought impact. These were labeled as positive instances for the dataset of each impact identification module. In Appendix C we can see the criteria used for selecting each sentence type. If any sentence contained more than one possible impact, it was discarded, to ensure each sentence referenced only one type of impact so as to ease the training of the systems. Subsequently, the final collection of sentences was also revised by two supervisors to ensure its correctness.

In order to obtain a set of negative instances for these datasets, first, we collected all positive sentences from every other dataset, and then combined these with another collection of 502 manually selected sentences that contained no discernible drought impacts. From this combined pool of sentences, we randomly selected several of them to be used for the negative set of each of the datasets. The number of total sentences (positives and negatives) for each training set can be seen in Table 2.

4.3. Dataset for end-to-end evaluation

This is an additional test dataset that was created for the purpose of an end-to-end evaluation strategy, whose goals and motivations are explained more in detail in Section 5 and its results in Section 6.2.

[12] These imbalances were due to having derived the initial collection of negative articles from random newspaper pieces. As a result of this data collection strategy, some topics present in the positives corpus – e.g. agriculture – could potentially be missing in the negatives one. This could cause our system to incorrectly assume that, for instance, all agriculture-related articles are of our interest — even if no drought is discussed.

[13] Due to copyright issues, this raw pool of newspaper articles cannot be shared as-is; instead, we provide the URLs for the retrieved articles – though not their raw text content –, for reproducibility purposes, in the following URL: https://doi.org/10.20350/digitalCSIC/16540.

[14] An example of this would be 'cosechas' ('crops') or 'campo' ('field') for Agriculture impacts.

Our goal when creating this corpus was to make it a representative sample of the type of articles that our system would likely find in a real-world application. We focused on the subset of articles that contained the keyword 'sequía' ('drought') within our original pool of newspaper articles collected in Section 4.1. This specific selection of articles was chosen because it is a close match to the kind of newspaper reports that the system would likely find by relying on the keyword-based filter module from Section 3.2. Additionally, since a number of these articles might be false positives, this would also be a good opportunity to test the ability of our model to detect this type of articles.[15]

Following that, we took a look at the newspaper sections where each of those articles originally came from and derived some percentages on the distribution of the categories in the overall corpus. This way, we obtained a statistical representation of the content and themes of all articles containing the keyword 'drought'. Sections that contained less than 1% of the overall corpus were excluded. Each percentage, rounded up, was then multiplied by 2, with the objective to collect around 200 news articles — with the final number being 194. This number was chosen with the idea to keep the balance between the effort needed for human annotation and the amount of test cases needed for an illustrative evaluation. The annotation process was performed by one expert. Each of these articles were then manually read and labeled as either drought-related or not. In the case of positive articles, these were also annotated with a series of detected drought impacts. The corpus also contains several positive articles that do not have any reported impact.

The final dataset consists of a series of 194 news articles, 159 of which are drought-related articles, while the remaining 35 are negatives.

4.4. Drought impact locator database

This collection consists of a series of files – used by the drought impact locator module presented in Section 3.5 – containing geographical information, available for download from different reliable sources, and whose contents stem from the online platform Geonames (Geonames, 2022) and from several geographical datastores provided by Spanish national institutions — namely those of the Ministerio para la Transición Ecológica (2024)[16] and Instituto Geográfico Nacional (2024).[17] Some additional information, such as town names, also originates from 'Instituto Nacional de Estadística' ('INE'; Spanish National Statistics Institute),[18] as well as from Wikidata via the use of SPARQL queries.

The file formats in this database are heterogeneous, due to them being created by different agents, and range from simple TSV files to geographical-specific data formats such as KML/KMZ, GFS or GML files. Our framework contains suitable codes that load these different data formats and combine them into a single, unified, in-memory lookup list.

As a whole, our database contains references to urban settlements (cities, towns, villages...), dams, reservoirs, river basins, provinces and autonomous communities (a Spanish-specific regional entity).

5. Evaluation methods

In this section we will present a series of different methodologies we have used to evaluate the performance of our proposed framework.

We make use of the following widespread evaluation metrics: Accuracy, Precision, Recall and F1. These are commonly used in the literature to evaluate the performance of machine learning-based solutions (Pedregosa et al., 2018). They are calculated as follows:

- Accuracy: Number of correctly predicted instances compared against the rest of predictions and their true values.
$$Acc = \frac{True\ Positives + False\ Negatives}{All\ predictions}$$

- Precision: metric that measures how many instances classified as positives are actually positives (i.e. true positives) and how many were actually false positives.
$$Prec = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall: similar metric to Precision, but measures, from all the positives in a test corpus, how many of them a classifier has been able to correctly identify and how many it has omitted as false negatives.
$$Rec = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- F1: mean value of the Precision and Recall values obtained when measuring a system. This provides a single score that is meant to measure both metrics.
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

These metrics were used to evaluate the drought news detector and drought impacts identification modules — by comparing each individual module against the test set that was part of their respective training dataset. These were used as well to evaluate the results obtained for an end-to-end dataset (see Section 4.3) that was built to test our framework as a whole — i.e. with all its modules working in combination. In this experiment we would run our framework through an unseen corpus of news articles and report its obtained results in a multilabel style. Other than using the metrics already mentioned above – Accuracy, Recall, Precision and F1 –, we also relied on two additional metrics that are commonly used in the context of multilabel classification: 'Receiver Operating Characteristic - Area Under the Curve' (ROC-AUC) score (micro-averaged) and Hamming loss. These consist on the following criteria:

- ROC-AUC measures the ability of a model to differentiate between the different labels it supports. It consists of two steps:

    1. Calculation of a 'ROC' curve, which is created by first dividing the Recall value of a label with its False Positive Rate (FPR), and then by plotting that result against a set of thresholds. The obtained curve shows the correspondence between the number of true and false positives for a specific label. False Positive Rate (FPR) is calculated as follows:
    $$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

    2. Once the curve is obtained, we calculate the 'Area Under the Curve'. As the name implies, it represents how much of the plotted area lies under the ROC curve. The broader the area under the curve is, the better the model is at identifying a specific label. It produces a single value, ranging from 0 to 1, where higher scores mean better performance.

[15] We acknowledge that this selection of newspaper articles, which was based on the criteria of containing the Spanish keyword 'sequía', might involve a certain degree of bias with this evaluation strategy, as a selection of random articles would have been preferable due to being more representative. However, due to time constraints, this is an issue that we are leaving for future study.

[16] Available at https://www.mapama.gob.es/app/descargas/descargafichero.aspx?f=RiosCompPfafs.zip and https://www.mapama.gob.es/app/descargas/descargafichero.aspx?f=egis_embalse_geoetrs89.kmz.

[17] Available at https://centrodedescargas.cnig.es/CentroDescargas/documentos/atom/au/lineas_limite_gml.zip.

[18] Derived from https://www.ine.es/daco/daco42/codmun/codmun11/11codmun.xls.

For each label we calculate its corresponding ROC-AUC value, and we then obtain the micro-average of them all.

- Hamming loss, on the other hand, is a metric that obtains the proportion of labels that have been incorrectly identified by a model. It can be seen as the opposite of Accuracy and, in contrast with this metric, is able to handle labels from each classified instance individually, hence making multilabel classification results more interpretable.

  Hamming loss L_H for a set of labels y, and their corresponding predictions $\hat{y}$, is calculated as follows (Pedregosa et al., 2018):

$$L_H(y, \hat{y}) = \frac{1}{n_{\text{samples}} * n_{\text{labels}}} \sum_{i=0}^{n_{\text{samples}}-1} \sum_{j=0}^{n_{\text{labels}}-1} 1(\hat{y}_{i,j} \neq y_{i,j}),$$

  where:

  - n_{samples} is the number of samples.
  - n_{labels} is the number of labels present in each sample (multilabel-style classification).
  - $y_{i,j}$ is the truth value of a label j in a sample i.
  - $\hat{y}_{i,j}$ is the predicted value of a label j in a sample i.

  It yields a number from 0 to 1, where a lower score is preferable, as it indicates fewer mistakes.

6. Results

In this section we will present the results of the different evaluation metrics presented in Section 5, which either aimed to test the performance of each of the individual modules in our framework, or that of our complete framework in a simulated real-world scenario, with a custom built end-to-end dataset (see Section 4.3), which is meant to test our framework when exposed to its intended end task of deriving drought impacts information from raw newspapers.

6.1. Results for individual modules

6.1.1. Evaluation of drought news detector

In Fig. 3(a), we can see the confusion matrix obtained from running our test set against our Transformer-based drought news detector. The results shown are excellent, with only 14 misclassified instances, and with our module achieving excellent metrics, which are listed in the first row of Table 3.

We have also measured the classification performance of a simple, keyword-based classifier – based on the use of the word 'sequía' ('drought') – to serve as a baseline. We also present another baseline that makes use of a broader set of keywords: 'sequía', 'agua' ('water') and 'lluvia' ('rain'). We must account for the fact that these two baselines – in particular the one centered on the use of the keyword 'sequía' – are expected to perform well with our specific test set, due to the fact that our evaluation dataset was partially built around a keyword-based strategy (see Section 4.1). However, the keyword-based technique alone cannot filter out the news articles that use the word 'drought' metaphorically and, on the other hand, it will also filter out any articles dealing with droughts but not using said keywords at all.

Additionally, we also compare the performance of both keyword-based filters when they are individually combined with our Transformer-based drought news detector.

We can see the obtained results for our keyword-based baselines, as well those that combine keywords with the Transformer news detector, in rows 2 to 5 of Table 3, as well as in a series of confusion matrices in Fig. 3(b).

Regarding the results from the keyword-based search with the term 'drought' only, the metrics perform well (0.869 F1 score), but are still inferior to the Transformer-based technique (0.98 F1 score). Critically, we can also see that there is a number of false negatives – 14 of them – that coincides with a set of ten positive articles within the test set that do not contain the 'drought' keyword at all.

If we take a look at the results of the keyword-based baseline that uses a broader set of keywords, we can see that the results are lower, showing a precision score of 0.745 – far from the 0.9 range of the other models – coupled with a high recall score of 0.95. This is a sign of the model underperforming and accepting too many false positives, something that can be observed in its corresponding confusion matrix. This demonstrates the need for the binary classifier module from Section 3.3, as without its use a big number of false positives and false negatives would be considered as drought-related and non-drought articles, respectively, by the system.

The keyword-based filter with the 'drought' keyword only, coupled with the Transformer classifier, is the solution that features the highest precision value (0.986) among the presented alternatives. The keyword-based filter with a broader set of keywords, combined with the Transformer classifier, features a slightly lower precision value but its recall score is much higher than its counterpart making use of the 'drought' keyword exclusively. Overall, this means that the optional keyword filter can be activated by the user if they wish to encourage higher precision in their results — and vice versa.

Table 3
Metrics for the drought news detector test set when run against our Transformer-based solution as well as two keyword-based baselines — with and without running them in combination with our Transformer-based classifier.

Model	Acc.	Prec.	Recall	F1
Transformer-based(w/o keywords filter)	**0.978**	0.978	**0.983**	**0.980**
'drought' keyword	0.865	0.954	0.797	0.869
('drought'|'water'|'rain') keywords	0.790	0.745	0.95	0.835
'drought' + Transformer-based	0.877	**0.986**	0.792	0.878
('drought'|'rain'|'water') + Transformer	0.940	0.976	0.915	0.944

Table 4
Metrics obtained during the training of the different drought impact identification modules.

Model	Acc.	Prec.	Recall	F1
Agriculture	0.939	0.932	0.802	0.862
Livestock	0.961	0.773	0.953	0.854
Hydrological resources	0.954	0.833	0.777	0.804
Energy	0.959	0.803	0.918	0.857

6.1.2. Evaluation of drought impact identification modules

For the impact identification modules, we can see in Fig. 4 the confusion matrices for their test sets.

The metrics obtained during training for drought impact systems are listed in Table 4. The results are reported in Accuracy, Precision, Recall and F1 (micro-averaged) scores.

For each individual drought impact identification module, the accuracy scores range between 0.93 and 0.95, with the F1 scores having a bit of a wider variation between 0.804 and 0.862. The latter can be attributed to some of the individual precision and recall scores being a bit lower in some cases. The most notable is the recall value for the hydrological resources label, with a score of 0.777, less than the rest of the impacts — which display values between 0.8 and 0.9. To better understand the types of classification errors made by the drought impacts identification module, we have included in Appendix D a list of sentences commonly misinterpreted by our system.

6.2. Complete framework evaluation results: end-to-end dataset

We will now present the results of testing our framework against a custom-annotated end-to-end dataset, presented in Section 5, whose

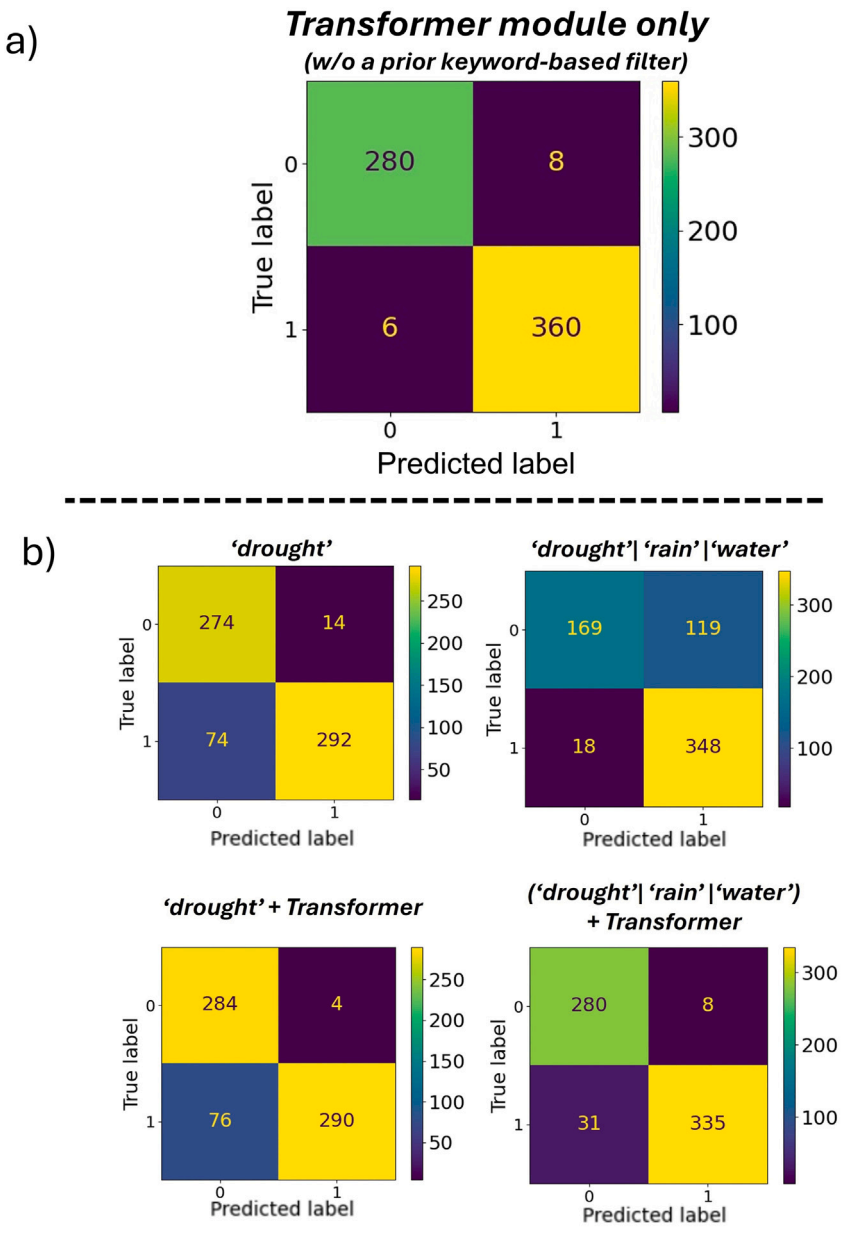

Fig. 3. Confusion matrix obtained for (a) the drought news detector module, compared against (b) a series of keyword-based approaches used to filter drought-related news — with or without using the Transformer-based classifier.

Table 5
Metrics obtained for the end-to-end evaluation dataset on the complete framework, reported in a multilabel classification setting.

Acc.	Prec.	Recall	F1	**ROC-AUC**	**Hamming**
0.788	0.735	0.806	0.769	**0.883**	**0.057**

purpose is to provide a perspective on how the complete framework – and not only its individual modules – would perform when being deployed in a real-world scenario, such as being integrated in a larger drought monitoring system.

The results obtained are listed in the confusion matrices presented in Fig. 5, and in the metrics in Table 5. These results have been calculated in a multilabel classification setting, with micro-averaging being used for the precision, recall and F1 scores.

As can be seen from the results, the metrics are lower overall in comparison to those outlined in sections 6.1.1 and 6.1.2, notably in regards to the Accuracy metric — 0.788. Among these results, those related to 'Agriculture' show a noticeable bias, especially when compared to the other categories, which means the system might not be performing well when replicating this category.

We must consider, however, that in a multilabel scenario the Accuracy metric is not the most indicative of the performance of a model. If we observe some of the other metrics – Precision, Recall and F1 –, we can see respectable scores, between 0.7 and 0.8. However, if we consider the scores obtained in the ROC-AUC – 0.883 – and Hamming Loss – 0.057 – metrics, we can see some relatively high scores which are promising.

If we regard the confusion matrices in Fig. 5, these show, in most cases, a good number of well-classified instances. This leads us to believe that, in a real-life research scenario, the end user would only have to filter out a relatively minor selection of false positives from their collected datasets — while only missing out some potential impacts (i.e. false negatives). Overall, this entails a much more reduced

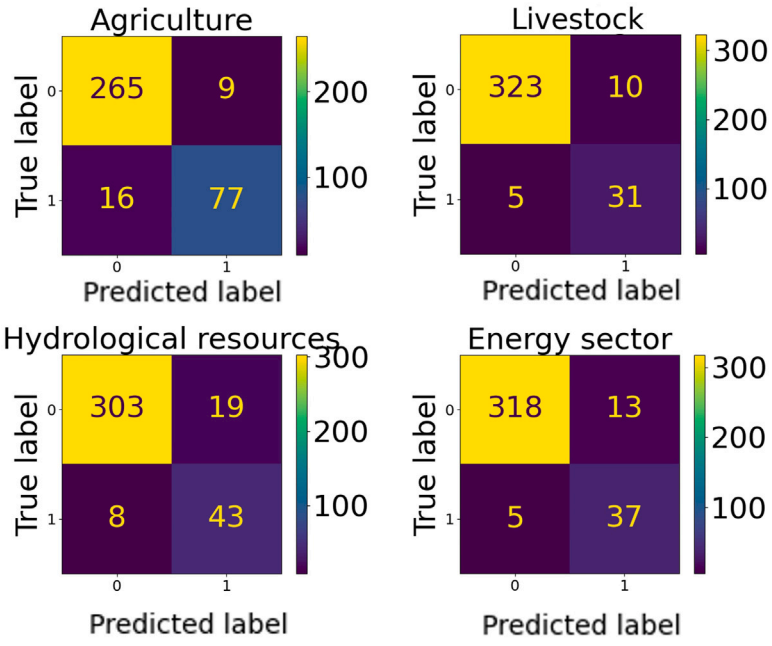

Fig. 4. Confusion matrices for the different drought impacts of the test dataset.

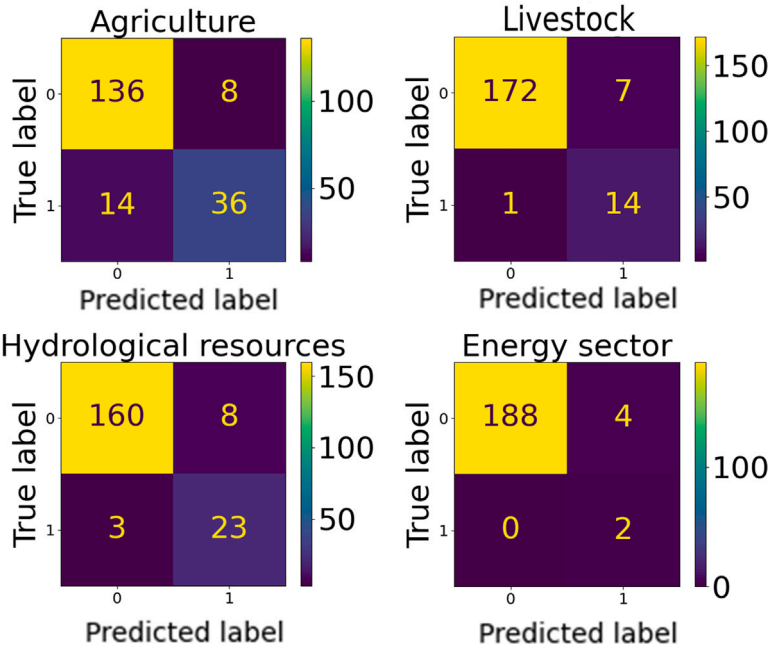

Fig. 5. Confusion matrices for the different drought impacts of the end-to-end evaluation dataset.

workload on the part of the researcher – compared to performing the entire classification work manually –, therefore, we consider that the benefits outweigh the outlined disadvantages.

7. Analysis of the 1991–1995 drought in Spain

As additional support to our provided results, while showing the output and application of SeqIA to a real case, we applied our framework to the period 1991–1995 from the newspaper El País. This period was selected because it represents one of the most significant droughts of the last century in Spain. In the drought catalog based on the Standardized Precipitation Index (SPI), compiled by Trullenque-Blanco et al. (2024), this period is divided into two dry pulses: the first one spanning from August 1991 to May 1993, and the second one from June 1993 to December 1995, with more than 40% of the territory showing SPI-12 values below −0.84. This resulted in streamflow reductions exceeding 70% in some basins, and water reservoirs below 10% of its capacity (del Guadiana, 2018). This drought event caused significant economic losses exceeding 600 million euros, and 8 million people experienced water restrictions as well, which in some cases reached 12-hour duration (Llamas, 2000). Furthermore, this drought marked a shift in mindset regarding the value of water, and triggered significant government-level measures against drought. However, there are no systematic studies of the impacts of the 1991–1995 drought like the one presented here.

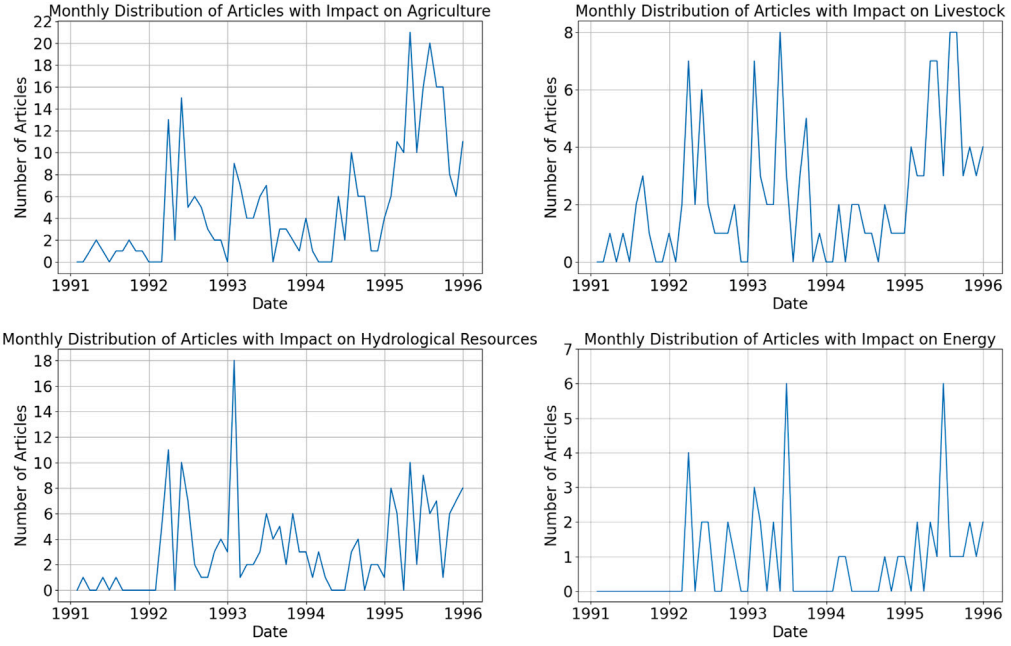

Fig. 6. Monthly rate of newspaper articles with detected drought impact in Spain in the period from 1991 to 1995.

Table 6
Number of drought impacts detected in individual news articles in the complete set of articles from newspaper El País from 1991 to 1995.

	1991	1992	1993	1994	1995	Combined
Agriculture	10	53	50	37	151	301
Livestock	9	24	34	13	57	137
Hydrological resources	3	47	55	17	70	192
Energy	–	11	13	5	19	48

From our corpus of news articles, we were able to extract a series of drought impacts from different categories, whose numbers are presented in Table 6. From a total of 303,784 articles published between 1991 to 1995 in newspaper El País, 1071 articles were detected by our system as being drought-related, and 550 of them contained at least one type of detected drought impact. In Fig. 6 we can see the number of articles per month for each of the detected drought impacts.

Most types of detected drought impacts did not start getting large coverage in the press until early 1992, with the exception of some impacts reports in agriculture and livestock in 1991. Around March 1992, however, all types of drought impacts, including hydrological resources and energy ones, started getting major press coverage.

From early 1992 to the end of 1993, the number of reported impacts is found to increase across all impacts, with a slight decrease in mid 1992. There is then a large reprise at the beginning of 1993 in all impacts as well, with the numbers of reports then declining in the middle of that year. The exception is livestock, where the impacts during all the year of 1993 are higher than ever. The year 1994 is found across all detected impacts to be when there seems to be less drought impacts reports compared to all the other years — except for that of 1991. The numbers for agriculture in 1994, however, are slightly higher compared to the rest of the impacts. The number of overall reports on drought rises again in the year 1995, except for hydrological resources, where there are less number of reported impacts compared to those of the same impact in early 1993 — yet are still relatively high. Energy sector impacts also show an increased number of reports in early 1993, coinciding with hydrological resources, but unlike the former it also displays a large increase in news reports towards mid 1995.

Interestingly, the evolution of impacts on agriculture and livestock is similar. The same can be said of energy sector impacts in regards to hydrological resources ones. This is explained due to the fact that the development of an agricultural drought will equally affect crops as well as cattle due to a decreasing number of pastures. On the other hand, the impacts in the energy sector will be more related to hydrological drought and a consequent reduction in streamflow and water reservoirs.

Fig. 7 shows the geographical location of the detected impacts. Each point corresponds to a different news article and type of impact. In order to better represent this information in our map, broader areas such as provinces or river basins were reduced to latitude/longitude pairs, which corresponded to the centroid point of their polygon shape. This was done to better graphically convey the number of occurrences of each specific area in regards to each drought impact type.[19]

The data provided in Fig. 7 represents the raw output from our model, which represents the geographical spread of a series of impacts in our studied time period, albeit some errors are also to be expected. As an example, we can mention the references which are usually found in the press regarding Madrid, the capital city of Spain. Its name is often used to metaphorically refer to the central government, which happens to be located there, so this may lead to some false positives being included.[20]

In 1991 all impacts types are fundamentally found in the northern part of the Iberian peninsula, with the exception of a few localized hydrological resources impacts in the south-east area. Then, in 1992, agriculture and livestock impacts started seeing more occurrences – mostly concentrated in the south-east region of Spain and sparsely located in the rest of the country –, and hydrological resources and energy impacts were mostly localized in the eastern and center areas. In 1993 all types of impacts were predominantly found in the center and south of Spain, with some spread concentrations as well throughout specific areas of the north — except for livestock.[21] 1994 sees a large

[19] Also, if a same type of impact happened in the same geographical area, this was represented as a pile-like reference.

[20] The exclusion of these types of references, however, is left for manual examination, and is not enforced as a design feature in our framework, since actual drought impacts happening in Madrid could be accidentally left out if an automatic exclusion was to be implemented.

[21] Exceptionally, the area surrounding Zaragoza, in north-east Spain, also showed great concentrations of hydrological resources and energy impacts.

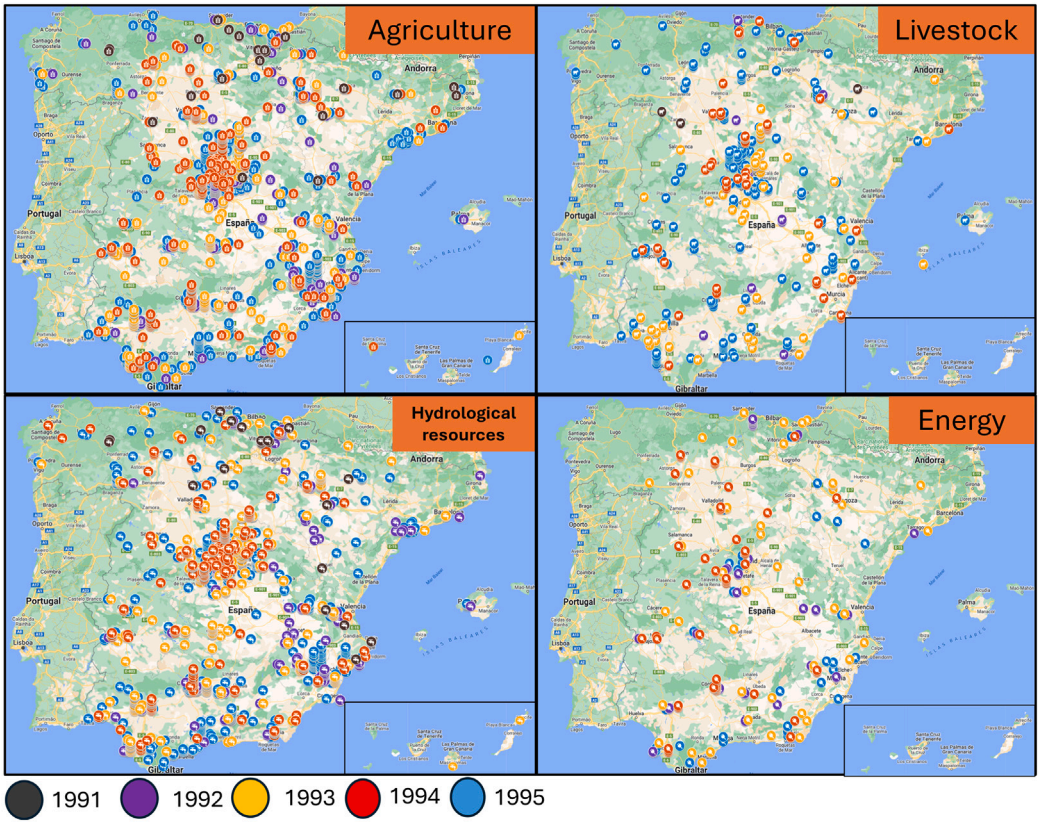

Fig. 7. Map with drought impacts of different type in Spain from 1991 to 1995.

increase of all types of impacts in the entire Spanish territory, with most of them found in center and southern areas, specially for agriculture and livestock impacts. Hydrological resources and energy sector impacts were also found that year in central northern areas, albeit more sparsely located. Finally, for the year 1995, we can see a similar spread and extension as those of 1994 between all types of impacts in the central and southern parts of Spain. The number of reported impacts, however, is much larger in some areas compared to the prior year — most notably, the south-east of the country as well as the Mediterranean coastline of Andalusia in southern Spain. The center area shows a larger number of livestock impacts.

In conclusion, as seen throughout this section, the detection of drought impacts is an interesting area of research that help us both understand the temporal and geographical spread of specific drought events, as well as link them to existing traditional drought metrics – e.g. SPI – in order to have a broader understanding of the effects of drought on the environment as well as on society. The results presented here for the 1991–1995 drought period in Spain, however, are merely illustrative, since a more in-depth analysis of the obtained results was out of the scope of this work. We wished instead to provide the reader with a glimpse of the possibilities that our framework and its ultimate objective – i.e. the extraction of socioeconomic drought impacts from press – can provide for the study of this weather hazard.

8. Discussion

Our architecture presents a few advantages in its design over other systems used for detecting weather hazards – different to drought – from newspaper articles or other sources of text – e.g. social media posts. For instance, as a first step in our system, we separately detect all drought-related articles, and filter out any unrelated ones, via a Transformer-based text classifier. We believe this design choice is crucial, as other similar frameworks for detecting weather hazards rely on other solutions, which are less ideal, for their detection of relevant texts. Most notably, many other works propose the use of keywords (Gopal et al., 2023; Sodoge et al., 2023; Zhou and Liu, 2024; Madruga de Brito et al., 2025) to perform such a step,[22] which is a method that is more brittle – e.g. due to having to manually define sets of keywords which could be polysemous or belonging to topics other than that of a hazard – and likely to introduce more unrelated texts in a corpus. Other works vouch for a custom classification method based on rule or ontology-based heuristics (Akbari et al., 2022; Lai et al., 2022), models which in some cases may prove complex to build, owing to their domain-specificity, and may also be prone to potential errors. Furthermore, all these approaches may require a significant amount of additional manual work to filter out any unrelated articles, such as is the case with Sodoge et al. (2023).[23] Meanwhile, our proposal does this same task with a Transformer-based solution, which is capable of detecting drought articles automatically.

Additionally, our framework is designed in a way that it helps deal with some unique characteristics of drought, which otherwise complicate their detection by standard information extraction systems.

[22] Interestingly, Zhou and Liu (2024), for the detection of tropical cyclones, proposes to combine a keyword-based approach with the use of Transformer-based models. This is done by first detecting name places via neural models and then comparing them against a predefined list of keywords containing the most common places for cyclones. This method, however, relies on the source texts being weather hazard reports from the start, so this information is already filtered out in their case.

[23] This work first uses a keyword-based approach (based on the use of the German words for drought 'Dürre' and 'Trockenheit') to filter out articles, and then rely on topic modeling techniques – such as LDA (Latent Dirichlet Allocation) – to manually remove false positives. The latter step, however, requires manual examination of the output of those models for each new corpus of text being handled.

In this sense, compared to other types of weather hazards, drought events present a very sparse effect across multiple different socioeconomic sectors, with press reports on this hazard being very varied and different one from the other. This great variety of themes makes it difficult for existing frameworks, not tailor-suited for drought, to accurately retrieve information on this hazard. This issue is solved in our case thanks to our drought news detector being trained on an expert-annotated collection of texts, which contain as many reports on drought from as many areas as possible — thus being potentially representative of the many different types of reports on drought that can exist in the media.

Also, since the impacts of drought and their consequent reflection in the press can spread over weeks, months or even years, existing software approaches for information extraction might present some difficulties as well in dealing with this type of information. This is not an issue with systems focusing in the detection of other types of weather hazards from text, such as floods (Lai et al., 2022), since these events tend to be reported in the press immediately after they have happened and are thus relatively simpler to locate.

To the best of our knowledge, there is only one other framework specific for drought impact detection from newspapers: Sodoge et al. (2023). Our research has an advantage over this work, however, in our use of Transformer models, state-of-the-art language models that are known to achieve high accuracy in tasks related to language. Sodoge et al. (2023), on the other hand, rely on more traditional text classification strategies to detect drought impacts in written press – in their case, lasso regression –, which do not necessarily achieve better results.[24] The main difference between these two proposals, however, is that the target language of our framework is Spanish, whereas the other one is based on German newspaper articles. This provides "SeqIA" with the potential to detect drought impacts within Spain, making it the first software framework effort of this type in this country.

At the time of writing, our framework supports the detection of drought impacts in the following four areas: agriculture, livestock, hydrological resources and the energy sector. However, the modular architecture of our system allows for the inclusion of other types of drought impacts in the future. This is accomplished thanks to the use of multiple individual binary classifiers for drought impact identification step, an architectural choice that facilitates the addition of new drought impacts, as it only requires the training and addition of a new classifier module to the stack — while keeping the existing impact identification modules intact. This would be more difficult to achieve if these modules were implemented in a multiclass or multilabel classifier setting instead — as used by other approaches such as Zou et al. (2024), in their case for detecting in social media texts different types of damages caused by typhoons.[25] These methods present a few disadvantages, such as needing more carefully curated training datasets to be built – as these models can prove sensitive to imbalanced data and some categories may require in some cases to be subjected to oversampling (Sáez et al., 2016) – and a full retraining of the classifier each time a new type of impact was to be added to the system. Moreover, in our framework, by having each drought impact being detected by a different individual classifier, it becomes easier to manually unearth any possible misclassifications that may arise and determine more easily what changes should be made to each module — something that would be more difficult to carry out in a multiclass or multilabel-style classifier.

Further regarding the architecture of the package, by classifying drought impacts through the individual sentences of an article – and not via the complete, raw article texts – we can obtain more fine-grained information about each of the detected impacts. This data would otherwise be more difficult to obtain for the researcher if the analysis were to be performed at the global text level. A practical example of this in our work is the detection of location names for drought impacts which, as explained in Section 3.5, is performed at the sentence level. It also makes it easier for the human researcher to investigate the decision-making strategies of the system, as well as to perform a more in-depth manual supervision of the classification results.

Finally, the development of accurate drought impact datasets are essential to advance in the assessment of the environmental, economic and social vulnerability and risk associated to droughts. In this line, our choice of relying on newspaper articles is crucial, since it is an excellent source of information – contrasted, objective and reporting on daily lives occurrences – to derive this sort of information, not only in a real-time setting but also for the last years and even centuries. From the perspective of drought monitoring, SeqIA could synergize with other additional data sources such as citizen science-inspired studies – e.g. Lackstrom et al. (2022) –, which encourage individuals to share information from their own personal devices, such as mobile phones, for obtaining fine-grained information on ongoing drought events.[26] In this same line, ongoing work on the real-time monitoring of drought could benefit as well from the inclusion of data extracted by IoT-based devices (Internet of Things), which could provide information on climate and environmental variables related with drought such as precipitation, soil moisture, temperature, radiation or streamflow, among others. There are existing examples of research on drought that leverage information from IoT devices — e.g. Shabbeer et al. (2016), Hoang et al. (2020) or Dahir et al. (2023).

9. Conclusions and future work

'SeqIA' is a software framework that enables the collection of drought impacts, from newspaper reports, across several socioeconomic sectors. It accomplishes this task by automatically processing a set of newspaper articles, detecting all drought-related ones and deriving from them a series of drought impacts, including their type and location, into a structured format. This makes it possible to apply our framework in a real-time scenario, where it could be integrated into a broader system that provides it with an incremental number of daily newspaper articles, helping create ongoing drought impacts datasets.[27]

The results displayed for our system indicate a good overall classification performance, with our individual modules, on the one hand, showing high classification scores, but on the other hand also observing good results in a series of preliminary tests ran in our framework with an independent, annotated dataset built to evaluate our system in an end-to-end fashion. Our analysis of a major drought event in Spain from 1991 to 1995 is also significant, as it demonstrates a practical application of our framework in a potential real-world scenario.

[24] The lasso logistic regression technique they used also requires some additional preprocessing steps that further complicate a text's interpretability or add unnecessary complexity, such as the need for lemmatization, removal of stopwords or calculation of TF-IDF scores.

[25] This research work also presents a separate step for the detection of texts of their interest, as we do in our solution, although with a different implementation: an LSTM classifier working with the raw output from a Transformer model.

[26] In part, there are existing proposals that attempt to do this to some degree: for instance, Zhang et al. (2022) performs this task for drought relying on posts from X/Twitter, and Otudi et al. (2024) does the same for general weather extremes, for which they combine contextualized embeddings – extracted from Transformer-based models after running Twitter posts through them – with meteorological sensor data. Beyond these works, however, this is something that should be further addressed in future work.

[27] As a quick demonstration, we ran our framework against a set of every single article, whether related or unrelated to drought, that was published in the Spanish newspaper El País from 1980 to 1982 — amounting to a total of 124,929 articles. It took our system around 12.5 h, running on a single NVDIA RTX A4000 GPU, to process the entirety of the dataset.

To further showcase the capabilities of SeqIA, we have released an accompanying visualization system.[28] This website allows users to input a news article URL, from which it extracts and highlights drought impacts within the text from the article. Additionally, the system maps the locations mentioned in the article, providing a geographical visualization of the impacts.

Our system, as of today, is tailor-suited for detecting contemporary drought impacts in newspapers, corresponding to what the classifier modules have seen in their training data. Written reflection of drought on the media, however, is bound to change as time passes and societies evolve, with newer types of drought impacts appearing or these affecting some socioeconomic sectors in which drought effects, as of today, are not as pronounced. Our framework, similarly to other frameworks that rely on training data – e.g. Liu et al. (2018), Gopal et al. (2023) –, is limited by any potential changes in the types of impacts or even the way these are reported in the press. In order to address this issue, we would recommend a periodical updating effort – e.g. every 5 to 7 years – of the training datasets in order to keep up with changing social visions on drought.[29] This would be alleviated by our framework's modular construction, which enable a straightforward identification of possible misclassifications in specific parts of the system and thus enables it to be changed accordingly — e.g. by upgrading the training datasets for an affected module. Additionally, also regarding our system's architecture, if a new type of drought impact started appearing in the future, it would also be feasible to add a new classification module to handle these new types of impacts.

Whereas our system is currently limited to the scope of Spain and the Spanish language, our framework could also be expanded with ease to detect drought impacts in other regions of the world. An example of this would be the detection of drought impacts in Latin American countries, in those where Spanish is spoken. This could be feasible thanks to Spanish being the current target language of our framework. Although the differences between European and American Spanish could not allow for a straightforward application of our framework to the case of these countries, we believe that this would be nevertheless feasible while only performing minor manual adaptation work: (i) adapting the existing training datasets presented in this work – where the existing Spain-based news sources in those datasets would be complemented with others originating from a specific country – and consequent retraining of the classifier modules; and (ii) the creation of a suitable database of geographical coordinates, one for each individual country. This would allow for our framework to be applied in potentially all Spanish-speaking countries in Latin America, and thus could prove helpful in aiding the drought research effort internationally, with a suitable working tool for detecting drought impacts in a language that is spoken across several dozen countries. The adaptation of this framework to non-Spanish speaking countries is also feasible, albeit more costly due to the need to gather annotated training data. The annotation process presented in this research, however, could serve as a starting point and inspiration for researchers willing to adapt this software framework to the language of their use.

Furthermore, there are several other features that will enhance SeqIA's usability. For instance, more sophisticated preprocessing of the articles will make it possible to detect duplicate texts that are not strictly identical, eliminating the need to classify articles from different news outlets that cover the same drought impacts. Currently, articles longer than the maximum token length of the base model of the binary classifier are discarded. Enabling the input of longer articles into the binary classifier will increase the number of articles processed and the number of impacts extracted.

In addition to the extraction of impacts and their locations from the articles, future studies might include the quantification of these impacts (e.g., specific crops affected by the drought, the percentage decrease in reservoir water levels, and financial losses incurred). This quantification would provide better insights into the impacts.

[28] https://siminelaki.org/seqia

[29] In a similar sense, historical texts – i.e. stepping back to several centuries ago – are likely to present this problem as well, due to these reflecting different socioeconomic realities – e.g. different types of impacts as those of today –, but also presenting other issues such as a markedly different language style, as well as a limited access to suitable datasets containing these sources and the texts themselves having been potentially subjected beforehand to a poor digitization effort — due to the use of faulty OCR techniques, among others. In future work we would like to address these issues as well in order to be able to handle historical texts on drought, which could help us derive drought impact information from text sources originating in time periods where corresponding drought metrics – e.g. SPI – did not exist yet.

CRediT authorship contribution statement

Miguel López-Otal: Writing – original draft, Validation, Software, Investigation, Data curation. **Fernando Domínguez-Castro:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Borja Latorre:** Writing – review & editing, Visualization, Validation, Software, Resources, Formal analysis. **Javier Vela-Tambo:** Writing – review & editing, Software. **Jorge Gracia:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the research projects PID2019-108589 RA-I00, TED2021-129152B-C41 and TED2021-129152B-C42, financed by the Spanish Ministry of Science and FEDER. This paper has been also partially supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón *y* Cajal" program (RYC2019-028112-I). It has also been partially funded by DGA Government predoctoral fellowship. We would also like to thank the Centro de Supercomputación de Galicia and CSIC for access to computer resources (CESGA and DRAGO respectively).

Appendix A. Training details

Training environment

For all the Transformer-based models mentioned below, training was performed on a non-GPU machine, with a set of 48 CPU cores, PyTorch compiled for CPU use, IPEX (Intel Extensions for PyTorch) enabled and 1 TB available RAM.

A.1. Drought news detector training details

- Architecture: Longformer-based Spanish checkpoint (PlanTL-GOB-ES/longformer-base-4096-bne-es at HuggingFace)
- Target task: Binary text classification
- Epochs: 5
- Learning rate: 5e−5
- Batch size: 8

A.2. Drought impact identification modules training details

A total of four RoBERTa-based models were trained, starting from the same Spanish-based checkpoint at https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne. All models shared the following hyperparameters:

- Architecture: RoBERTa-based Spanish checkpoint (PlanTL-GOB-ES/roberta-base-bne at HuggingFace)
- Target task: Binary text classification
- Epochs: 8
- Learning rate: 5e−5
- Batch size: 8

Appendix B. Drought news detector training dataset annotation guidelines

Here we reproduce the exact annotation statement established for when creating the drought news detector training dataset from Section 4.1:

> 'For this part of building the training dataset, the main problem lies in determining what articles we choose to be part of our set, and what articles we do not. On a general basis, we established that all articles that talk about droughts – whether direct impacts that have an influence in some area, or the possibility of droughts in the future – should be considered. For instance, if we find an article that simply mentions crop failures, but which fails to mention if it is due to a ongoing drought, is to be omitted. However, one that mentions that a crop was compromised due to an ongoing drought – even if the reference to that drought might be very slight and even circumstantial – is to be considered.

Additionally, we should also not consider an article that talks about preventive measures taken to address the impacts of future droughts, but that do not tackle a drought that is currently in progress. An example of this is the building of a dam as a long-term plan, but when a drought is not yet occurring.

In the cases, however, when an article mentions measures taken to alleviate a drought after it has happened, that article can be added to the corpus. That is because it does mention a consequence of an actual drought event — even if it is done at a later date. For instance, there can be an article that talks about public funds being given to livestock producers to recover from a particularly harsh drought period that has just finished.

We are also interested in detecting articles that mention the possibilities of hypothetical droughts in the future, which is a common place for many reports on global warming and similar issues.

Overall, all articles that mention any drought in progress are to be considered, no matter how slight the reference might be.'

Appendix C. Drought impacts training datasets annotation statements

Below we reproduce as-is the overall guidelines posed for the drought impacts annotation team, in order to collect a set of suitable sentences during the building of the training dataset from Section 4.2. Additionally, in Table C.7 we can see the manually crafted list of keywords used to preliminary select a series of articles depending on their potential drought impacts content.

- Agriculture:

1. We must find sentences that clearly mention that a drought has had an effect or damage – no matter which – in crops.
2. If a sentence fails to mention that a specific damage to crops is due to droughts, we have to carefully consider its inclusion.
   In some circumstances, we can include it due to the fact that, despite not mentioning droughts, it is implicitly understood that a specific circumstance is due to a drought happening. We have to remember that the sentences we leveraged for annotating this dataset stem from a preselected corpus of drought-related news. For instance, in this sense, the following sentence could be a viable option:
   → 'La semana pasada los agricultores de Les Garrigues ya daban casi por perdida la cosecha de este año, que comenzará la próxima semana.'
   → 'Last week farmers from Les Garrigues were almost certain about the loss of this year's crop, which is set to begin next week.'
   The sentence above has some clear impacts on Agriculture, even if drought is not mentioned directly.
   Other sentences, on the other hand, have no clear mention of actual agriculture impacts, and should therefore be removed. This is an example:
   → 'Rodríguez Leal reconoció, no obstante, que ha habido problemas puntuales, sobre todo en al agricultura, en la zona del Guadalhorce.'
   → 'Rodríguez Leal admitted, nevertheless, that there had been some occasional problems in the area of Guadalhorce, specially in agriculture.'
3. Even if drought is mentioned in a very slight manner, only as a secondary cause of a negative impact on agriculture, we should consider it.
   For example:
   → 'La superproducción arruinó la pasada campaña citrícola *y* en la que ahora comienza es la sequía la que amenaza la calidad de la fruta.'
   → 'Overproduction ruined the last citrus fruit season, and for the upcoming one it is now drought that threatens the quality of the fruit.'
4. If there is a sentence that mentions drought impacts but does not clearly tell us that it has had its effects on crops, we must avoid it — even if, by context, we are somehow sure it could be related to agriculture. We have to be completely sure that the decontextualized sentence is, indeed, related to agriculture.
   For instance, a sentence of this type is to be omitted:
   → 'En este caso, al problema de la sequía se han sumado los graves daños por las heladas en toda la mitad norte *y* en parte del centro.'
   → 'In this case, other than the issue of droughts we have also suffered severe damages due to cold spells in many places of the country.'
5. Water irrigation issues or plans regarding crops are ambiguous, so unless these refer to plans currently in place to alleviate an ongoing drought, they are to be omitted.
   For instance:
   → 'A pesar de ese riesgo, en los últimos años se han producido avances discretos en la utilización de sistemas de riego más eficientes.'
   → 'Despite that risk, in these last years there have been slight advancements towards the use of more efficient irrigation systems.'
   Even then, careful consideration has to be given to the fact of whether the sentence could actually be part of the Hydrology label. We must remember at this point that, if a sentence belongs to more than one label, it must be discarded.

6. If crop production levels have dropped, but this not in direct correlation with an ongoing drought, this sentence is to be omitted as well.
   For example:
   → 'La compañía que dirige ha vendido unas 15000 toneladas de esta fruta, casi la mitad de la producción nacional, unas 32000 toneladas.'
   → 'The company he/she leads has distributed 15,000 tons of this fruit, almost half of that of the national production, which is at around 32,000 tons.'

• Livestock:

1. As with Agriculture, in this category we will find a series of sentences in which we can tell that cattle or livestock facilities have undergone damages or severe circumstances due to a drought. An ongoing drought must be explicitly mentioned as the cause behind the event in order for a sentence to be considered.
2. Overall, some of the same indications for the Agriculture label mentioned above also apply here. These include, for instance, the fact that a drop in overall livestock production – that is, if it is not explained as a consequence of a drought – will not be considered a candidate for our positives dataset.

• Hydrological resources:

1. All sentences belonging to this category refer, mainly, to water reservoirs or dams whose levels of stored water have been affected by drought.
   For example:
   → 'Según el último informe, los pantanos de la cuenca del Segura están al 55,6% de su capacidad *y* los del Júcar, al 52,2%, por debajo del 66,8% de media española.'
   → 'According to the last report, water reservoirs in the Segura basin are at a 55.6% of its overall capacity, and those of the Júcar basin are at a 52,2%, below that of the usual 66.8% of the rest of Spain.'
2. Sentences that refer to a more restricted availability of either surface or underground waters – as long as they are explicitly mentioned to be caused by drought – are also to be considered.
   For example:
   → 'La cuenca del Segura se encuentra sólo al 18% de su capacidad *y* sólo tiene el agua del Tajo garantizada (fundamental para dar de beber a la cuenca) hasta julio.'
   → Translation: 'The Segura basin has only 18% of its usual water capacity and the water flow originating from the river Tajo – which is an essential source for this river – is only guaranteed to be present until July'
3. Planned construction or maintenance work on dams or reservoirs, even if they are mentioned to be caused by drought, are not to be part of this category.
   For instance:
   → 'El consejero avanzó que la próxima semana comenzarán las reuniones para coordinar las obras de emergencias con motivo de la sequía.'
   → Translation: 'The adviser revealed that the meetings for putting forward some emergency works for alleviating the current drought will begin next week'
4. Water supply cuts issued as part of ordinances or decrees due to droughts are also not to be part of this category, since it is a different drought impact altogether.
   For example:
   → 'El decreto de la Comunidad que ha prohibido durante un año regar los parques *y* jardines no históricos se derogó el pasado 10 de junio.'
   → Translation: 'The state ordinance that forbade the watering of non-historical parks and gardens for a year is out of effect since June 10th'
5. Consumption levels of water being reduced due to drought are not to be part of this label.
   This includes examples such as the following:
   → 'Del volumen total del consumo, el 60% corresponde a la agricultura, el 24% al consumo urbano *y* el 35% restante al industrial entre otros.'
   → 'Out off the total volume of water consumption, around 60% of it comes from agriculture, 24% from urban consumption, and a remaining 35% comes from industrial uses.'

Table C.7
Manually selected keywords used to preliminary detect a series of articles on drought impacts.

Agriculture	Livestock	Hydrological r.	Energy
'agricultores'	'ganado'	'río'	'hidroeléctrica'
'regantes'	'ganadería'	'embalse'	'electricidad'
'campo'	'cabezas'	'cuenca'	'energía'
'cultivos'	'ayudas'	'confederación'	'turbinas'
'cosechas'	'ganaderos'	'reservas'	–
'campañas'	–	'caudal'	–
'producción'	–	–	–
'campaña'	–	–	–
'regadío'	–	–	–

• Energy sector:

1. The reviewers should retrieve sentences that contain references to hydroelectrical power production being affected by a current drought happening.
2. Any sentence referring hydroelectrical power production in some way, but not mentioning droughts is to be omitted.

Appendix D. Examples of misclassified sentences by the drought impact identification module

While the results presented for the experiments on the end-to-end test dataset from Section 6.2 were promising, some of the labels presented a number of misclassified instances — either false positives or false negatives. We present on this Appendix some examples of sentences that are false positives for each type of drought impact, as well as some explanations behind the observed results.

• Agriculture:

  – Sentences dealing with drought, but not specifically agriculture:

    * 'La Junta andaluza invertirá 2.368 millones en obras para paliar la sequía'

      · 'The Andalusian Government will invest 2368 million in works to alleviate the drought'

    * 'No obstante, el caso que más preocupa al momento es el de Baja California, pues casi el 100% del territorio se encuentra bajo condiciones de sequía; precisamente el 99,7%.'

      · 'However, the case that is most worrying at the moment is that of Baja California, since almost 100% of the territory is under drought conditions; precisely 99.7%.'

- Sentences dealing with agriculture, but not necessarily drought:
  - ‘Las lluvias golpean, por lo general, en verano, próximo a la cosecha, cuando el productor ya puso toda la plata *y* necesita recuperar la inversión, dice Mántaras.’
    - ‘The rains generally hit in the summer, close to the harvest, when the producer has already invested all the money and needs to recover the investment, says Mántaras.’

- Livestock:
  - References to animals, but neither cattle nor drought:
    - ‘Los osos han empezado a alimentarse de bellotas muy pronto, desde principios de septiembre, *y* por eso se concentran en zona de robles, como la que ahora está siendo cercada por el fuego.’
      - ‘The bears have begun to feed on acorns very early, since the beginning of September, and that is why they concentrate in areas of oak trees, like the one that is now being surrounded by the fire.’
    - ‘El problema añadido es que los animales autóctonos, como conejos *y* pájaros, se han quedado sin alimentos *y* acuden a los viñedos a comerse las uvas.’
      - ‘The added problem is that native animals, such as rabbits and birds, have run out of food and go to the vineyards to eat the grapes.’
  - Cattle, but not drought:
    - ‘La moratoria – que fue rechazada por el Ejecutivo *y* por el Parlamento –, también prohíbe a los suizos criar animales transgénicos en sus granjas.’
      - ‘The moratorium – which was rejected by the Executive and Parliament – also prohibits the Swiss from raising transgenic animals on their farms.’

- Hydrological resources:
  - Weather or climate change being referenced instead of available hydrological resources:
    - ‘La temperatura media mundial en 2017 fue 0,46 grados superior al promedio del periodo 1981–2010.’
      - ‘The global average temperature in 2017 was 0.46 degrees higher than the average for the period 1981–2010.’
    - ‘Desde hace 650.000 años no había una concentración similar de CO2 en la atmósfera, según estos expertos.’
      - ‘There has not been a similar concentration of CO2 in the atmosphere since 650,000 years ago, according to these experts.’
  - Water bodies being referenced, but not talking about lack of hydrological resources:
    - ‘Si se compara con los 60 metros cúbicos por segundo de aforo más abundante del río Júcar, nos damos cuenta de lo que es un aluvión, dijo.’
      - ‘If it is compared to the 60 cubic meters per second of the most abundant capacity of the Júcar River, we realize what an alluvium is, he said.’
    - ‘Los índices de anhídrido sulfuroso *y* partículas en suspensión por metro cúbico se situaban la semana pasada por debajo del 50%.’
      - ‘The levels of sulfur dioxide and suspended particles per cubic meter were below 50% last week.’

- Energy:
  - Energy production, but not related directly to drought:
    - ‘La presa se construyó en 1962 para la producción de energía eléctrica *y* solo evacua agua por las compuertas altas cuando hay crecidas.’
      - ‘The dam was built in 1962 for the production of electrical energy and only evacuates water through the high gates when there are floods.’
  - Climate change due to energy production:
    - ‘El organismo, creado por la ONU, concluye además que el calentamiento se debe, con un 90% de certeza, a la actividad humana, en especial por el uso masivo de energía basada en combustibles fósiles.’
      - ‘The organization, created by the UN, also concludes that warming is due, with 90% certainty, to human activity, especially the massive use of energy based on fossil fuels.’
    - ‘Por otro, aunque España es líder en energías renovables, las emisiones de efecto invernadero no dejan de subir.’
      - ‘On the other hand, although Spain is a leader in renewable energy, greenhouse emissions continue to rise.’

Appendix E. Listing of heuristic syntactic rules used for location names disambiguation

The syntactic heuristics used for the module in Section 3.5 are relatively simple, and are based on the analysis of the dependency parse trees of each of the sentences. These are listed below:

- Location name is preceded by a keyword. A check is performed to see if, for an ambiguous location name, its head in a dependency tree happens to be any of the following keywords: ‘río’ (‘river’), ‘cuenca’ (‘basin’), ‘presa’ (‘dam’), ‘embalse’ (‘reservoir’) and ‘provincia’ (‘province’). If that is the case, then it is assumed that the location type matches that of the keyword.
- Location name is part of an enumeration alongside other location names, and they are all headed by a keyword in plural form. An example would be the following sentence:
  ‘La Subdelegación del Gobierno llegó a negar que las obras estuvieran afectadas por la anulación del decreto que pretendía corregir la sequía en **las cuencas del Guadiana, Guadalquivir** *y* **Ebro**"
  (‘The government subdelegation came as far as claiming that the works had not been affected by the drought alleviation ordinance

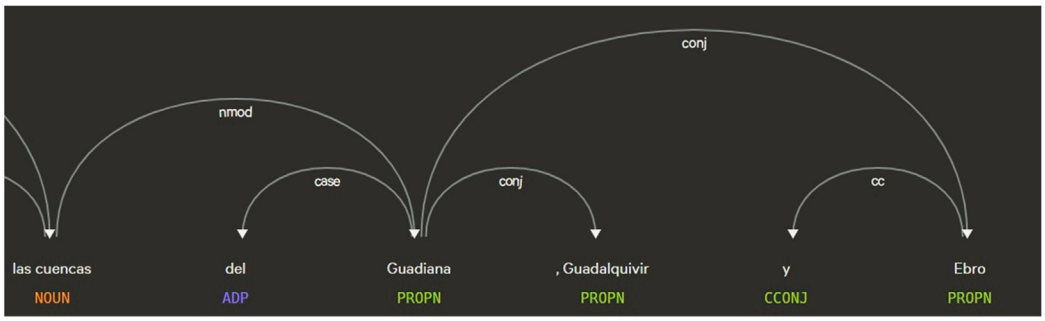

Fig. E.8. A visualization of a dependency parse tree for a sample noun phrase that consists of a keyword in plural form followed by a set of adjacent location names in an enumeration, where these ambiguous names' type can be determined thanks to their syntactic head.

put in place for **the river basins of Guadiana, Guadalquivir and Ebro**")

For this sentence, we are interested in the noun phrase marked in bold above, consisting on the enumeration of the proper names 'Guadiana', 'Guadalquivir' and 'Ebro' headed by the keyword 'cuencas' ('basins'). The dependency parse tree visualization for this noun phrase can be found in Fig. E.8. For this type of NP, some proper names may have a direct link with a keyword ('basins', in this case), making it straightforward to disambiguate them. Others, however, might be linked instead to other location names in the list; nevertheless, if they are either directly or indirectly linked to a proper noun whose head happens to be a keyword, then our method assumes that they share the same location type as that keyword.

- In absence of a keyword, for a location name shared by a town/city and a river, we perform a disambiguation technique based on the fact whether the name happens to be preceded by a determiner article. This follows a rule specific to the Spanish language: when native speakers are referring to an ambiguous location name as that of a river, they always include a determiner article in front of the toponym to mark it as such. This article is either standalone (such as 'el') or contracted with another form such as a preposition (such as in 'del' or 'al').
- Another heuristic rule consists on the disambiguation of location names shared by cities and provinces, such as that of 'Zaragoza'. In most Spanish texts, province names are usually enclosed by parentheses, and as such a simple test is performed to see if the named location in this case is surrounded by these marks, in which case it is considered as a province.

Finally, we also included a manual list of exceptions to our list for manually disambiguating some problematic entries that we detected during the testing process of building this system.

Data availability

The training datasets and model weights are available in open access in a link listed in the software and data availability statement in our paper. Our framework is accessible from our GitHub page.

References

Akbari, P., Gabriel, M., MacKenzie, C.A., 2022. Retrieving and disseminating information about disasters through natural language processing tools. In: IIE Annual Conference. Proceedings. pp. 1–6.

Bell, S., 2009. The driest continent and the greediest water company: newspaper reporting of drought in Sydney and London. Int. J. Environ. Studies 66 (5), 581–589. http://dx.doi.org/10.1080/00207230903239220.

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. http://dx.doi.org/10.48550/arXiv.2004.05150.

Dahir, A., Omar, M., Abukar, Y., 2023. Internet of things based agricultural drought detection system: case study Southern Somalia. Bull. Electr. Eng. Inform. 12 (1), 69–74.

Dayrell, C., Svensson, C., Hannaford, J., McEnery, T., Barker, L.J., Baker, H., Tanguy, M., 2022. Representation of drought events in the United Kingdom: Contrasting 200 years of news texts and rainfall records. Front. Environ. Sci. 10, 760147. http://dx.doi.org/10.3389/fenvs.2022.760147.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. In: Long and Short Papers, vol. 1, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Domala, J., Dogra, M., Masrani, V., Fernandes, D., D'souza, K., Fernandes, D., Carvalho, T., 2020. Automated identification of disaster news for crisis management using machine learning and natural language processing. In: 2020 International Conference on Electronics and Sustainable Communication Systems. ICESC, pp. 503–508. http://dx.doi.org/10.1109/ICESC48915.2020.9156031.

Domínguez-Castro, F., Vicente-Serrano, S.M., Tomás-Burguera, M., Peña-Gallardo, M., Beguería, S., El Kenawy, A., Luna, Y., Morata, A., 2019. High spatial resolution climatology of drought events for Spain: 1961–2014. Int. J. Climatol. 39 (13), 5046–5062. http://dx.doi.org/10.1002/joc.6126.

Dow, K., 2010. News coverage of drought impacts and vulnerability in the US Carolinas, 1998–2007. Nat. Hazards 54, 497–518. http://dx.doi.org/10.1007/s11069-009-9482-0.

Duarte, S., Corzo Perez, G.A., Santos, G., Solomatine, D.P., 2024. Application of natural language processing to identify extreme hydrometeorological events in digital news media: Case of the Magdalena river basin, Colombia. Adv. Hydroinformatics: Mach. Learn. Optim. Water Resour. 283–318. http://dx.doi.org/10.1002/9781119639268.

Eva, J., Arlene, C., Natascha, S., Therese, M., Laura, S., Conor, M., Robert, M., Francis, L., Csaba, H., 2022. Irish Drought Impacts Database v.1.0 (IDID) [dataset]. Zenodo, http://dx.doi.org/10.5281/zenodo.7216126.

Fandiño, A.G., Estapé, J.A., Pàmies, M., Palao, J.L., Ocampo, J.S., Carrino, C.P., Oller, C.A., Penagos, C.R., Agirre, A.G., Villegas, M., 2022. MarIA: Spanish language models. Proces. Del Leng. Nat. 68, 39–60. http://dx.doi.org/10.26342/2022-68-3.

Garcia-Garizabal, I., Causapé, J., Abrahão, R., Merchán, D., 2014. Impact of climate change on mediterranean irrigation demand: Historical dynamics of climate and future projections. Water Resour. Manag. 28, 1449–1462. http://dx.doi.org/10.1007/s11269-014-0565-7.

Geonames, 2022. Geonames. https://www.geonames.org/. (Accessed 14 March 2024).

Gonzalez-Hidalgo, J., Vicente-Serrano, S., Peña-Angulo, D., Salinas, C., Tomas-Burguera, M., Beguería, S., 2018. High-resolution spatio-temporal analyses of drought episodes in the western Mediterranean basin (Spanish mainland, Iberian Peninsula). Acta Geophys. 66, 1–12. http://dx.doi.org/10.1007/s11600-018-0138-x.

Gopal, L., Prabha, R., Vinodini Ramesh, M., 2023. Developing information extraction system for disaster impact factor retrieval from web news data. pp. 357–365. http://dx.doi.org/10.1007/978-981-19-0098-3_35.

del Guadiana, C.H., 2018. Plan especial de sequía. Demarcación Hidrográfica del Guadiana. https://www.chguadiana.es/sites/default/files/2018-12/PESCHGn.pdf. (Accessed 24 January 2025).

Hoang, V.P., Nguyen, M.H., Do, T.Q., Le, D.N., Bui, D.D., 2020. A long range, energy efficient Internet of Things based drought monitoring system. Int. J. Electr. Comput. Eng. 10 (2), 1278–1287.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., 2020. spaCy: Industrial-strength natural language processing in python [software]. http://dx.doi.org/10.5281/zenodo.1212303.

Hurlimann, A., Dolnicar, S., 2012. Newspaper coverage of water issues in Australia. Water Res. 46 (19), 6497–6507. http://dx.doi.org/10.1016/j.watres.2012.09.028.

Instituto Geográfico Nacional, 2024. Hidrografía de españa [dataset]. https://centrodedescargas.cnig.es/CentroDescargas/documentos/atom/au/lineas_limite_gml.zip. (Accessed 14 March 2024).

Lackstrom, K., Farris, A., Ward, R., 2022. Backyard hydroclimatology: citizen scientists contribute to drought detection and monitoring. Bull. Am. Meteorol. Soc. 103 (10), E2222–E2245.

Lai, K., Porter, J.R., Amodeo, M., Miller, D., Marston, M., Armal, S., 2022. A natural language processing approach to understanding context in the extraction and GeoCoding of historical floods, storms, and adaptation measures. Inf. Process. Manage. 59 (1), 102735. http://dx.doi.org/10.1016/j.ipm.2021.102735.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174, URL: http://www.jstor.org/stable/2529310.

Liu, X., Guo, H., Lin, Y.r., Li, Y., Hou, J., 2018. Analyzing spatial-temporal distribution of natural hazards in China by mining news sources. Nat. Hazards Rev. 19, 04018006. http://dx.doi.org/10.1061/(ASCE)NH.1527-6996.0000291.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692.

Llamas, M., 2000. Some lessons learnt during the drought of 1991–1995 in Spain. In: Drought and Drought Mitigation in Europe. Springer, pp. 253–264.

Llasat, M.C., Llasat-Botija, M., Barnolas, M., López, L., Altava-Ortiz, V., 2009. An analysis of the evolution of hydrometeorological extremes in newspapers: the case of Catalonia, 1982–2006. Nat. Hazards Earth Syst. Sci. 9 (4), 1201–1212. http://dx.doi.org/10.5194/nhess-9-1201-2009.

Madruga de Brito, M., Sodoge, J., Kreibich, H., Kuhlicke, C., 2025. Comprehensive assessment of flood socioeconomic impacts through text-mining. Water Resour. Res. 61 (1), http://dx.doi.org/10.1029/2024WR037813, e2024WR037813.

Ministerio para la Transición Ecológica, 2024. Ríos completos clasificados según pfafstetter modificado [dataset]. https://wms.mapama.gob.es/sig/Agua/RiosCompPfafs/wms.aspx?. (Accessed 14 May 2024).

Mukherjee, S., Wang, S., Hirschfeld, D., Lisonbee, J., Gillies, R., 2022. Feasibility of adding Twitter data to aid drought depiction: Case study in colorado. Water 14 (18), http://dx.doi.org/10.3390/w14182773.

Musaev, A., Stowers, K., Kam, J., 2018. Harnessing data to create an effective drought management system. In: ISCRAM 2018 Conference Proceedings – 15th International Conference on Information Systems for Crisis Response and Management. pp. 544–552, URL: https://idl.iscram.org/files/aibekmusaev/2018/2130_AibekMusaev_etal2018.pdf.

Nobre, C., Marengo, J., Seluchi, M., Cuartas, L., Alves, L., 2016. Some characteristics and impacts of the drought and water crisis in southeastern Brazil during 2014 and 2015. J. Water Resour. Prot. 08, 252–262. http://dx.doi.org/10.4236/jwarp.2016.82022.

O'Connor, P., Murphy, C., Matthews, T., Wilby, R.L., 2023. Relating drought indices to impacts reported in newspaper articles. Int. J. Climatol. 43 (4), 1796–1816. http://dx.doi.org/10.1002/joc.7946.

Otudi, H., Gupta, S., Albarakati, N., Obradovic, Z., 2024. Classifying severe weather events by utilizing social sensor data and social network analysis. In: Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM '23, Association for Computing Machinery, New York, NY, USA, pp. 64–71. http://dx.doi.org/10.1145/3625007.3627298, URL: https://doi.org/10.1145/3625007.3627298.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2018. Scikit-learn: Machine learning in python. http://dx.doi.org/10.48550/arXiv.1201.0490.

Peña-Gallardo, M., Vicente-Serrano, S.M., Domínguez-Castro, F., Beguería, S., 2019. The impact of drought on the productivity of two rainfed crops in Spain. Nat. Hazards Earth Syst. Sci. 19 (6), 1215–1234. http://dx.doi.org/10.5194/nhess-19-1215-2019.

Pita Costa, J., Rei, L., Bezak, N., Mikoš, M., Massri, M.B., Novalija, I., Leban, G., 2024. Towards improved knowledge about water-related extremes based on news media information captured using artificial intelligence. Int. J. Disaster Risk Reduct. 100, 104172. http://dx.doi.org/10.1016/j.ijdrr.2023.104172.

Ruiz Sinoga, J.D., León Gross, T., 2013. Droughts and their social perception in the mass media (southern Spain). Int. J. Climatol. 33 (3), 709–724. http://dx.doi.org/10.1002/joc.3465.

Rutledge-Prior, S., Beggs, R., 2021. Of droughts and fleeting rains: Drought, agriculture and media discourse in Australia†. Aust. J. Polit. Hist. 67 (1), 106–129. http://dx.doi.org/10.1111/ajph.12759.

Sáez, J.A., Krawczyk, B., Woźniak, M., 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognit. 57, 164–178. http://dx.doi.org/10.1016/j.patcog.2016.03.012, URL: https://www.sciencedirect.com/science/article/pii/S0031320316001072.

Shabbeer, S.A., Srijan, T.R.P., Yewale, R., Karande, S.C., 2016. Drought detection using internet of things. Int. Res. J. Eng. Technol. (IRJET).

Sodoge, J., Kuhlicke, C., de Brito, M.M., 2023. Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. Weather. Clim. Extrem. 41, 100574. http://dx.doi.org/10.1016/j.wace.2023.100574.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J.H., Tallaksen, L.M., Kampragou, E., Van Loon, A.F., Barker, L.J., Melsen, L.A., Bifulco, C., Musolino, D., de Carli, A., Massarutto, A., Assimacopoulos, D., Van Lanen, H.A.J., 2016. Impacts of European drought events: insights from an international database of text-based reports. Nat. Hazards Earth Syst. Sci. 16 (3), 801–819. http://dx.doi.org/10.5194/nhess-16-801-2016.

Stahl, K., Szillat, K., Blahova, M., Blauhut, V., Rossi, L., Masante, D., Maetens, W., Toreti, A., 2023. European Drought Impact Database. European Drought Observatory for Resilience and Adaptation, URL: https://drought.emergency.copernicus.eu/tumbo/edid.

Svoboda, M.D., Fuchs, B.A., et al., 2016. Handbook of drought indicators and indices, vol. 2, World Meteorological Organization Geneva, Switzerland.

Tounsi, A., Temimi, M., 2023. A systematic review of natural language processing applications for hydrometeorological hazards assessment. Nat. Hazards 116 (3), 2819–2870. http://dx.doi.org/10.1007/s11069-023-05842-0.

Trullenque-Blanco, V., Beguería, S., Vicente-Serrano, S.M., Peña-Angulo, D., González-Hidalgo, C., 2024. Catalogue of drought events in peninsular Spanish along 1916–2020 period. Sci. Data 11 (1), 703.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates, Inc., pp. 6000–6010, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vicente-Serrano, S.M., McVicar, T.R., Miralles, D.G., Yang, Y., Tomas-Burguera, M., 2020. Unraveling the influence of atmospheric evaporative demand on drought and its response to climate change. WIREs Clim. Chang. 11 (2), e632. http://dx.doi.org/10.1002/wcc.632.

Vicente-Serrano, S.M., Peña-Angulo, D., Beguería, S., Domínguez-Castro, F., Tomás-Burguera, M., Noguera, I., Gimeno-Sotelo, L., El Kenawy, A., 2022. Global drought trends and future projections. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 380 (2238), 20210285. http://dx.doi.org/10.1098/rsta.2021.0285.

van Vliet, M.T.H., Sheffield, J., Wiberg, D., Wood, E.F., 2016. Impacts of recent drought and warm years on water resources and electricity supply worldwide. Environ. Res. Lett. 11 (12), 124021. http://dx.doi.org/10.1088/1748-9326/11/12/124021.

Wilhite, D., 2000. Drought as a natural hazard: Concepts and definitions. In: Drought: A Global Assessment, vol. 1, Routledge, pp. 3–18.

WMO, 2021. WMO atlas of mortality and economic losses from weather, climate and water extremes (1970–2019). Technical Report WMO-No. 1267, WMO, Geneva, URL: https://library.wmo.int/idurl/4/57564.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, pp. 38–45. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.

Yzaguirre, A., Warren, R., Smit, M., 2015. Detecting environmental disasters in digital news archives. In: 2015 IEEE International Conference on Big Data (Big Data). pp. 2027–2035. http://dx.doi.org/10.1109/BigData.2015.7363984.

Zhang, B., Schilder, F., Smith, K.H., Hayes, M.J., Harms, S., Tadesse, T., 2022. TweetDrought: A deep-learning drought impacts recognizer based on Twitter data. In: ICML 2021 Workshop: Tackling Climate Change with Machine Learning. URL: https://www.climatechange.ai/papers/icml2021/32/paper.pdf.

Zhao, X., Huang, G., Li, Y., Lu, C., 2023. Responses of hydroelectricity generation to streamflow drought under climate change. Renew. Sustain. Energy Rev. 174, 113141. http://dx.doi.org/10.1016/j.rser.2022.113141.

Zhou, Y., Liu, P., 2024. Assessing multi-hazards related to tropical cyclones through large language models and geospatial approaches. Environ. Res. Lett. 19 (12), 124069.

Zou, L., He, Z., Zhou, C., Zhu, W., 2024. Multi-class multi-label classification of social media texts for typhoon damage assessment: a two-stage model fully integrating the outputs of the hidden layers of BERT. Int. J. Digit. Earth 17 (1), 2348668.