

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Duarte, S., Corzo Perez, G. A., Santos, G., & Solomatine, D. P. (2023). Application of Natural Language Processing to Identify Extreme Hydrometeorological Events in Digital News Media: Case of the Magdalena River Basin, Colombia. In *Advanced Hydroinformatics: Machine Learning and Optimization for Water Resources* (pp. 285-318). Wiley. <https://doi.org/10.1002/9781119639268.ch10>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

APPLICATION OF NATURAL LANGUAGE PROCESSING TO IDENTIFY EXTREME HYDROMETEOROLOGICAL EVENTS IN DIGITAL NEWS MEDIA: CASE OF THE MAGDALENA RIVER BASIN, COLOMBIA

**Santiago Duarte^{1,2}, Gerald A. Corzo Perez¹, Germán Santos³, and
Dimitri P. Solomatine^{1,2,4}**

¹IHE Delft Institute for Water Education, Delft, The Netherlands

²Delft University of Technology, Delft, The Netherlands

³Colombian School of Engineering Julio Garavito, Bogotá, Colombia

⁴Water Problems Institute of the Russian Academy of Sciences, Moscow, Russia

Human behavior and decision making are dynamically influenced by digital media, making digital news a vast data source of events and points of view. Meanwhile, artificial intelligence now has advanced capacity for natural language processing (NLP) through tools such as sentiment analysis and topic identification. This research uses machine learning algorithms to prove the hypothesis that it is possible to find correlations between news information and water resource problems. The 207 water bodies in the Magdalena River Basin, Colombia, were analyzed alongside 19,490 news articles published between 2016 and 2020 in 42 online newspapers. A platform for the visualization of spatiotemporal information was developed as a proof of concept. There has been a noticeable increase in digital

news about the Magdalena River Basin in recent years and some correlation with extreme events, but the comparison between the sentiments and measured physical variables was inconclusive. This is likely due to the difficulty in filtering the topic from obtained news, as well as in precisely identifying the spatial location and temporal range of events. In the future, NLP techniques to analyze news articles could be used to provide additional information for decision making at the basin level.

10.1. Introduction

Integrated water resources management (IWRM) aims at making informed decisions with the involvement and information from all stakeholders. Over the last decade, digital media has implicitly increased data-based information on most societal activities carried out by humans (Nielsen, 2015). Water resources, extreme conditions, and hydrometeorological events are not exceptions. For example, flood-related digital media information has been collected in the last decades, from tweets (Murthy & Longwell, 2013) to social media (Rossi et al., 2018), and used to provide information toward a smart flood-risk management (L. Smith et al., 2017) and improved policies (A. Smith et al., 2017). However, challenges posed by conflicting interests are not easily solved (Giupponi & Gain, 2017). The problem of decision making in water management using IWRM support from stakeholders is based on several criteria. While multiple actors participate in hydrology, the possibilities to involve citizens' science in hydrology has increase, as presented detailed in Nardi et al. (2020) where a framework to address citizens' science and hydrology problems are approached. It would seem that the relation between the information provided by citizens and extreme events is obvious, but in practice, the technologies require integration, as well as validation of this relation. Citizens' data collected from local observations to crowdsourcing (Le Coz et al., 2016), measured directly or not (Kam et al., 2019), it's nowadays a new database in decision-making process. In this context, the decision-making process in IWRM is based on history, documentation, and key elements of past and current decisions (Pires et al., 2017). In Avellaneda et al. (2020), crowdsourcing information regarding volunteers' calibration data improved hydrological model results. In the same way, analysis of media (digital and analogical) regarding historical hydrological disaster event that affected the Paraná coastal region in Brazil concluded in an improved understanding and characterization of floods and debris flows (Picanço et al., 2017).

One of the main challenges in decision making is the fact that human behavior and decisions are influenced by digital media (Keegan & Rowley, 2017). In general, decision-making results are presented as all kinds of policies and human interventions around water bodies, from forecasting estimations and emergency alerts, to nature-based solutions or artificial structures (Watts, 2016). Moreover, the implications of these actions on digital media behavior is starting to be used to validate decision-making results. Yu et al. (2019) discussed the applications of social media in hydroinformatics, including detection of fake news, analysis of floods management, and optimization of water resource planning, among others. Further, relations between digital media (e.g., tweets) and disaster severity could improve risk planning management when distributing resources in an emergency (Kankanamge et al., 2020). Thus, it is necessary to quantify the short-term and the long-term effects of these decisions, while measuring and looking for physical variables, social equity, environmental impacts, and so on (Badham et al., 2019). Consequently, new tools have been developed and used to study decision-making connections with digital media. For example, Bai et al. (2020) applied a machine learning technique (ML-KNN) to evaluate flood disaster risk based on social media (Weibo).

In this research, we address a relatively new paradigm in relating digital information (digital news) to historical decision making about water resources. We do so by analyzing digital citizens' media information across space and time. To understand these relations in other areas of science, artificial intelligence (AI) has developed some branches of data analysis in the semantics of language and how they can be integrated and quantified, the so-called natural language processing (NLP). This area explores AI tools to understand how natural language can be used for analyzing data (Chowdhury, 2003). Most of the advances focus on machine translation, speech recognition, and speech synthesis (Hirschberg & Manning, 2015). NLP has also been used to build concepts, such as sentiment analysis (Bello, 2020; Cambria et al., 2017). Sentiment analysis is a technique that provides a feeling-like index that relates to the semantics of positive and negative emotions, combined with AI. NLP learning algorithms allow providing a real-time analysis of data sets. The work by Mishev et al. (2020) shows some examples of how this has been implemented in finance problems. Garvey and Maskal (2020) use the NLP Google tool to quantify positive or negative effects on the news about AI. In general, research in this area has shown an important potential. All these advances motivate researchers to study the relation between digital media and water resources management. In this context, in a country such as Colombia, where most of its Gross Domestic Product (GDP) depends on water resources, decisions regarding

the Magdalena River Basin are thus quite important, especially in an economic recovery scenario after COVID-19 (Borja-Vega et al., 2020). Moreover, digital media may help identify social patterns related to the environment and water resources. In this research, we evaluate particular cases (e.g., Hidroituango) using NLP, from which results are obtained and conclusions are drawn. This will also be relevant for future researchers who can analyze the effects of decision-making processes as proposed solutions to specific water-related problems and their consequences on human behavior.

10.2. Case Study

The description of this case study is divided into water resources and a digital socioeconomic perspective. The description of the water resources covers all important and general hydrometeorological information about the river basin in the Colombian context. The socioeconomic and digital perspective covers the information about not only the importance of the river basin but also the characteristics of this digital information in terms of quantity, quality, and availability.

The Magdalena River Basin is located in Colombia, South America, as shown in Figure 10.1. The basin extension is 273,459 km², and the main river length is 1,558 km. The average discharge flow is 7,293 m³/s. The annual mean precipitation is 2,050 mm, with wet seasons between March to May and September to November, and dry seasons between December to February and June to August.

In total, this macro basin integrates 724 municipalities of 19 departments of the country, which produce 83% of Colombia's GDP (USD 260 billion in 2018), 70% of its hydropower, and 70% of its agricultural production. In recent years, this basin has been highly subjected to human interventions, involving a wide variety of stakeholders (public and private). Adding the capacities of the projects in operation and under construction, it has an installed capacity of around 9.97 GW, distributed among 35 large- and medium-scale projects, and approximately 0.5 GW in minor hydraulic generation projects. This research has been developed as an entry to the competition "Conectate con el Río Magdalena" under the auspices of the Netherlands' Ministry of Foreign Affairs.

The Hidroituango project's location (Cauca River) is also presented in Figure 10.1. This is one of the most important infrastructure projects in Colombia in recent years, with a power capacity of 2,400 MW. In 2018, the collapse of a diversion tunnel due to a geological fault resulted in a major emergency on the Cauca River. This caused an accumulation in the

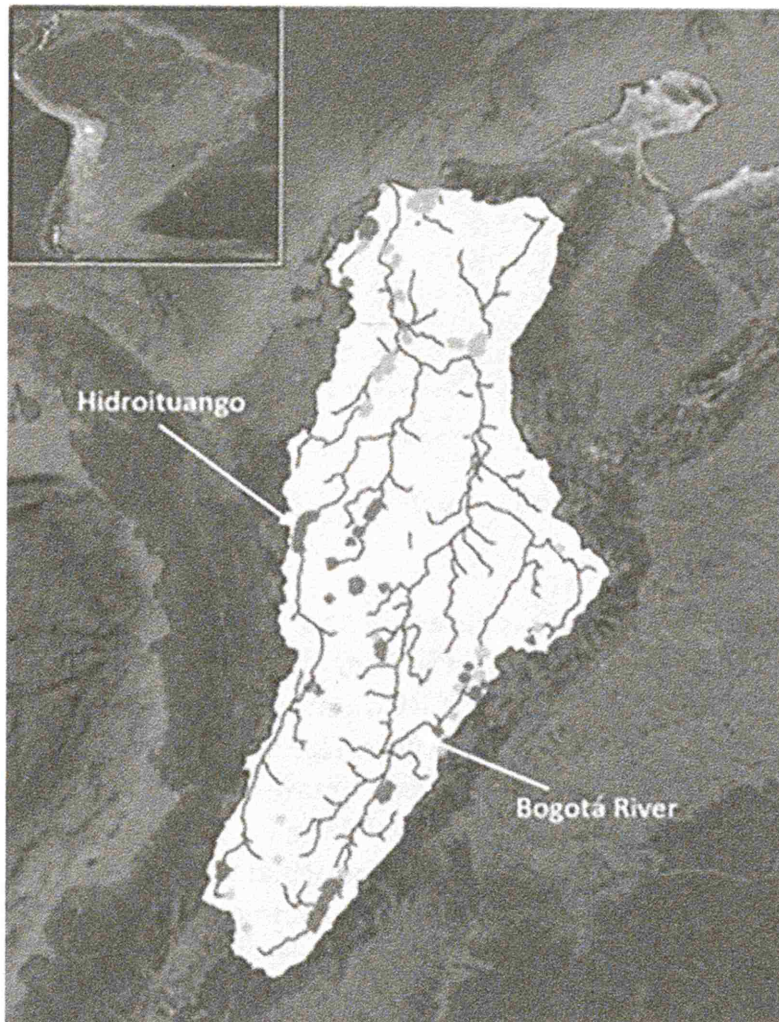


Figure 10.1 Magdalena River Basin scheme: Rivers and water bodies. WGS-84. dams (dark grey), rivers (lines), lakes (white), swamps (light grey).

reservoir and a decrease in the level of the Cauca River. For this reason, the comparison of the sentiment analysis and the measurable physical impact is an objective of our study. Figure 10.2 shows the impact of the emergency.

Three hydroclimatic variables (precipitation, temperature, and river levels) were used in this research. Two river-level stations in situ were selected to evaluate the sentiment index. For both stations, the monthly minimum and maximum river levels were considered for the analysis. The first one was the Puerto Valdivia hydrological station located next to Hidroituango Dam. The average minimum and maximum river-level values were found to be 300.12 and 423.69 cm, respectively. The second station, La Campina in Girardot Municipality, recorded the data from Bogotá River's flows. Both stations presented the results on the flows and the levels (minimum, average, and maximum) at monthly intervals from



Figure 10.2 Journal article shows the emergency (from Tiempo, 2018).

2016 to 2020. The average minimum and maximum river-level values were found to be 200.40 and 493.82 cm, respectively. This information was obtained from the IDEAM (Colombia's Institute of Hydrology, Meteorology and Environmental Studies) database.

Likewise, ERA5 data on the precipitation and the temperature at monthly intervals, with a resolution of $0.25^\circ \times 0.25^\circ$ for the analysis period, were used to compare the results. The information was obtained through Google Earth Engine. The average medium precipitation in the course of the Magdalena River Basin was 233.59 mm for the period, and the average maximum precipitation was 1,805.38 mm. For the same period, the average medium, maximum, and minimum temperatures in the course of the Magdalena River Basin were 23.65°C , 30.83°C , and 9.68°C , respectively.

Among the multiple sources of digital citizens' information (e.g., Facebook, Twitter, news, text messages), only digital newspapers were selected for this research. According to *Prensa* (2015, Dec 17):

The penetration of the press in Colombia is 78% and regional newspapers and the popular press play an important role. Although there appear to be no studies on the time Colombians spend reading the newspaper, recent research revealed that the number of newspaper readers in Colombia reached 6.3 million in 2013 versus 6.1 million in 2012. As of 2015, there are only 4 newspapers with national circulation. There are about 90 different regional newspapers between paid, popular and free press.

Additionally, Perdomo (2022) found:

Colombian media has historically been dominated by publishers and broadcasters with strong ties to the political and business elites. That influence

seems to be transferring to the digital world, but new players and social media are clearly changing consumption trends. “El Tiempo” newspaper website is by far the most used news site (40%), followed by the online versions of “El Espectador” and “Caracol TV” (27%). But some digital native media are growing in importance. (5 Jan, p. 121)

10.3. Methodology

Figure 10.3 shows the main tasks performed to apply, evaluate, and validate the use of NLP in the mapping between water resource problems and digital news. These steps allowed us to extract, detect, and analyze qualitative and quantitative information from newspaper articles related to water bodies.

10.3.1. Step 1 (Digital Citizens’ Search for Information Sources)

In this study, digital newspapers were selected as the sources of digital citizens’ information. In this sense, the digital news information

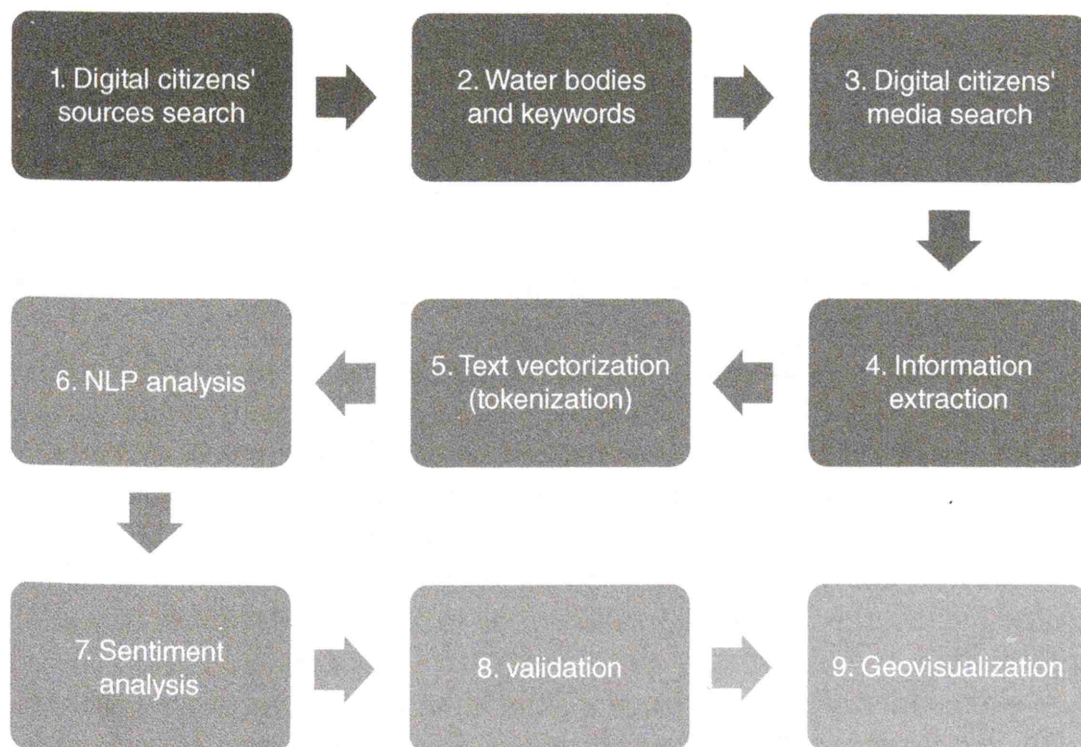


Figure 10.3 Methodology process.

provided text records of water-related events (environmental, extreme climatic events, supply). First, all available open sources of information (digital newspapers) were identified. No journalist blogs, newsletters, or non-water-related newspapers were considered for the analysis. The digital newspaper sources were selected under the following criteria: spatial coverage (regional or national), digital information availability, Web access permission, and number of articles. Nevertheless, because of the exploratory nature of this research, the newspaper selection was limited, and not all available newspapers were used. Only departments in close proximity to the Magdalena River Basin were selected. Two regional newspapers were picked for each department, and four national newspapers were used.

Once the newspaper sources were identified, the uniform resource locator (URL) of each newspaper was obtained. For all selected digital newspapers' websites, the types of available information (articles, photos, videos, and documentaries) were studied. Thus, the news text section was identified for the purpose of obtaining information only from each article's author. No photos or videos were analyzed.

Although the most representative and oldest newspaper in the country has provided digital information since 1990 and even earlier, most of this information is poor and mainly related to sociopolitical problems, not to water resources. In contrast, the objective of our research was to test the hypothesis that it would be possible to relate water resource events and associated problems to news information. Therefore, the time range was reduced; the digital news search included only news articles from 1 January 2016 to 31 May 2020.

10.3.2. Step 2 (Water Bodies and Keywords)

Next, the most important water bodies in the basin were identified (rivers, reservoirs, lakes, and swamps). An analysis of the water bodies in the Magdalena River Basin was then carried out. The analyzed characteristics are (1) size (area or length); (2) importance for the region (environmental, economic, and social); and (3) information availability from digital citizens' sources. This analysis allowed us to identify and prioritize the elements. Consequently, this information allowed us to select local sources of digital information.

Understanding that the basin is a fundamental and transversal element in the development of many communities, a filtering process to omit irrelevant or unrelated news was carried out. Then, a set of keywords to be used in the news search process and filtering related to the main events and

Table 10.1 Keywords defined for the news search.

Climate change and climate variability	Seasonal events	Environment	Related human activities
<i>Cambio climático</i> (Climate change)	<i>Sequía</i> (Drought)	<i>Ecosistema</i> (Ecosystem)	<i>Minería</i> (Mining)
El Niño	<i>Verano</i> (Summer)	<i>Vegetación</i> (Vegetation)	<i>Energía</i> (Energy)
La Niña	<i>Inundación</i> (Flood)	<i>Peces</i> (Fish)	<i>Riego</i> (Irrigation)
	<i>Invierno</i> (Winter)	<i>Cuenca</i> (Basin)	<i>Contaminación</i> (Pollution)
	<i>Nivel</i> (Water level)		<i>Navegabilidad</i> (Navigability)

problems in the basin was defined. The keywords are presented in Spanish in Table 10.1 because of the language used in the case study. The aggregation of these words in groups is also included.

Because of the exploratory context of this research, other keywords that might provide additional results were excluded. This was also intended to reduce the possible redundancy of newspaper articles. For this purpose, words such as *desbordamiento* (overflow) might be included in articles related to the word *inundación* (flood) or *invierno* (winter). Consequently, the newspaper articles related to *desbordamiento* were already included in the search results.

10.3.3. Step 3 (Digital Citizens' Media Search)

We adapted the custom search API (application programming interface) of Google Cloud and a custom search engine (CSE) from Google Search Console for the news search process. For this case, only the results from the selected newspapers' websites were obtained. The search process was performed as the combination of a water body, a keyword, and a date range in a process called *Web crawling*, for example: "Río Bogotá" "Inundación" "01/06/2016 to 30/06/2016".

The specifications of these parameters had these objectives: Specifying the full name of the water body allowed us to locate the spatial region of analysis. For example, if the word *rio* (river) was not included, the information obtained would be related to the City of Bogotá, not the Bogotá River Basin.

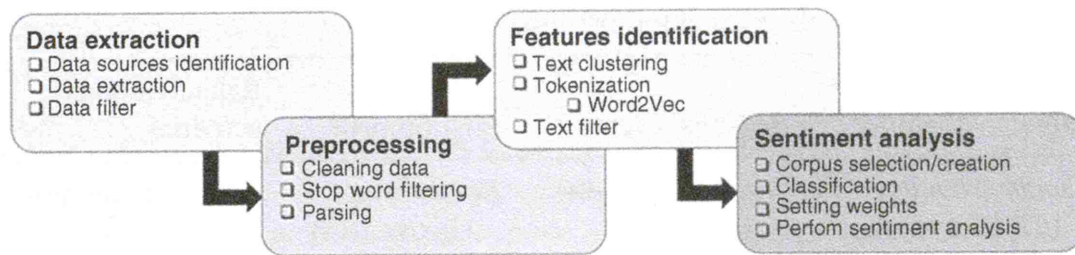


Figure 10.4 Data-mining process.

The keyword filtered information regarding the water resource main events and problems. In this sense, our search excluded any irrelevant or unrelated news, for example, “Soccer match on the banks of the Medellin River.” In this case, while the news might mention the water body, the focus of the main article was related to other social topics. The time period selection allowed us to filter information regarding its digital publishing date for further analysis.

Figure 10.4 summarizes the data mining and analysis process developed in Steps 4 to 7.

10.3.4. Step 4 (Data Extraction and Preprocessing)

As a result of Step 3, for every newspaper article related to each water body, the URL was obtained through the Web-crawling process. The consequent step was to perform Web scraping, based on the parameters of each newspaper website obtained in Step 1. In this process, the information was filtered, downloaded, and processed from an HTML format to a comma-separated values (CSV) file as unstructured data. Python scripts with Googleapiclient, BeautifulSoup, and Panda libraries were used to download, extract, and transform the digital news. These scripts are available in a GitHub repository at <https://github.com/sduarte09/buscandoelmagdalena>.

Next, the unstructured data were transformed into a structured data format, so the text of each article, the title, the date, and the ID of the news websites were extracted. This information was then used to filter the news articles in order to remove duplicates, empty website pages, music/video links, or image galleries, as illustrated in Figure 10.5.

Once the structured data were filtered, the data were cleaned, checking the accent marks, changing uppercase into lowercase letters, and removing advertising in the middle of the text. Likewise, for the implementation of NLP techniques, a stop word filter was needed. In the NLP context,

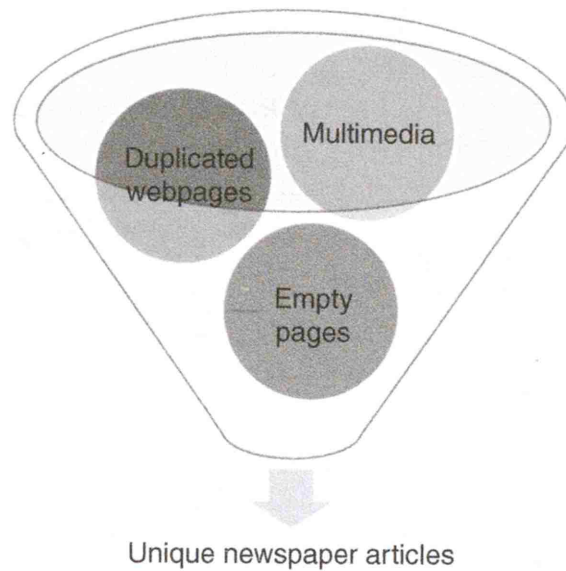


Figure 10.5 News-filtering process.

these are words that do not add much meaning value to a sentence or a document. The bag of stop words can be obtained from the GitHub collective repository at <https://github.com/Alir3z4/stop-words>. Other stop words were added to the bag, following a trial-and-error procedure on the article text, regarding the dialect.

10.3.5. Step 5 (Text Vectorization–Tokenization)

Subsequently, the feature text identification or tokenization started with the word embedding, applying the Word2Vec function, which provides concepts for the word embedding (Goldberg & Levy, 2014). The Word2Vec method identifies text mapping and then creates a model for reproducing such mappings to generate “word embedding” (given a text corpus), which is the process of mapping words to a vector space to realize statistical and probabilistic operations. The most important objective of this method is to preserve the relation between words, this is minimizing the distance between vectors within a similar context. In example, while through a tokenization, it is possible to represent the word *flood* in a text corpus X as $[0,0,0,0,0,0,1,0,0,0,0,0,0]$ and the word *drought* as $[0,0,1,0,0,0,0,0,0,0,0,0,0]$, where the vector size is equal to the vocabulary in the text corpus, these vectors do not show any relation between the meaning of the words *flood* and *drought*, both extreme event opposites within the hydrological cycle.

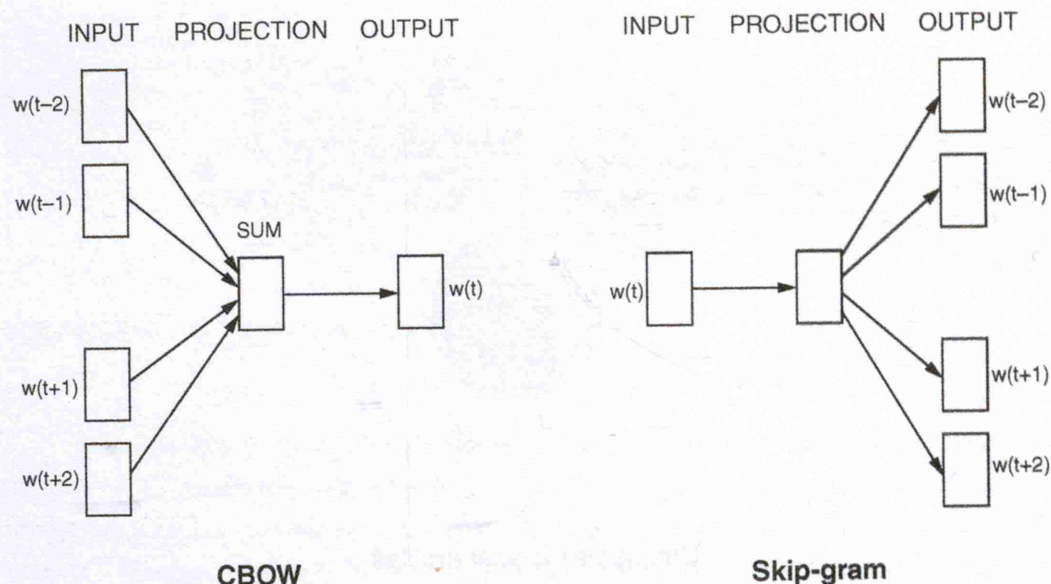


Figure 10.6 Graphical representation of the CBOW model and the Skip-gram model (from Mikolov et al., 2013).

Additionally, large values of data (e.g., big data corpus) could present a challenge to computational calculations (Mikolov et al., 2013). In this sense, the Word2Vec function looks for words that occur in similar contexts to have similar embeddings, that is, represent the meaning of the word, for example:

The flood in the Bogotá River triggered emergency responses from the Colombian government.

The drought in the Bogotá River triggered remedial actions taken by the Colombian government.

The words *flood* and *drought* will have similar word vectors due to a similar context. Two architectures/algorithms are widely used to represent the “meaning” of a word according to a context: CBOW (Continuous Bag of Words) (predict the target word from the bag-of-words context) and Skip-gram (predict the context words, given the target) (Jang et al., 2019). Both implementations use neural networks, as shown in Figure 10.6, where $w(t)$ denotes the analyzed word, and $w(t \pm s)$ represents the word at a distance of s words in the sentence context.

For example:

Sentence: “Several parts in Kuala Terengganu were inundated with flood waters, following continuous rain.”

Distance: -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4

If $w(t) = \text{flood}$, then $w(t+1) = \text{waters}$, $w(t-2) = \text{inundated}$, and $w(t+4) = \text{rain}$. In practice, Skip-gram gives better word representations when the monolingual data are small. However, CBOW is faster and more suitable for larger datasets (Mikolov et al., 2013). In the Skip-gram approach, the objective of the neuronal network is to maximize the probability of any context word, given the current central word, mathematically formulated as follows:

$$F.O. = \frac{1}{T} \sum_{t=1}^T \left(\sum_{\substack{j=-k \\ j \neq 0}}^k \log p(w_{t+j}|w_t) \right),$$

where k denotes the size of the training window (which can be a function of the central word), and $p(w_{t+j}|w_t)$ represents the probability of the word $w(t+j)$, given $w(t)$:

$$p(w_{t+j}|w_t) = \frac{\exp(u_w^T * v_{wj})}{\sum_{n=1}^N \exp(u_n^T * v_{wn})},$$

where N signifies the number of words in the vocabulary.

10.3.6. Step 6 (NLP Analysis)

Once the text had been tokenized and the Word2Vec function had been applied to embed the words in vectors, NLP tools were used to create a word cloud for each water body, so the main words linked to each article could be represented in a graph. This allowed us to obtain a better picture of the most representative elements (social, economic, environmental, and climatic) related to each water body. The Natural Language ToolKit (NLTK), Scipy, and Word Cloud libraries were used in this process.

Counting the number of news items for each time step and each water body, as well as identifying the main topics about each water body, were also part of this study. Through this information, patterns between extreme events and digital news could be explored. Later, this information was aggregated on monthly and annual levels. However, most of the NLP free libraries work only for English-language texts or have a paragraph word-count limitation for non-English languages (because of an implicit third-party translation process). This makes the application of NLP

difficult in non-English regions. Nevertheless, some public packages and libraries allow NLP analysis in Spanish-language texts.

10.3.7. Step 7 (Sentiment Analysis)

An important analysis that could be derived from NLP tools is sentiment analysis. Sentiment analysis shows the perception of an author when writing the article, interpreting this sentiment with a number that ranges between 0 (negative sentiment) and 1 (positive sentiment). For this purpose, a neuronal network is built to classify the writer's sentiment through a numerical value; a summary of the involved processes is shown below. This requires statistical data sampling, data normalization, input variable selection, training, and validation.

Algorithm 10.1 Sentiment analysis pseudo code

Input: $x=[x_1, \dots, x_i]$, $x_t \in \mathbb{R}^m$ and $s=[s_1, \dots, s_n]$, x_t and $s_t \in [0,1]$

Initialize parameters: $w=[w_1, \dots, w_j]$, $w_t \in \mathbb{R}^n$

for $t=1, \dots, n$ do

 Calculate $y=w \bullet x$

 Update w

 Calculate error(y, s)

end for

Output: $s_{n+1} \in [0,1]$

In this research, we used the *spanish_sentiment_analysis* library (Hofman, 2017/2021), which was already trained and tested in its open source version. The library used almost one million reviews of users of the following webpages: *Trip Advisor*, *Pedidos Ya*, *Apestan*, *QuejasOnline*, *Mer-cadoLibre*, *SensaCine*, *OpenCine*, *TASS*, and *Twitter* (Spanish-language version) for its calibration and validation. TensorFlow and Marisa-trie libraries were used for building the sentiment neuronal network. In this process, the model developed a regression between the users' reviews of the websites and their corresponding scores on a scale from 0 to 5 (websites' 5-star reviews). Then, this information was represented as a numerical value between 0 and 1, where a 5-star review represented a sentiment value of 1. This calibrated model was used to predict the news sentiment over the Magdalena River Basin.

The *spanish_sentiment_analysis* package is an open-source library built on scikit-learn and NLTK in Python. The model works as a simple pipeline that includes a vectorizer (TfidfVectorizer), a feature selector (SelectKBest), and a classifier (MultinomialNB). The feature selector will reduce

the number of features generated by the vectorizer. The feature selector function selected was `SelectKBest` from `scikit-learn`, performed by checking whether a feature was relevant or irrelevant. The classifier model used was a multinomial naive Bayes with best parameter $\alpha = 0.03$. The parameters and hyperparameters of this pipeline were found by using a `GridSearchK` cross-validation, with $K = 10$. Additionally, the `Marisa-trie` library was used for a static memory-efficient structure. This model was applied to the digital news information of each article. The results were compiled on monthly and annual aggregation levels. Moreover, the model was tested with reviews of users of the pages *eltenedor*, *decathlon*, *filmaffinity*, and *ebay* in their Spanish language for two data sets (positive and negative) of 50,418 reviews each one using the Mean Absolute Error (MAE).

10.3.8. Step 8 (Validation)

The results obtained from the NLP analysis were compared with measured physical variables (e.g., river-level flows) and observed events (social and environmental emergencies), as part of the validation process. This comparison was necessary to verify the relations between events and digital citizens' media. For this purpose, different parameters were validated, as listed below.

1. Sentiment index versus monthly river levels' range next to the Hidroituango Dam and the Bogotá River (Girardot). The monthly river-level range was used as a simple representation of a sudden change in the flow of the river (e.g., Hidroituango emergency).
2. Correlation of monthly climatic variables and sentiment index. The correlation was performed using the monthly sentiment index for the water bodies with information for all the time ranges (at least one article per month). Furthermore, the climatic events were extracted as the maximum and the minimum values of precipitation and temperature for the spatial area of the water body.
3. Confusion matrix of extreme events identified by climatic variables (precipitation and temperature) versus the number of articles over time. Extreme events were identified as monthly values smaller than 5% of the selected range values of the climatic variables. Likewise, months with values that exceeded 95% of the selected time series were classified as extremes. For this analysis, annual time ranges were selected.

10.3.9. Step 9 (Geovisualization)

Finally, these results were incorporated in interactive maps using the Google Earth search engine, in which the largest amount of the information obtained was represented. This allowed a second verification through visualization of the georeferenced data and an analysis of the transitions in the water bodies in the region that had occurred over time.

10.4. Results

10.4.1. Step 1

Initially, 81 digital newspapers were identified after conducting a search of newspapers published in every department of the country. However, only 42 newspapers were selected. A list of the newspaper sources is shown in Table 10.2.

Table 10.2 Selected newspaper sources grouped by region.

Department	Newspapers
All	<i>El Tiempo - El Espectador - Publimetro</i> <i>- La República - Semana</i>
Antioquia	<i>El Colombiano - El Mundo</i>
Atlántico	<i>Al día - El Heraldo</i>
Bolívar	<i>El Universal - El Bolivarense</i>
Boyacá	<i>El Diario - Siete días Boyacá</i>
Caldas	<i>La Patria - Eje21</i>
Caquetá	<i>Florencianos - Diario del Caquetá</i>
Cauca	<i>El Nuevo Liberal - Diario del Cauca</i>
Cesar	<i>El Pílon - Diario del Cesar</i>
Córdoba	<i>El Meridiano de Córdoba - Tu Prensa Digital</i>
Cundinamarca	<i>Extra - Periodismo Público</i>
Huila	<i>La Nación - Opanoticias</i>
La Guajira	<i>Diario del norte</i>
Magdalena	<i>Hoy diario del Magdalena - Seguimiento</i>
Norte de Santander	<i>La Opinión - Así es Cúcuta</i>
Quindío	<i>La crónica del Quindío - El Quindiano</i>
Risaralda	<i>El Diario - El Expreso</i>
Santander	<i>Vanguardia Liberal</i>
Sucre	<i>El Meridiano de Sucre</i>
Tolima	<i>El Nuevo Día - El Olfato</i>
Valle del Cauca	<i>El País - Diario Occidente</i>

10.4.2. Step 2

In total, 207 water bodies were identified and distributed, as shown in Figure 10.7. Because of the scale and the area covered by each regional newspaper source, little or no information was found for small basins within the expected time search.

An examination between the newspaper circulation size (region covered) and the extension of the water body was carried out. The news articles about the largest water bodies reported more information from more sources, while the news items about small water bodies presented the most accurate information at the local level. It was also not possible to access some newspapers because of website restrictions or property law restrictions.

10.4.3. Step 3

As a result of the digital citizens' media search, 19,490 newspaper articles from January 2016 to July 2020 were found. The largest number of articles regarding the analyzed type of water body featured the Magdalena River (4,405 news articles), followed by Hidroituango Dam (2,600 news articles), Cienaga Grande de Santa Marta (swamp, 621 news articles), and Fúquene Lake (94 new articles).

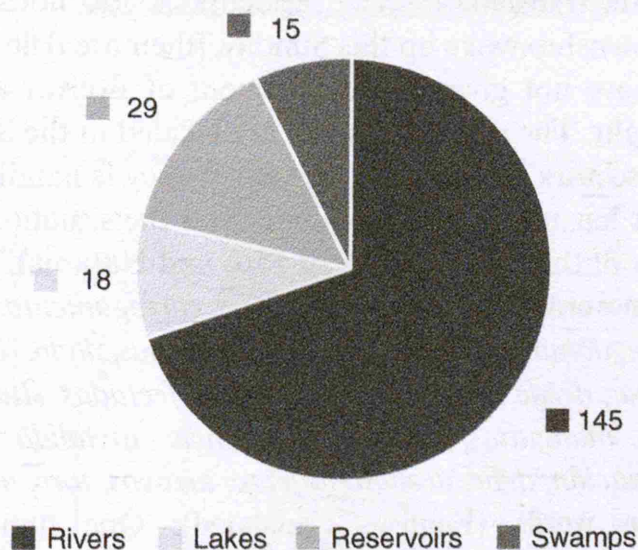


Figure 10.7 Distribution of 207 water bodies.

10.4.4. Step 4

A summary of the Web-scraping process is shown in Figure 10.8, transforming a website into an HTML format, and then into a CSV file. An example of the information extracted from a news article is illustrated in Figure 10.9.

As the main result, the distribution of articles featuring the Magdalena River Basin (by type of water body) is shown in Figure 10.10. The rivers are the main water bodies covered in digital news.

10.4.5. Step 5 (Tokenization)

A news article from Arroyo Alférez is shown as an example of the tokenization performed. The selected article was published in the newspaper *El Universal* on 19 November 2016, with the title, “Inundaciones en San Cayetano y deslizamiento de tierra en la Troncal.” Below is the first paragraph of the article and its tokenization (including the removal of stop words):

- **Text (Spanish, original):** “*Cien casas del corregimiento de san cayetano amanecieron este domingo inundadas por las fuertes lluvias que no le han dado tregua al departamento de bolívar desde la noche del sábado. Las casas afectadas están situadas en los sectores san martín los manguitos y bocatá. la emergencia es atendida por la defensa civil que informó que la situación se debe al desbordamiento de los arroyos el toro y naranjal.*”
- **Text (English, translated):** “The residents of 100 houses in the San Cayetano township woke up this Sunday, [their area] flooded by heavy rains that have not given the department of Bolívar a respite since Saturday night. The affected houses are located in the San Martín los Manguitos sectors and Bocatá. The emergency is handled by the civil defense that has informed [the public] that the situation is caused by the overflow of the river streams El Toro and Naranjal.”
- **Tokenization words (Spanish):** *Cien, casas, corregimiento, san, cayetano, amanecieron, domingo, inundadas, fuertes, lluvias, dado, tregua, departamento, bolivar, desde, noche, sábado, casas, afectadas, situadas, sectores, san, martín, manguitos, bocatá, emergencia, atendida, defensa, civil, informó, situación, debe, desbordamiento, arroyos, toro, naranjal*
- **Tokenization words (English, translated):** One hundred, houses, San, Cayetano, township, woke, Sunday, flooded, heavy, rains, given, department, Bolívar, respite, since, Saturday, night, affected, houses, located, San, Martin, Manguitos, sectors, Bocatá, emergency,

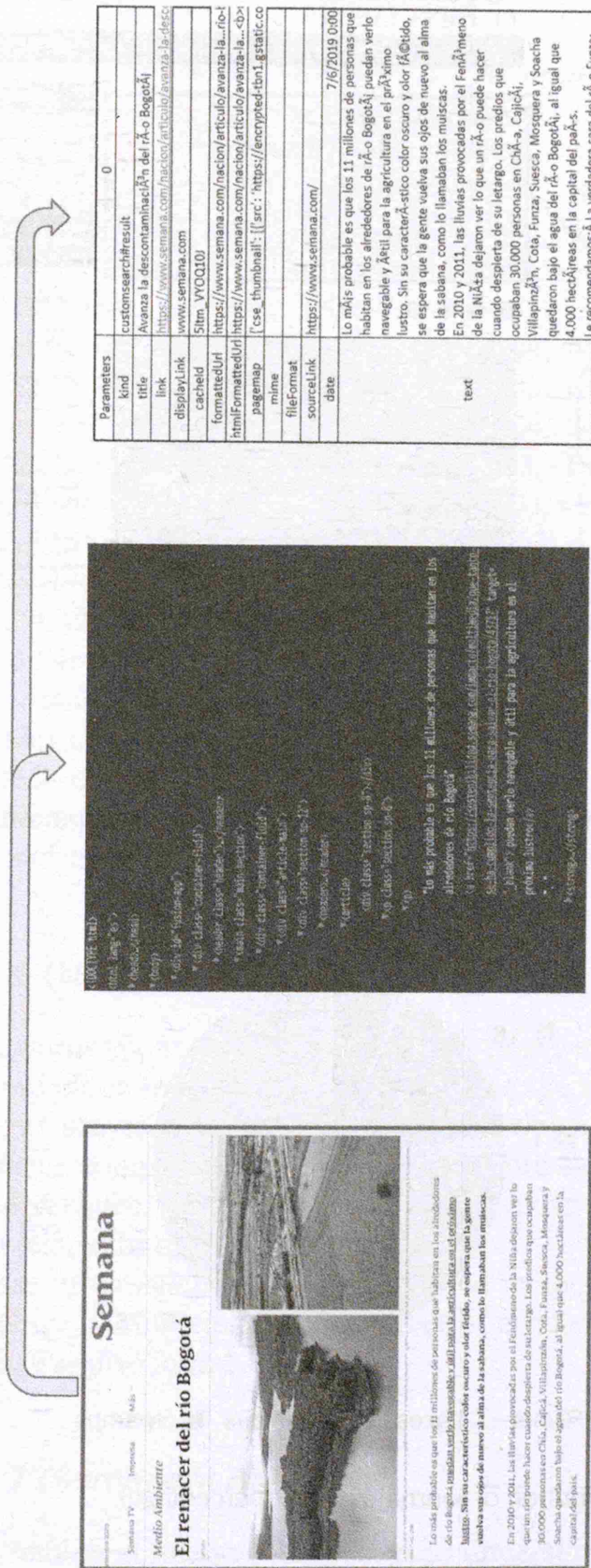


Figure 10.8 Web-scraping simplified process.



Figure 10.9 Example of information extracted from a newspaper website (<https://www.elespectador.com>).

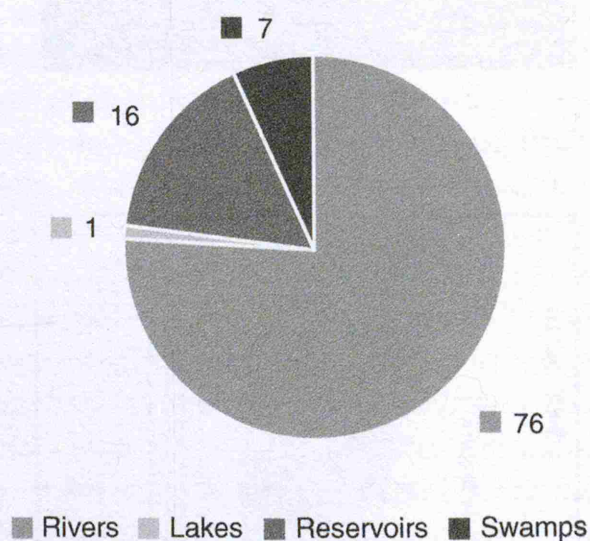


Figure 10.10 Distribution of journal articles (percentage).

attended, civil, defense, informed, situation, generated, overflow, river, streams, Toro, Naranjal

- **Tokenization vector (word: “overflow”).** *This result is calculated for each word given the paragraph as corpus:* [1.61e-03, -1.59e-03, -4.26e-03, -4.87e-03, -1.46e-03, 1.09e-03, -3.36e-03, -1.99e-03, 8.40e-04, -2.99e-03, 2.48e-03, -3.71e-03, 2.06e-03, 2.68e-03, -1.53e-03, 4.51e-03, 4.11e-03, 2.23e-03, -2.47e-03, 1.03e-03, -4.58e-03, 3.82e-03, -4.65e-03, -9.20e-04, 2.32e-03, -2.91e-03, 4.91e-03, -4.60e-04, -4.42e-03, 3.40e-03, -3.38e-03, -2.47e-03, -3.05e-03, 2.19e-03, 2.48e-03, -2.50e-03, -3.09e-03, 2.98e-03, 2.23e-03, -1.60e-03, -3.25e-03, -2.60e-04, 6.70e-04, -4.09e-03, -6.20e-04, 4.08e-03, -3.78e-03, 1.33e-03, 4.60e-04, 2.55e-03, 1.01e-03, -8.60e-04, 4.79e-03, 3.01e-03, 3.02e-03, -1.76e-03, 9.10e-04, 1.79e-03, -3.66e-03, 3.73e-03, 9.90e-04, 2.00e-05, 1.19e-03, -4.63e-03, -1.30e-04, -1.81e-03, -2.02e-03, 3.05e-03, 1.33e-03, 2.33e-03, 2.49e-03, -1.82e-03, 7.40e-04, 1.96e-03, -3.34e-03, 1.90e-04, 1.42e-03, 1.86e-03, 4.70e-04, 3.95e-03, -2.77e-03, 1.17e-03, 1.88e-03, 4.19e-03, 2.70e-03, -3.21e-03, -4.90e-04, -4.94e-03, -4.49e-03, -3.90e-04, 2.70e-03, 4.70e-03, 4.92e-03, -3.30e-03, 3.32e-03, 3.54e-03, -1.27e-03, -3.43e-03, 3.16e-03, -3.77e-03]
- **Top three words similar to “overflow,” given the vector distance (minimum distance):**
 1. affected - distance: 0.275
 2. township - distance: 0.183
 3. rains - distance: 0.135

10.4.6. Step 6 (NLP Analysis)

The main words about each water body over time were found and represented as word cloud images. Figure 10.11 shows the word cloud image for Quebrada el Carmen, with the word *Ecopetrol*, which is the largest and primary petroleum company in Colombia. In the word cloud, we also identified the word *derrame*, which means “oil spill,” illustrating the region’s main problem related to the river.

An increase in newspaper articles published in the last years of the study period (Fig. 10.12), tripling the number of articles written between 2016 and 2019, was also found.

10.4.7. Step 7 (Sentiment Analysis)

Applying sentiment analysis allowed us to understand the social perception about the water bodies. For the case, the *spanish_sentiment_analysis*



Figure 10.11 Word cloud image for Quebrada el Carmen representing the news articles about the water bodies (rivers) composing the Magdalena River Basin.

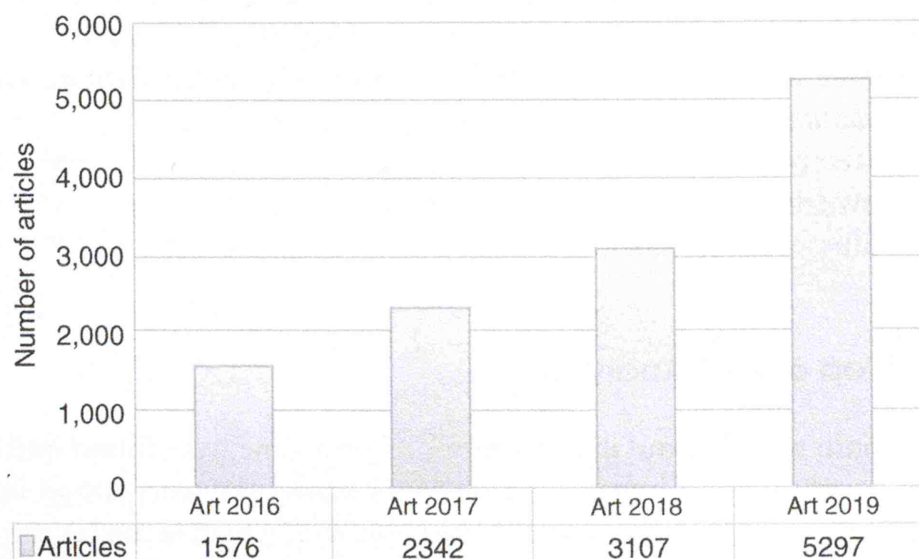


Figure 10.12 Number of newspaper articles about the rivers composing the Magdalena River Basin.

trained model was tested and MAE values of 0.068 and 0.198 were found regarding positive and negative reviews respectively. These values indicate the improved performance of the model regarding positive reviews. This is a logical consequence of the data sources of calibration, as positive reviews are more frequent in these websites. For example, we analyzed the sentiment toward the water body (Hidroituango Reservoir) over time. This sentiment shows the news writer's opinion, decreasing after the emergency,

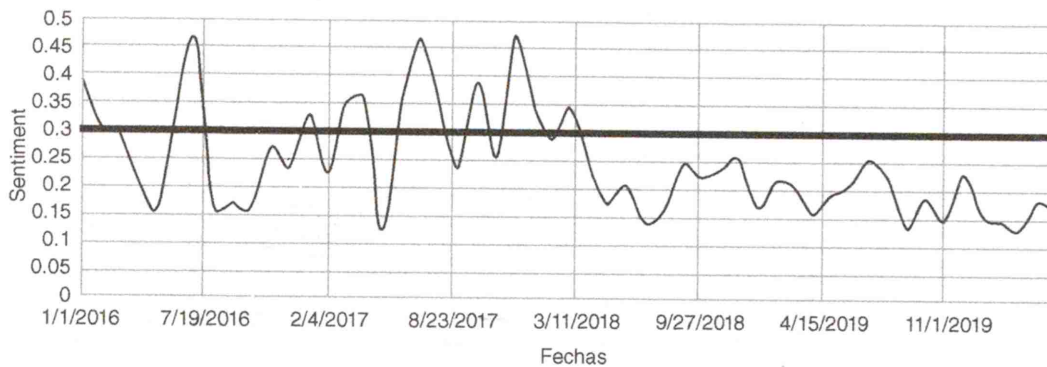


Figure 10.13 Sentiment change over time: Hidroituango.

in comparison with the previous months, as presented in Figure 10.13. This result indicates a negative mainstream opinion about the project due to the emergency. The y-scale values in Figure 10.13 are limited to those founded in Hidroituango Reservoir, while the horizontal line, intend to represent clearly the changes in sentiment before and after the emergency.

10.4.8. Step 8 (Validation)

1. *Sentiment index versus monthly river-level ranges next to Hidroituango Dam and Bogotá River.* The graphical results of the comparison between the sentiment index and the monthly river-level ranges of the Cauca River stations (located next to the dam) related to the Hidroituango Dam emergency in April 2018 are shown in Figure 10.14. Likewise, the Bogotá River sentiment results over time were validated through a river-level station at the river mouth, as shown in Figure 10.15.

In both cases, larger monthly level ranges are related to decreases in the sentiment toward the water body (dark dashed lines, Figs. 10.14 and 10.15). Nevertheless, due to the characteristics of each water body, its associated problems, and main physical variables, it is not possible to infer a direct relation between sentiments toward water bodies and flow/level changes. This suggests detailed research on each water body to understand the nexus between communities and water bodies.

2. *Correlation of climatic events and sentiment index.* Due to the lack of available news covering the complete time range, only Bogotá River, Cauca River, and Magdalena River were used in this analysis. The results of the correlation of climatic variables (precipitation and temperature) and the sentiment index are shown in Table 10.3.

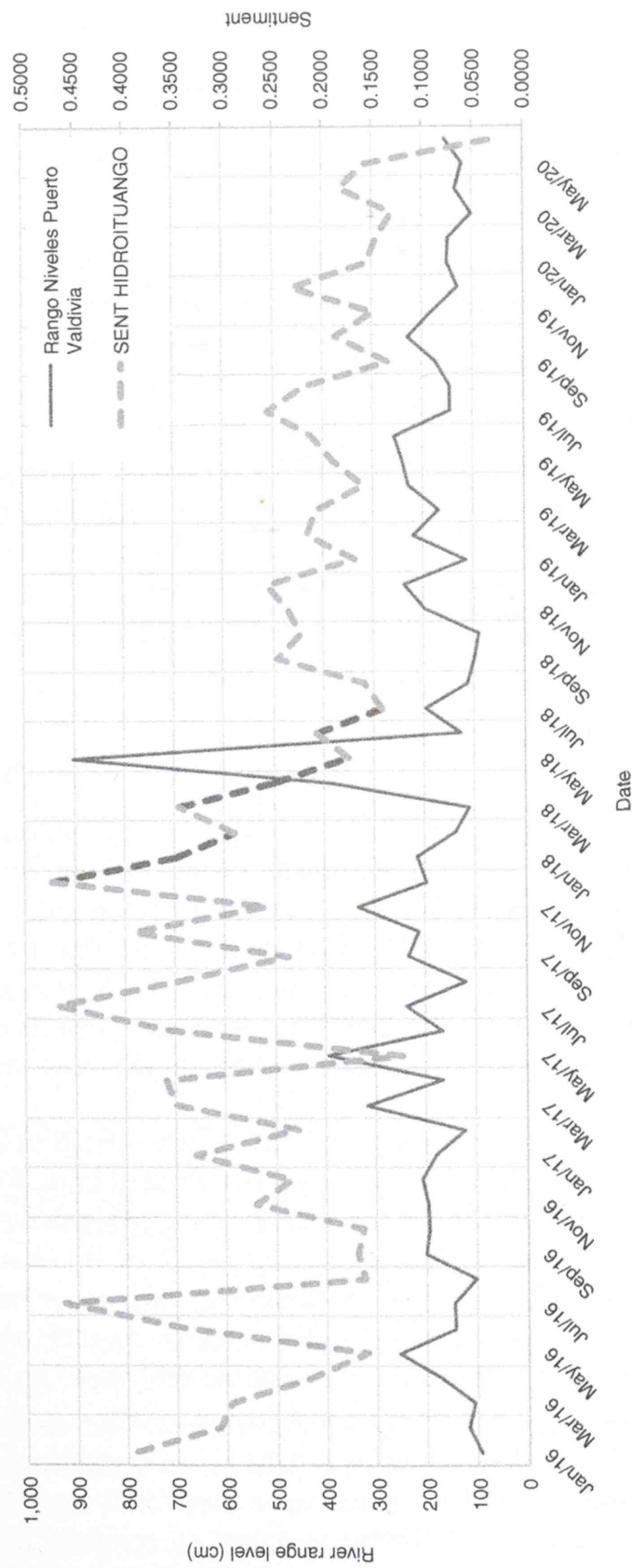


Figure 10.14 Monthly river-level range, Puerto Valdivia station in the Cauca River (Continuous line) versus Hidroituango sentiment (Dash line).

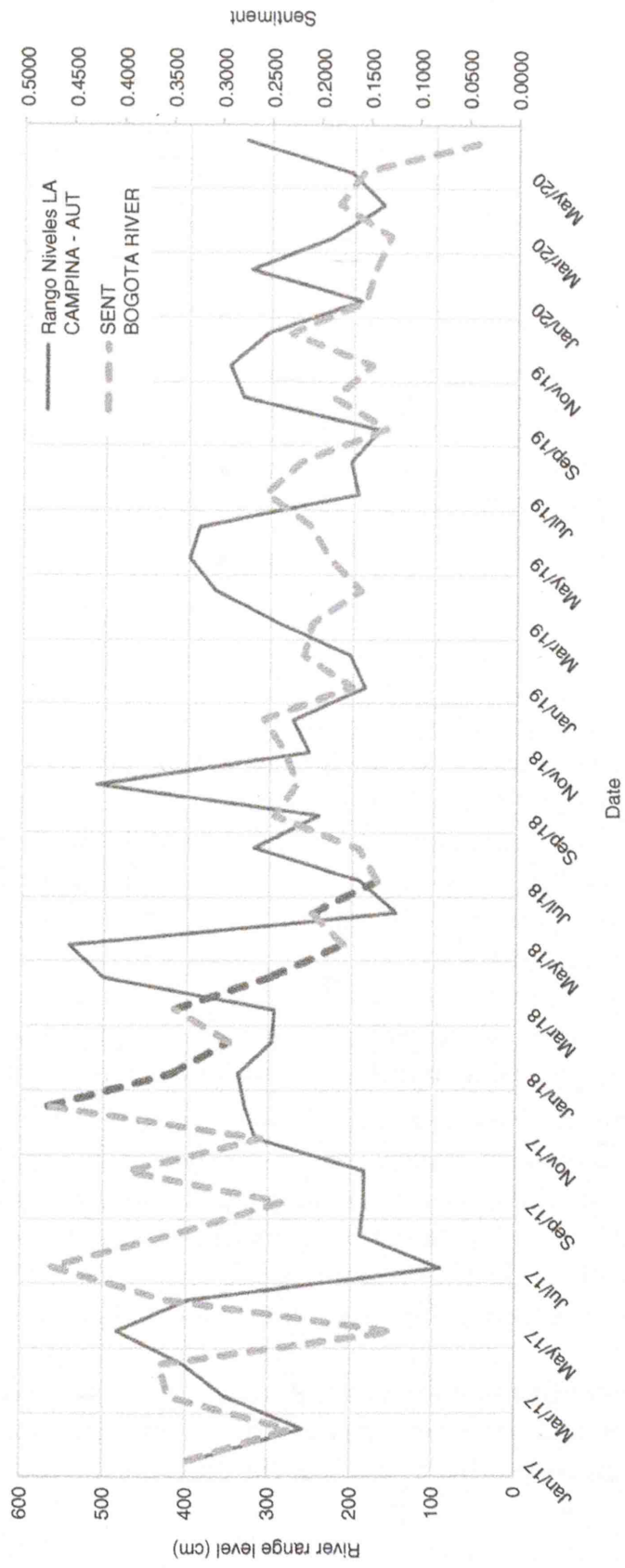


Figure 10.15 Monthly river-level range, La Campina station (Continuous line) versus Bogotá River sentiment (Dash line).

Table 10.3 Correlation of monthly sentiment index and monthly climatic data.

Element	Precipitation		Temperature	
	Min	Max	Min	Max
Bogotá River	-0.050	0.010	0.038	0.023
Cauca River	0.032	-0.099	-0.006	-0.130
Magdalena River	-0.101	-0.217	0.018	-0.197

Table 10.4 Confusion matrix results from Ayapel Swamp, showing high and low values of Precipitation (Pr.) and Temperature (Temp.) versus sentiment index.

	True negative	False positive	False negative	True positive
Pr. min	34	15	1	2
Temp. min	35	17	0	0
Pr. max	33	16	2	1
Temp. max	35	17	0	0

These results are explained by the difficulty in filtering the topic from the obtained news related to climatic conditions (e.g., higher temperatures and precipitation). This is an important statement, considering that most of the newspaper articles include components that cannot be easily measurable by means of physical variables. Likewise, a precise identification of the spatial location and the temporal range of the events remains a challenge.

These results reveal that the spatial distribution, the amount of information, the time range, and the integration of multiple topics are key components for using digital news information in an IWRM.

3. *Confusion matrix of extreme events identified by climatic variables (precipitation and temperature) versus number of articles over time.* Two examples of the confusion matrix application are presented in Tables 10.4 and 10.5. The first one refers to the Ayapel Swamp in the northern area of the Magdalena River Basin. The Betania Reservoir results are also presented. This reservoir is located in the upper river region.

While the confusion matrix results are inconclusive because of the multiple climatic, social, or environmental variables that can influence a news article, it is possible to analyze the patterns for which the True

Table 10.5 Confusion matrix results from Betania Reservoir, showing high and low values of Precipitation (Pr.) and Temperature (Temp.) versus sentiment index.

	True negative	False positive	False negative	True positive
Pr. min	42	7	1	2
Temp. min	40	9	3	0
Pr. max	40	9	3	0
Temp. max	42	7	1	2

Positive area is presented. In the first case, the low and the high monthly precipitation values imply the relation and the dependence between the Ayapel Swamp operations and the precipitation variable; in comparison, the temperature variable is not that relevant. In contrast, in the second example, low precipitation and high temperatures coincide with the patterns that affect the operation of a reservoir. Low precipitation may be related to water shortage, while high temperature values are related to the water evaporation process. It is also important to note that the increase in the number of articles over time is a very relevant parameter for the confusion matrix. The latest years with most of the articles may not be compared directly with earlier years, so the available information is less. Thus, using the number of articles as a variable in the matrix confusion should be checked in further analysis.

10.4.9. Step 9 (Geovisualization)

Finally, through the Google Earth search engine, it was possible to represent the indexing of the water bodies (swamps, reservoirs, lakes, and rivers) with their respective sentiments over time. The maps can be found at <https://santiagoduarte09.users.earthengine.app/>. Figure 10.16 shows an example of a sentiment change regarding the river basins between 2016 and 2019.

It was also possible to simultaneously visualize the sentiment variation and the spatial distribution of the ERA5 information (precipitation and temperature) for the same period, as shown in Figure 10.17 for precipitation.

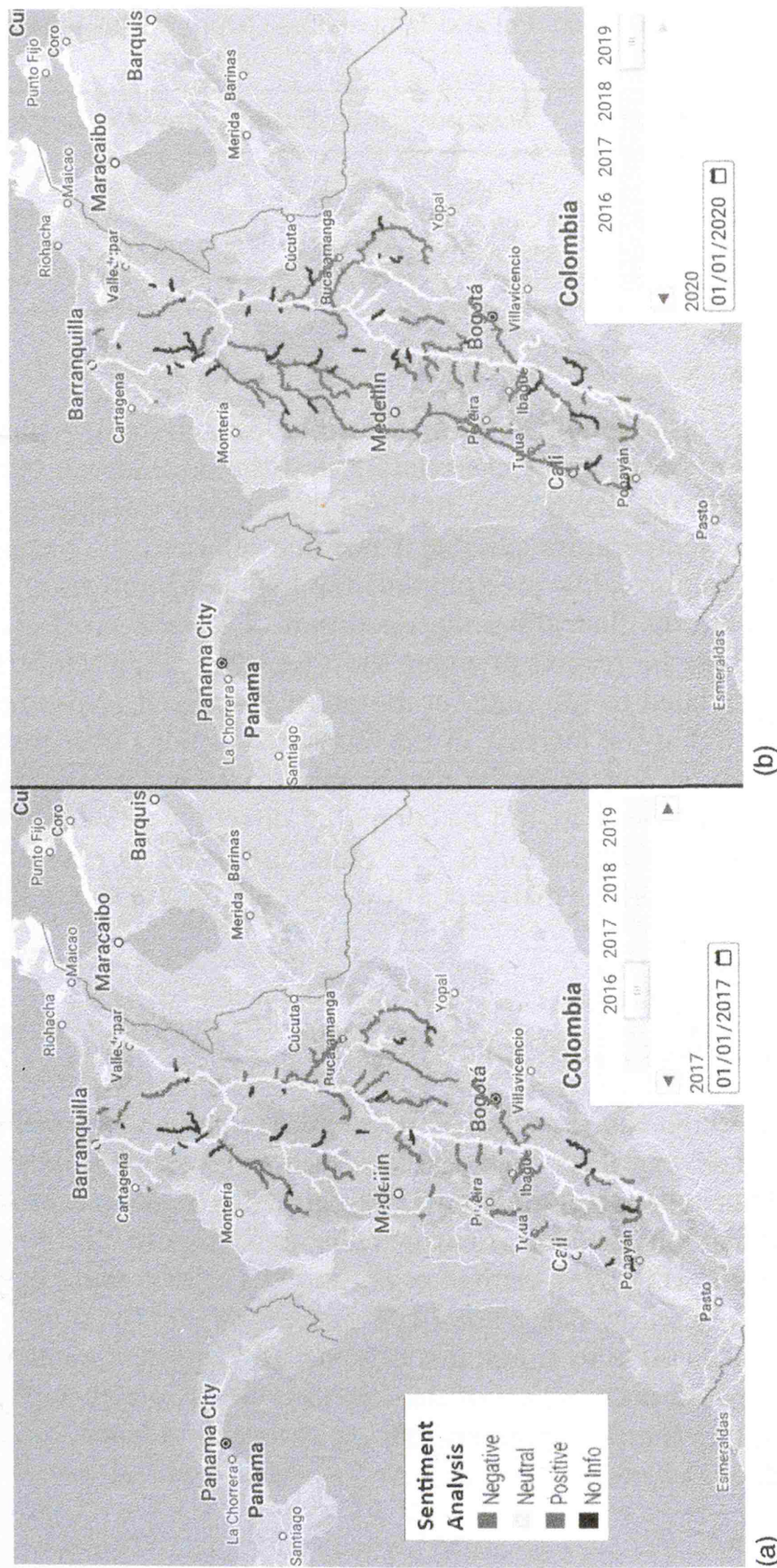


Figure 10.16 Spatial sentiment changes regarding the Magdalena River Basin, between (a) 2016 and (b) 2019.

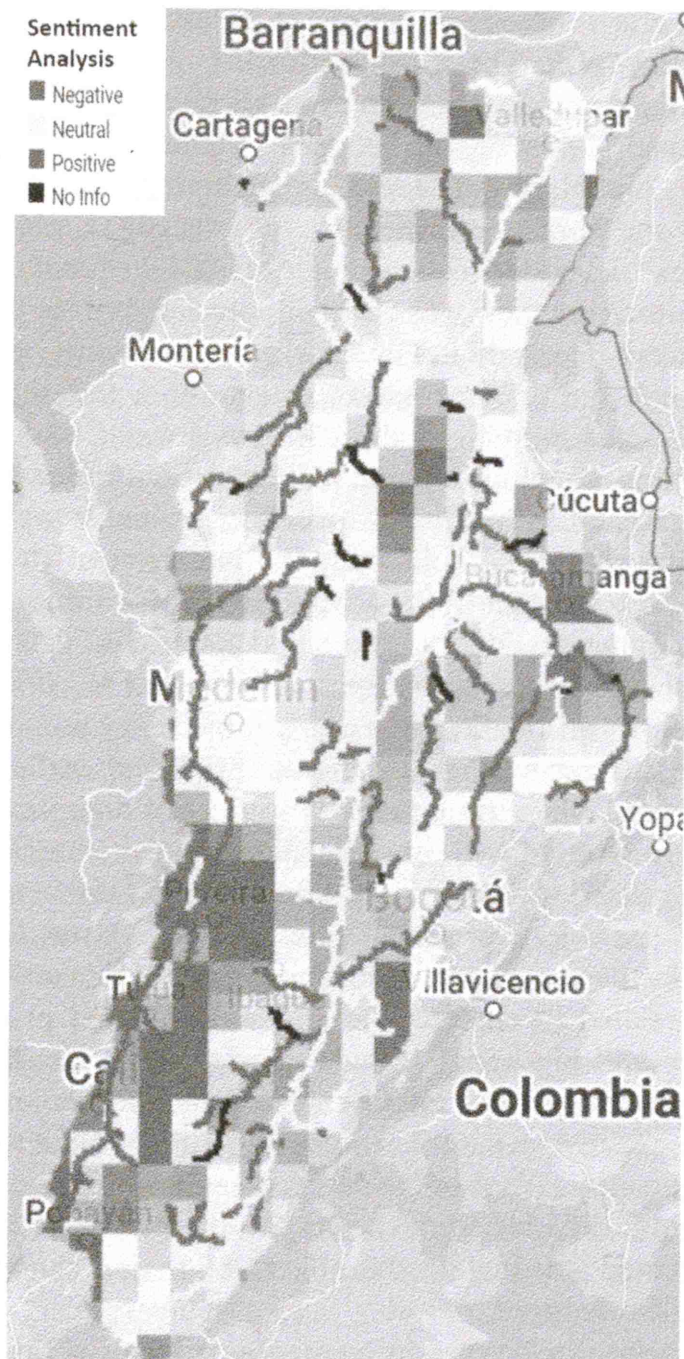


Figure 10.17 Spatial comparison sentiment and climatological variables regarding the Magdalena River Basin, precipitation, 2018 (Dark grey = wet; Light grey = dry).

10.5. Conclusions

Analysis of digital citizens' information allows us to understand the existing relation between communities and water bodies, as well as their cultural appropriation. We found that physical changes in water bodies

with significant effects on communities can be analyzed over time through news articles and NLP. The identification of main topics about water bodies included in newspaper articles can illustrate the big picture of water resources problems for a region, as presented in Quebrada el Carmen. In this context, digital news is representing digital citizens' information from a part of the population, in this case journalists, which describe a mainstream view of a water-related topic for a given spatial and temporal range. This mainstream view was clearly appreciated in the Hidroituango emergency analysis. Here, digital citizens' information has been captured indirectly through the documentation of one or a few individuals and shared with the community in digital news format. Additionally, it's also observed and concluded about editorial bias influence on digital news results using NLP. In other words, the journalist writer's political or environmental affinities could twist the intention of the information documented, and consequently it's advised to include subjectivity analysis in further developments. Other authors had also found and analyzed this bias in their works regarding newspapers (Gangula et al., 2019; Valdeavilla & Pulido, 2019). In this work, the writer's subjectivity was not measured, nevertheless, it was expected that individual political and environmental biases were reduced when using multiple sources of digital newspapers, as well as use information of well-established newspaper editorials. Additionally, politically related topic news was filtered within the search news step. Likewise, given the objective and activities of this research, it was deduced (as illustrated in the geovisualization step, within the maps) that most of the news was related to emergencies and water-related problems. These negative stories and big impact events allowed us to identify challenges in IWRM. Furthermore, this implies a coverage bias, as journalists tend to write about negative events. This bias should be studied further when evaluating the positive effects of the decision-making process, and applying NLP outputs in a broader context. Despite these drawbacks, newspaper articles represented adequately the effects of major changes on water bodies (e.g., Hidroituango).

The comparison between sentiments toward water bodies and physical variables is essential for a better understanding of the relations in the water-energy-food nexus. Sentiment index and monthly river-level ranges comparison in Hidroituango Dam and Bogotá River allowed us to satisfactorily identify drastic sentiment changes over. From the correlation between climatic events and sentiment index, two main statements are obtained. First, it is necessary to have an improved geolocation methodology. Spatial location derived from water bodies shows it is not enough when analyzing

long rivers, as point-located news are not easily geolocated in the river. Second, it was found that the date of publication of a newspaper article is not representative of the time occurrence of a hydrological event in small water bodies with few data sources. In these cases, the publication date denoted the timestep when the hydrological event was a mainstream topic for the community. Nevertheless, the visualization of georeferenced sentiment analysis results has proven to be a useful tool to understand the spatial changes in water bodies over time. Additionally, the results of the confusion matrix of extreme events provides a new understanding regarding the complexity when linking physical variables with digital media, as external factors could affect the sentiment index regarding hydrological events. In the context of NLP analysis on newspapers, other authors suggest the existence of sets of assumptions that imply a different interpretation of sentiment analysis results regarding the field of study (Puschmann & Powell, 2018), while Baccarella et al. (2018) exhibits the risk of social media and fake news. These statements reinforce the relevance of a digital news filtering process toward a newspaper database of water-resources topics. Consequently, this discussion proposes that a deeper analysis should be carried out when analyzing the perceptions of digital news and our understanding of it. Finally, we conclude by stating that NLP analysis of newspaper articles can provide additional information for decision making at the basin level, based on digital news information about the Magdalena River Basin's water bodies, however, this should be checked against physical variables and improved in spatial and temporal resolution when implementing action plans.

It is foreseen that other sources of information about extreme hydrological events could be explored through NLP (e.g., Twitter). Further, a specialized word-corpus based on water resource management is highly recommended for improved results. Furthermore, additional water cycle components should be incorporated in NLP applications, as aquifers, subsurface flows, and humidity. In future works, a deeper geospatial analysis between digital citizens' information and measured physical variables could provide additional tools for decision-making processes. We also recommend to explore new machine learning approaches for NLP involving water resources, as new advances could improve their results. Moreover, while new machine learning developments could provide new results, new additional challenges arise when trying to implement the work results in a decision-making process for community development and as a new methodology within the broader academic study of water-resource management.

References

- Avellaneda, P. M., Ficklin, D. L., Lowry, C. S., Knouft, J. H., & Hall, D. M. (2020). Improving hydrological models with the assimilation of crowdsourced data. *Water Resources Research*, 56(5), e2019WR026325. <https://doi.org/10.1029/2019WR026325>
- Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious! Understanding the dark side of social media. *European Management Journal*, 36(4), 431–438. <https://doi.org/10.1016/j.emj.2018.07.002>
- Badham, J., Elsayah, S., Guillaume, J. H. A., Hamilton, S. H., Hunt, R. J., Jakeman, A. J., et al. (2019). Effective modeling for integrated water resource management: A guide to contextual practices by phases and steps and future opportunities. *Environmental Modelling & Software*, 116, 40–56. <https://doi.org/10.1016/j.envsoft.2019.02.013>
- Bai, H., Yu, H., Yu, G., & Huang, X. (2020). A novel emergency situation awareness machine learning approach to assess flood disaster risk based on Chinese Weibo. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-020-05487-1>
- Bello, H. J. (2020). *Sentiment-analysis-spanish: Sentiment analysis for sentences in Spanish* (Version # 0.0.24) [Python; OS Independent]. Hugo Bello. <https://github.com/sentiment-analysis-spanish/sentiment-spanish>
- Borja-Vega, C., de Groot, K., & Serrano, H. A. (2020). *Colombia, turning the tide: Water security for recovery and sustainable growth*. Snapshot (Spanish). Water Security Diagnostics, Volume 1, Report 152351. Washington, DC : World Bank Group. <http://documents.worldbank.org/curated/en/212231615975980795/Snapshot>
- Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), 74–80. <https://doi.org/10.1109/MIS.2017.4531228>
- Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89. <https://doi.org/10.1002/aris.1440370103>
- Gangula, R. R. R., Duggenpudi, S. R., & Mamidi, R. (2019). *Detecting political bias in news articles using headline attention*. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 77–84. <https://doi.org/10.18653/v1/W19-4809>
- Garvey, C., & Maskal, C. (2020). Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. *OMICS: A Journal of Integrative Biology*, 24(5), 286–299. <https://doi.org/10.1089/omi.2019.0078>
- Giupponi, C., & Gain, A. K. (2017). Integrated water resources management (IWRM) for climate change adaptation. *Regional Environmental Change*, 17(7), 1865–1867. <https://doi.org/10.1007/s10113-017-1173-x>
- Goldberg, Y., & Levy, O. (2014). Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv:1402.3722* [Cs, Stat]. <http://arxiv.org/abs/1402.3722>

- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261–266. <https://doi.org/10.1126/science.aaa8685>
- Hofman, E. (2021). Ayllyote/senti-py [Python]. Elliot Hofman. <https://github.com/ayllyote/senti-py> (Original work published 2017).
- Jang, B., Kim, I., & Kim, J. W. (2019). Word2Vec convolutional neural networks for classification of news articles and tweets. *PLOS ONE*, 14(8), e0220976. <https://doi.org/10.1371/journal.pone.0220976>
- Kam, J., Stowers, K., & Kim, S. (2019). Monitoring of drought awareness from Google Trends: A case study of the 2011–17 California drought. *Weather, Climate, and Society*, 11(2), 419–429. <https://doi.org/10.1175/WCAS-D-18-0085.1>
- Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2020). Determining disaster severity through social media analysis: Testing the methodology with South East Queensland flood tweets. *International Journal of Disaster Risk Reduction*, 42, 101360. <https://doi.org/10.1016/j.ijdrr.2019.101360>
- Keegan, B. J., & Rowley, J. (2017). Evaluation and decision making in social media marketing. *Management Decision*, 55(1), 15–31. <https://doi.org/10.1108/MD-10-2015-0450>
- Le Coz, J., Patalano, A., Collins, D., Guillén, N. F., García, C. M., Smart, G. M., et al. (2016). Crowdsourced data for flood hydrology: Feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology*, 541, 766–777. <https://doi.org/10.1016/j.jhydrol.2016.07.036>
- Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *ArXiv:1309.4168* [Cs]. <http://arxiv.org/abs/1309.4168>
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2020.3009626>
- Murthy, D., & Longwell, S. A. (2013). Twitter and disasters: The uses of Twitter during the 2010 Pakistan floods. *Information, Communication & Society*, 16(6), 837–855. <https://doi.org/10.1080/1369118X.2012.696123>
- Nardi, F., Cudennec, C., Abrate, T., Allouch, C., Annis, A., Assumpção, T., et al. (2020). Citizens AND HYdrology (CANDHY): Conceptualizing a trans-disciplinary framework for citizen science addressing hydrological challenges. *Hydrological Sciences Journal*, 0(0), 1–18. <https://doi.org/10.1080/02626667.2020.1849707>
- Nielsen, R. K. (2015). *Local journalism: The decline of newspapers and the rise of digital media*. Bloomsbury Publishing.
- Perdomo, V. (2022). Colombia. Reuters Institute for the Study of Journalism. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/colombia>
- Picanço, J., Mesquita, M. J., & Soares, L. F. (2017). The hydrological disasters through historical survey in the Serra do Mar Range, Southern Brazil. In M. Mikos, B. Tiwari, Y. Yin, & K. Sassa (Eds.), *Advancing culture of living with landslides* (pp. 1017–1026). Springer International Publishing. https://doi.org/10.1007/978-3-319-53498-5_115

- Pires, A., Morato, J., Peixoto, H., Botero, V., Zuluaga, L., & Figueroa, A. (2017). Sustainability assessment of indicators for integrated water resources management. *Science of the Total Environment*, 578, 139–147. <https://doi.org/10.1016/j.scitotenv.2016.10.217>
- Prensa (2015, December 17). *De quién son los medios*. <https://www.monitoreodemedios.co/panoramaprensa/>
- Puschmann, C., & Powell, A. (2018). Turning words into consumer preferences: how sentiment analysis is framed in research and the news media. *Social Media + Society*, 4(3), 2056305118797724. <https://doi.org/10.1177/2056305118797724>
- Rossi, C., Acerbo, F. S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., et al. (2018). Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction*, 30, 145–157. <https://doi.org/10.1016/j.ijdrr.2018.03.002>
- Smith, A., Porter, J. J., & Upham, P. (2017). “We cannot let this happen again”: Reversing UK flood policy in response to the Somerset Levels floods, 2014. *Journal of Environmental Planning and Management*, 60(2), 351–369.
- Smith, L., Liang, Q., James, P., & Lin, W. (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modeling framework. *Journal of Flood Risk Management*, 10(3), 370–380. <https://doi.org/10.1111/jfr3.12154>
- Tiempo, C. E. E. (2018, May 17). *Nueva emergencia en Hidroituango mantiene en vilo a 12 municipios*. *El Tiempo*. <https://www.eltiempo.com/colombia/medellin/causas-de-la-nueva-emergencia-en-hidroituango-218440>
- Valdeavilla, D., & Pulido, M. (2019). *Bias in Filipino newspapers? Newspaper sentiment analysis of the 2017 Battle of Marawi*. Proceedings of the 4th International Conference on Internet of Things, Big Data and Security, 408–413. <https://doi.org/10.5220/0007752104080413>
- Watts, G. (2016). Hydrology with impact: How does hydrological science inform decision-makers? *Hydrology Research*, 47(3), 545–551. Scopus. <https://doi.org/10.2166/nh.2015.044>
- Yu, Y., Zhu, Y., Wan, D., Zhao, Q., Shu, K., & Liu, H. (2019). Applications of social media in hydroinformatics: a survey. *arXiv:1905.03035* [cs]. <http://arxiv.org/abs/1905.03035>