# An Exploratory Analysis of Digital Information using Natural Language Processing for the Planning and Decision Making Process of Water Resources in Bolivia

Camilo Andrés González Ayala

MSc Thesis Identifier WSE-IMHI 20-104
March 2021

**Updated version,** April 2021

# An Exploratory Analysis of Digital Information using Natural Language Processing for the Planning and Decision Making Process of Water Resources in Bolivia

Master of Science Thesis

by

**Camilo Andrés González Ayala**

Supervisors

Dimitri Solomatine

Germán Ricardo Santos Granados

Mentors

Gerald Augusto Corzo Pérez

Héctor Andrés Angarita Corredor

This research is done for the partial fulfilment of requirements for the International Master of Science degree at the IHE Delft Institute for Water Education, Delft, the Netherlands and

Escuela Colombiana de Inngeniería Julio Garavito, Bogotá, Colombia

**Delft**

20/04/2021

# Table of Contents

# List of Figures

# Abbreviations

| | | |
|---|---|---|
| AI | – | Artificial Intelligence |
| API | – | Application Programming Interface |
| CBOW | – | Continuous Bag of Words |
| IWRM | – | Integrated Water Resource Management |
| NER | – | Named Entity Recognition |
| NLP | – | Natural Language Processing |
| RBF | – | Radial Basis Function |
| RDS | – | Robust Decision Support |
| SA | – | Sentiment Analysis |
| SEI | – | Stockholm Environment Institute |
| SG | – | Skipgram |
| SVM | – | Support Vector Machines |
| TF-IDF | – | Term Frequency, Inverse Document Frequency |
| TRMM | – | Tropical Rainfall Measurement Mission |
| WRM | – | Water Resources Management |

# Acknowledgements

First of all I want to thank my mentors Gerald Augusto Corzo Perez, PhD, and Hector Andrés Angarita, PhD, for their knowledge, time, motivation, and mainly very good energy. They gave me all the necessary support in the different stages of this research in order to obtain promising results in an innovative project that can be studied in the future by new researchers.

My sincere thanks to the Stockholm Environment Institute who with their work team provided me with resources, knowledge and support material to understand the area of study in the research.

To Dimitri Solomatine, PhD, and Germán Ricardo Santos Granados, PhD, my supervisors, for their advice on the defense of the proposal and Mid-Term, that allowed me to push myself even more in my presentations. To the high quality institutions worldwide IHE Institute for Water Education and Escuela Colombiana de Ingeniería Julio Garavito for creating the International Master in Hydroinformatics program, it is an honor to be part of the first seniors of this program.

To each of the professors of Escuela Colombiana de Ingeniería Julio Garavito and IHE Institute for Water Education for their dedication, effort and for always keeping a smile in class in difficult times in the world due to Covid-19 situation.

This situation brought us closer to our family, to our friends, to the people who love and care about us. To my brothers, my father and my stepfather, my bachelor friends who were always there for me, childhood friends and Master friends, thank you for giving me words of encouragement while I was in the Netherlands, for giving me confidence and believe in me when I have the burden of the world on my shoulders.

Finally and the most important thanks to my mother Esperanza, who raised me with the best values. For motivate me every day despite my many concerns, for always supporting me in each of my projects and mainly for never letting me fall.

To each of you, thank you for being part of this process…

# Abstract

In recent years, the community is much more participatory in the planning and decision-making processes of Integrated Water Resources Management. However, differences between competing stakeholders prevent the identification of important variables in decision-making. In addition, the COVID-19 situation has prevented activities from being face to face with the community where fundamental information is collected for the planning process. Faced with this panorama, and with the aim of complementing the characterization of a water system, and provide an alternative that collaborates in the planning and decision-making process, this research focuses on analyzing digital information sources from the public media, obtaining useful information from articles associated with a basin. The case study corresponds to La Paz - Choqueyapu river basin in Bolivia. The information from 6 representative newspapers of that country, related to water resources, was extracted. An exploratory analysis of the information is executed and it is associated with historical information on hydrological phenomena such as precipitation in the last decade, finding a good correlation between both sources of information. Through the application of Named Entity Recognition, it was possible to identify different entities associated with bodies of water, dams, authorities and communities that are present in the basin.

Each of the articles is associated with a positive or negative sentiment according to its content in order to carry out a qualitative analysis of the basin. From the article and its associated sentiment, sentiment text classification models are build in the context of water resources with the extracted articles with different techniques of word embedding and classification machine learning algorithms. It was found that the model with the best performance corresponds to the SVM algorithm with linear kernel and Word2vec continuous bag of words word embedding, obtaining 84% accuracy. This result was compared with the value obtained through the Spanish Sentiment Analysis library of 63%, evidencing a high improvement in the classification of texts associated with water resources in the Spanish language. Finally, by finding the most frequent words in a positive or negative context, important variables can be evidenced for the improvement of the planning and decision-making process.

# Chapter 1 Introduction

The socio-economic development of a country depends mainly on adequate water resources management (WRM). Sectors such as mining, agriculture, navigation, and energy production are essential in local economies and negatively impact the water resource quality and availability. Planning and decision-making processes in the WRM have become, over the years, an interdisciplinary process to promote the public benefit of the society and boost socio-economic development. Participation and collaboration of experts, government authorities, and the community is one of the fundamental components considered by the Integrated Water Resource Management (IWRM) (Galvez & Rojas. 2019). The purpose of creating interdisciplinary knowledge assists more informed and creative decisions in a water system (Van Cauwenbergh et al. 2018).

The integration and collection of information are conventionally done through questionnaires, stories, perceptions and opinions of the community, debates, *etc*. This process is typically executed in person, which requires considerable time demand and financial resources. Besides, the community needs and priorities vary according to different factors such as its culture, location, economy, or particular interests of the government, which generates conflicts between competing stakeholders.

This conflict hinders the exchange of information between the different disciplines. Consequently, in the absence of information, some important variables for the planning and decision-making process in WRM may not be considered. Furthermore, this makes it difficult to understand the correlation between key actors and variables in a water system.

Also, The COVID-19 situation has prevented to collect information in person with all stakeholders. Considering this situation, it is proposed to explore alternative sources of information like digital information found on the web pages of communication media such as newspapers, radio, television in order to understand public opinion in the context of water resources.

According to the above, some authors have decided to explore digital information by applying artificial intelligence (AI) to understand public opinion in the context of water resources and transform it into useful information.

Authors like Murphy et al. (2014) used AI techniques such as Natural Language Processing (NLP) and Named Entity Recognition (NER) in order to find the interactions between actors and water management systems using information obtained from public media.

Other researchers used NLP as a text sentiment classifier. Reyes-Menendez et al. (2018) classified and related the tweets found in the hashtag #WordEnviromentDay on Twitter, for the year 2018, with the sustainable development goals (SDGs), by a sentiment analysis model. Xiong et al. (2020) with sentiment analysis, classify the topics from greatest to least concern in the community of Chennai, India in the shortage crisis of 2019. Duarte Prieto (2020) identify the main problems that have arisen in the main water bodies of the Magdalena river basin in Colombia using the Spanish Sentiment Analysis library.

Case study will be applied in the South-America country of Bolivia. Bolivia is one of the countries that is betting on improving the IWRM Every year, floods and droughts hit the most vulnerable populations in different regions of Bolivia. Floods represent the greatest hydroclimatological risk factor in the country along with landslides caused by heavy precipitation. Along with these challenges in the country, there is also inefficient water treatment for water supply which can lead to other problems like diseases.

To respond to these challenges, the Bolivia WATCH program, through a multi-participatory process, seeks to collect data that allows the identification of key actors, variables that generate water stress, and solutions that benefit the water resource management.

The purpose of this research is focused on contributing to the Bolivia WATCH program, recognizing how Bolivian society relates the management of water resources in La Paz – Choqueyapu hydrological basin, to improve the planning and decision-making process. The application of a SA classification model in Python language programming is developed with water keywords in the Bolivian context. The aim is to correlate the digital information[1] with the planning and decision-making process, identifying key elements, actors and frequent words in the IWRM field in order to see represented all the interests of the different stakeholders.

[1] For the purposes of this research, digital information will be understood as the information obtained from the web pages of the media such as newspapers, radio, television.

Following this introduction, the objectives, research questions, innovation, and practical value of this research are presented. Chapter 2 presents the literature review in more detail. The research methodology with a general description of the different methods developed is presented in chapter 3. Chapter 4 describes the case study in detail with the exploratory analysis of the available data and the specifics of the models set-ups. Results and discussion of the case study is presented in chapter 5. Chapter 6 includes conclusions and recommendations of this research. Finally, the references are presented.

## 1.1 General Objective

To obtain experience from digital information on how society perceives the water resource management in Bolivia.

## 1.2 Specific Objectives

- To understand the correlation between digital information and the planning and decision-making process in the water resource management in Bolivia context.
- To identify key actors and variables that generate water stress and the most frequent words related to water resources in La Paz – Choqueyapu basin.
- To develop a sentiment analysis classification model using artificial intelligence techniques through articles related to water resources context.

## 1.3 Research Questions

- What level of reliability has digital information in identifying key actors and variables that generate water stress in water resources management in Bolivia?
- Which text classification algorithm performs best in the application of a sentiment analysis model in the context of water resources?

## 1.4 Innovation

This is one of the few investigations that is being developed in the Natural Language Processing area in conjunction with Water Resources Management up to the knowledge of this researcher and on the literature review that has been explored so far.

Aside from that, this research is one of the first attempts to explore how terminology inside the concept of NLP for water resources can be incorporated and adapted to different languages. Also, applications of NLP in the context of Hydroinformatics are no so often use.

Moreover, this research can have a high impact on the concept of citizen science. The lack of knowledge in an expression that is transmitted through a text, a conversation, or an image, can be transformed into valuable information for decision-making.

## 1.5 Practical Value

The practical value of this research look at the implementation of a low-cost alternative that presents detailed information on how society perceives water resources management, by ranking frequent topics as negative or positive and their relation with locations, key actors, and water keywords in order to improve the planning and decision-making. Also, intent for motivate future researchers to go deeper into the application of NLP in the context of water resources management in different countries and cultures to reduce the gap of information. Finally, this research contributes to the need to implement innovative decision support tools considering new sources of knowledge for the Bolivia WATCH program.

# Chapter 2 Literature Review

This chapter cfocuses on the application of NLP as a text classifier in the context of environment and water resources.

Public media, social networks and questionnaires on the web are the main sources of digital information that express the perception and opinion on a specific topic in order to store it in a structured way. Recently, these means have been applied to complement studies related to the management of water resources and to monitor environmental quality factors.

Noga & Wolbring (2013) developed an online questionnaire of 37 multiple-choice questions in order to obtain people's opinion on water management in Calgary, Canada. The questions in the questionnaire consisted mainly of the perception of water use by the responsible authorities, daily consumption, water price, availability of the resource and access to drinking water. The results showed that the majority of people consider water as a private resource and identify parameters to improve the management of the resource by the competent entities.

On the other hand, Murphy et al. (2014) found a relationship between the actors, structure, and variation of water management systems in the Colorado River basin in the southwestern United States. This research focused on the application of two AI techniques such as natural language processing (NLP) and named entity recognition (NER) in more than 110,000 newspaper articles using water keywords. NER allowed the identification of dates (days of the week, months, years), personal titles (administrator, president, engineer), geographical terms, geographical classes, names of sites and bodies of water, among others in the basin under study. The Term-Frequency (TF) technique was applied in order to identify the most frequent words in the articles (related to the context under study) and thus find the most relevant documents. The main objective of the research was to develop networks, identifying the relationship between the different water systems and institutions, relating this information to the location through NLP. The study demonstrated that relevant information can be captured on the management of a water system in different locations, as well as the identification of entities and their interaction in time and space by applying NLP techniques.

One of the main uses of NLP and applied in the second part of the last decade in the context of water resources and the environment consists in the classification of texts and its main tool is sentiment analysis.

Using a sentiment analysis text classification model, using the Support Vector Machines (SVM) algorithm Wang et al. (2017) analyzed the opinion of the people in two very popular social networks in China towards the environment between 2015 and 2016, obtaining an 85.67% accuracy on average. This analysis made it possible to create a model for evaluating environmental quality in real time that, based on a person's feeling towards the environment, an Environmental Quality Index (EQI) can be obtained in a specific area and time.

In 2016, in the region of St. Louis, Missouri, United States, high levels of lead were found in school drinking water and several users expressed their disagreement through the social network Twitter. Ekenga et al. (2018) collected the tweets related to this topic in the period from August to December 2016 and classified the users in 5 categories such as the general public, news, government official, non-governmental authority and political-academic. Through these categories, sentiment analysis was applied to each of the tweets, finding wide differences and inequities between public opinion and academic and governmental opinion. These differences of opinion are a key factor when making more creative decisions in decision-making in the integral management of water resources.

Applying sentiment analysis and topic modeling, Zhang et al. (2018), monitored public opinion about the South-to-North Water Transfer Project (STNWTP) in China to determine the degree of support for the water conservation project in different regions. The sentiment classification model obtained an accuracy of 85.3% using the SVM algorithm. Once the public perception sentiment by regions was found, the impacts of the project on different social groups were analyzed and, therefore, more specific policies and more accurate decisions were formulated to prevent risk and threat events.

Performing a sentiment analysis model in Python, applying the SVM algorithm, Reyes-Menendez et al. (2018) classified and related the tweets found in the hashtag #WordEnviromentDay on Twitter, for the year 2018, with the 17 sustainable development goals (SDGs). The data were initially grouped according to the sentiment value found. Subsequently, through textual analysis, the tweets were classified in each of the sustainable development goals. Finally, the main issues of concern to the community regarding the environment and public health were identified.

Xiong et al. (2020) explore Twitter public opinion, identifying the most frequent discussion topics related to the water shortage crisis of 2019 in Chennai, India applying Latent Dirichlet Allocation model. To monitor the water crisis, sentiment analysis was developed using VADER to classify the topics from greatest to least concern in the community according to the text emotions as negative, neutral or positive.

Duarte Prieto (2020) in the contest "Conectate con el Río Magdalena (Connect with Magdalena River)" produce a web tool which consisted of identifying the main problems that have arisen in the main water bodies of the Magdalena river basin in Colombia. The base information consisted of newspaper articles from national and regional newspapers in Colombia to which Spanish sentiment analysis library was applied. The results obtained from the sentiment analysis were presented using web geovisualization tools.

From this chapter it can be concluded that recently the application of Natural Language in alternative information sources such as public and social media to improve the concept of water resource management is increasing. Also, the application of SVM as an algorithm for text classification is commonly used obtaining high precision values compared to other classification algorithms (Hernández & Gómez. 2013). However, most studies have been conducted in the English language. In Spanish, the only application of sentiment analysis in water resources, according to the literature explored by the researcher, is the one of Duarte Prieto in 2020. Consequently, this research is one of the first attempts to apply NLP and Sentiment Analysis in the context of water resources in the Spanish language.

# Chapter 3 Research Methodology

The research methodology consists of three approaches: learning, applicative and analytical.

The learning approach consists of the review of literature related to natural language processing and the experience related to its application in the context of water resources and the Spanish language. It seeks to identify the fundamental concepts of this technique for the execution of a text classification model through sentiment analysis.

The applicative approach consists of two phases: phase one seeks to perform an exploratory analysis of the information to be applied in this study, while the second focuses on developing a sentiment analysis classification model in the context of water resources.

Finally, the analytical approach presents an analysis of the results obtained from the applicative approach.

In Figure 1, the diagram corresponding to the research methodology is ilustrated. Subsequently, themethodology approaches are presented in more detail.

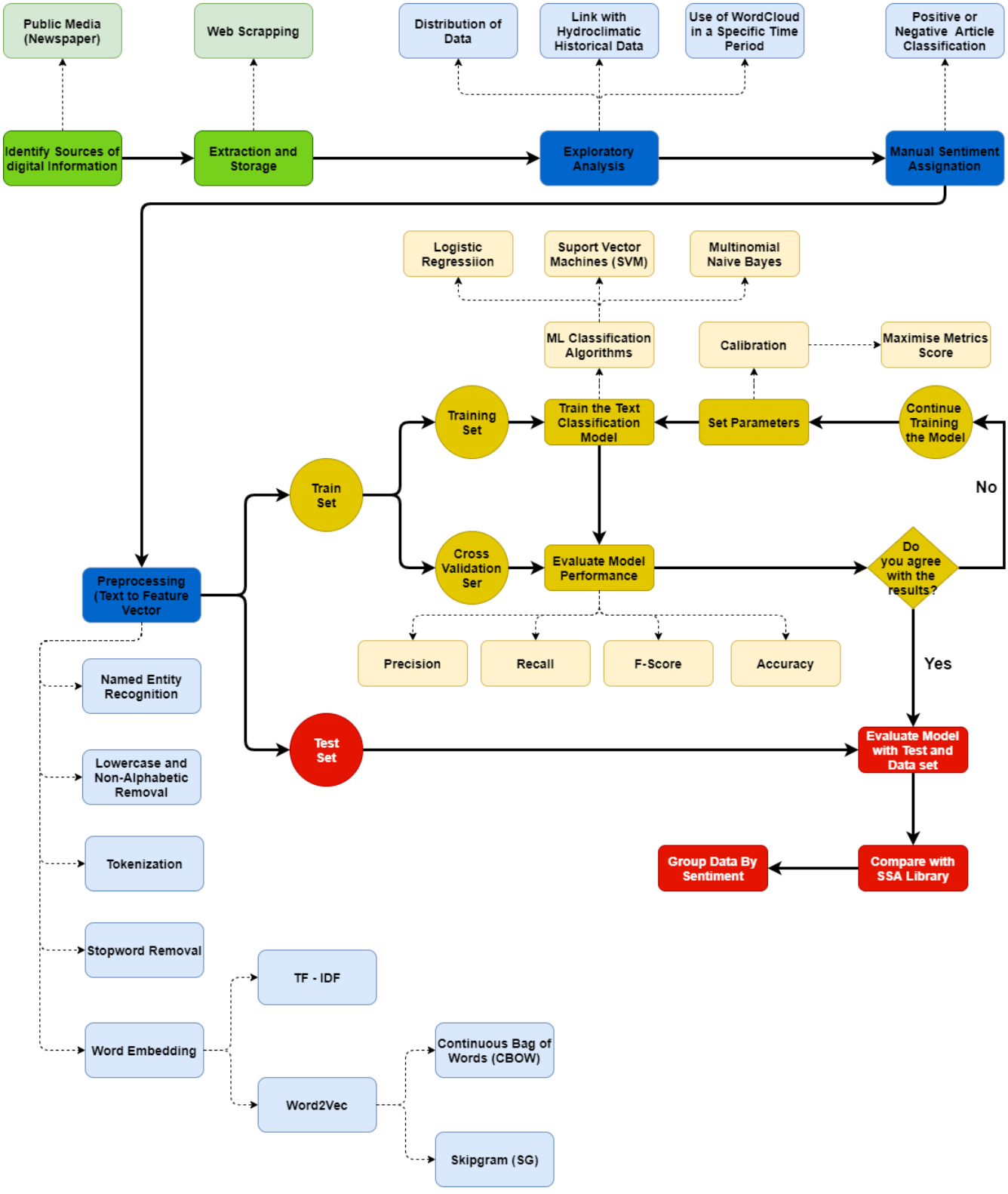



*Figure 1 - Research Methodology Diagram*

## 3.1 Learning Approach

The learning approach main objective is to understand the concepts in the Natural Language Processing and Sentiment Aanalysis process and which essential steps must be taken into account for their development in Python language programming. Python is chosen as language because it is open acces.

Libraries such as Natural Language Toolkit (NLTK) (Nalini et al. 2019) and Spacy (Sharma & Bansal. 2019) are known to perform the preprocessing of texts applying some important steps like tokenization, lematizing, stopword removal, stemming, Named Entity Recognition or Word Embedding to mention some of them. Scikit-Learn (Japhne & Murugeswari. 2020), VADER (Ilyas et al. 2020) and TextBlob (Subirats et al. 2020) have currently been used to develop sentiment classification models in different approaches such as marketing, politics and medicine. This libraries perform some classification, regression, clustering and also preprocessing algorithms.

This approach is summarized in chapter 2 corresponding to literature review in which the most common algorithms of text classification applied in the context of water resources were identified, such as Support Vector Machines.

In the applicative approach presented in chapter 3.2, some of the concepts and steps necessary to develop a sentiment analysis classification model are presented.

## 3.2 Applicative Approach

One of the main objectives of this approach is to obtain detailed information on topics frequently mentioned in the texts under study and their relationship with actors, authorities and water systems through a spatio-temporal analysis executing an exploratory analysis of the information. The other objective consists of build a sentiment analysis classification model in the context of water resources management. Typical steps to take when building a model are information extraction, information pre-processing, and classification (Kalaivani et al. 2019)

The information extraction is carried out using web crawling or web scraping techniques on web pages or social networks depending on the information that is required to be extracted.

It is important to transform the extracted texts to a feature vector that can be understandable for a machine. The information will be pre-processed using the libraries described in chapter 3.1. Word Embedding is one of the fundamental steps in information

preprocessing. Methods like Term Frequency - Inverse Data Frequency (TF-IDF) and Word2Vec are commonly used to convert text to vector.

Classification of information consists of assigning the text a value that represents it within a category according to its characteristics (e.g .: as positive or negative for sentiment analysis).

The execution of a model consists of two stages: Training and Testing. At training process, the model learns to associate a text with a category, based on the content of the applied sample texts. The testing process indicates how accurate the model is in classifying the text to the correct category.

Some important concepts in the steps of data extraction and storage, preprocessing, and classification algorithms applicable to this investigation are presented below.

### 3.2.1 Extraction and Storage of Information

This step consists of extracting and storing data, through web scraping techniques filtering the required articles by using keywords. It is important to have the seed links where the information will be extracted. The python libraries to apply correspond to BeauitifulSoup4 and Selenium. In this technique the structure of the web page must be considered since the extraction consists of inspecting the HTML code identifying the tags that store the information. Information like the title of the article, its content, the date of publication, the source of information and the reference link is extracted and stored as .csv format. The biggest advantage of using these libraries is that it does not have a search limit as with the Google Random Search API.

### 3.2.2 Preprocessing of the Information

The digital information that has been collected must be converted to a compatible format in order to be processed by a machine, in other words, preprocessed. This phase consists of representing the text present in a document numerically using a feature vector in order to perform an adequate numerical analysis of it, and apply different Machine Learning algorithms. Relevant information can be extracted from this process that allows understanding the historical relationships of water resource management and its perception. Based on the research of Kalaivani et al. (2019) and Murphy et al, (2014), some important steps that must be taken into account in the pre-processing of information are:

- ***Knowledge of Language:*** In order to perform the model is vital to know which language must be understood for the machine.
- ***Tokenization:*** This step split the words of a text into tokens, e.g., for the phrase "Water science and engineering." the tokens are [Water], [science], [and], [engineering], [.]
- ***Stop word removal:*** Stopword removal is performed in order to remove those words that do not add value to the text, such as articles or pronouns.
- ***Named Entity Recognition:*** It allows recognizing essential attributes of a text such as places, dates or names of entities or people.

These proposed steps are considered by this methodology accompanied by non - alphabetic characters removal.

As mentioned in the introduction to this section, **Word embeddings** are also considered and these are the methods that will be applied in this research:

- **Term Frequency - Inverse Document Frequency (TF-IDF):** Is a text characterization model based on words statistics to extract features from a text (Liu et al. 2018). Some semantic is preserved as uncommon words are given more importance than common words in a whole corpus. The idea of this model is to identify the term frequency (TF) value of a particular word on a text of a dataset, and the Inverse Document Frequency (IDF) value of the word in the whole data set. The text frequency of a word in a text is the number of times that the word appears. The IDF formula is presented on Equation 1.

*Equation 1 - Inverse Document Frequency*

$$IDF = log\left(\frac{Number\ of\ documents}{Number\ of\ documents\ containing\ the\ word}\right)$$

Then, the TF – IDF formula is presented on Equation 2.

*Equation 2 - TF-IDF*

$$\boldsymbol{TF - IDF = TF(word\ in\ a\ text) * log\left(\frac{Number\ of\ documents}{Number\ of\ documents\ containing\ the\ word}\right)}$$

The result is a matrix of decimal values (TF-IDF features) between 0 and ∞. Highest TFIDF value, represent more importance of a word in a dataset.

Scikit-Learn library contains a function that performs the word embedding process by TF-IDF called **TfidfVectorizer.** Some of the most important parameters are: "**Analyzer**"

that control if the feature word should be made of words or character n-grams. “**min_df**” that ignore all the words that not appear at least in the given min_df value documents in a data set. “**use_idf**” which enable de inverse document frequency reweighting, and “**smooth_idf**” which prevents zero divisions if an extra document was seen containing every word of the vocabulary exactly once. More information about how this model performs is found at Scikit-Learn webpage (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html).

- **Word2Vec:** Is a word embedding model introduced by Mikolov et al. (2013) that processes a text by vectorizing its words with the main idea that two words that are in the same context, share a similar meaning, in consequence, a similar vector representation. From this idea, The advantage of this model is that words that are present in similar contexts are related, mantining a higher semantic of the text compare to TF-IDF model.

  The objective of the model is to train a two layer neural network using pairs of words obtained from the vocabulary of words found in a text or set of texts. The size of the vocabulary is controlled by the parameter **“min_count**”, an integer which consider the word if it appears at least the value of min_count (e.g.: 2, 3, 4, 5). The pairs of words are obtained from the “**window**” parameter, an integer that searches for the words that are before and after a vocabulary word. For example: for the phrase, "The child plays soccer in the park with his friends every day" with a window size of 3, for the word park there are the pairs of words are (soccer, park), (in, park), (the, park), (park, with), (park, his), (park, friends).

  The input to the model is a one-hot encoded vector of dimensions equal to the size of the vocabulary representing the input word. Assuming that our phrase from the previous paragraph belongs to a list of phrases with a vocabulary of 5000 words, to represent the word 'park', the one-hot encoded vector will have 5000 dimensions with a value equal to 0 in all positions except in the position of the word 'park' which will be equal to 1.

  The first layer of the model corresponds to a hidden layer with neurons equal to the "**size**" parameter, which represents the dimensionality that each word will have when represented as a vector (e.g .: size = 100).

Another important parameter in the model is "**negative**", which specifies how many noise words should be drawn (this value is usually between 5-20). This parameter helps to a faster run of the model and also avoid overfitting which improve the quality of the resulting vectors. "**Workers**" parameter also helps to reduce the time of model training. This value depends on the cores of the machine (higher cores, faster running).

The output of Word2vec model is the input vocabulary associating to each word a vector of dimensions equal to the size parameter, with a softmax activation. "*The presence of softmax means that the model will actually output probabilities for 5,000 words. This probability is the probability of the word at that index being the nearbuy or a context word for the input/current word. Intuitively, words that occur near the input word multiple times in the corpus will have a larger probability than others.*" (https://kushalj001.github.io/black-box-ml/word2vec/glove/word-embeddings/nlp/2019/11/13/Understanding-Word-Embeddings.html#Word2Vec).

A graphical representation of how the Word2vec model perfroms the word embedding is presented in Figure 2.

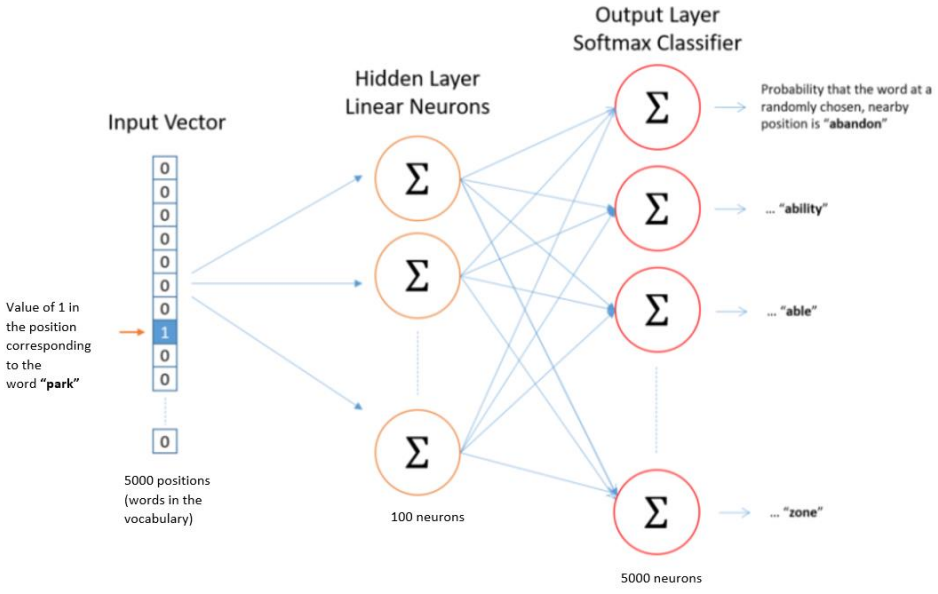



*Figure 2 - Architecture of Word2Vec Model*

The architecture of the model is presented in two configurations: Continuous Bag of Words (CBOW) and Skipgram. CBOW architecture predicts a target word based on the context or surrounding words, while Skipgram architecture estimates the context

from a center word (Al-Saqqa & Awajan, 2019). The graphical architecture of CBOW and Skipgram is presented on Figure 3.

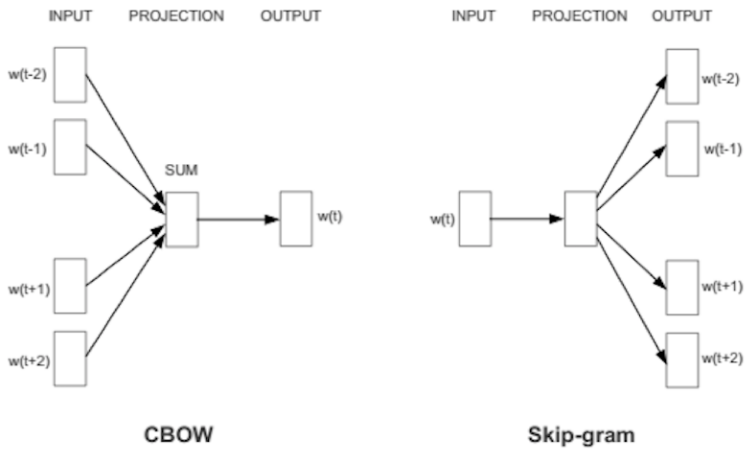



*Figure 3 - Architecture of Word2Vec CBOW and Skipgram*
*Source: Mikolov et al. (2013)*

Gensim is a open source library for representing the Word2Vec word embedding model in python. This library will be applied in this research. In the model you have to specify the architecture of the model. The CBOW architecture is the default model, where the parameter "**cbow_mean**" is equal to 1. To use Skipgram architecture, the parameter "**sg**" must be 1. More information about Word2Vec application with Gensim can be found at https://radimrehurek.com/gensim/models/word2vec.html.

### 3.2.3 Division of Data

The data will be divided into train sets that will be 80% of the whole data set, and test set with the remaining 20%. Cross – validation set is a subset of the train set in order to evaluate the model performance.

### 3.2.4 Text Classification Algorithms

In order to obtain a numerical value that allows to analyze the feature vector, it is necessary to apply a classification algorithm. For this research supervised classification algorithms will be used. The algorithms that will be used to perform NLP with sentiment analysis text classification are:

- **Logistic Regression:** Is a linear model for classification rather than regression and is typically used with a qualitative two – class or binary response (e.g.: positive and negative) (James et al. 2013). The outcoumes are modelled using a logistic function (Sigmoid function) where outputs are between 0 and 1. The Sigmoid function and its curve is presented in Equation 3 and Figure 4.

*Equation 3 - Sigmoid Function - Logistic Regression*

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Where $x_0$ correspond to the x value of the Sigmoid's midpoint (0 in Figure 4). L (1 in Figure 4) is the curve's maximum value and k is the logistic growth rate (1 in Figure 4).

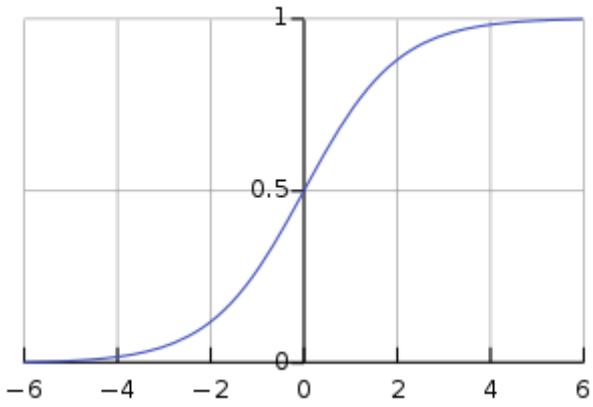



*Figure 4 - Sigmoid Function Curve*

- **Multinomial Naive Bayes:** According to Xu et al. (2017) Is an implementation of the Naïve Bayes algorithm for multinomially distributed data used in text classification. Are also known due to its fast an easy implementation. This algorithm works well when the features describe discrete frequency counts (e.g.: TF-IDF implementation). Scikit-learn webpage mention that "*The distribution in Multinomial Naïve Bayes is parametrized by vectors* $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ *for each class y where n is the size of the vocabulary and* $\theta_{yi}$ *is the probability* $P(x_i \mid y)$ *of feature i appearing in a sample belonging to class y.*"

- **Support Vector Machine (SVM):** Is a generalization of the maximal margin classifier. Is a representation of samples as points in space, mapped, so that samples of separate categories are divided by a hyperplane (solid line on Figure 5) which maximize the margin between the categories or classes. The vector points that touch the magin lines (dash lines on Figure 5) are known as support vectors (James et al. 2013).

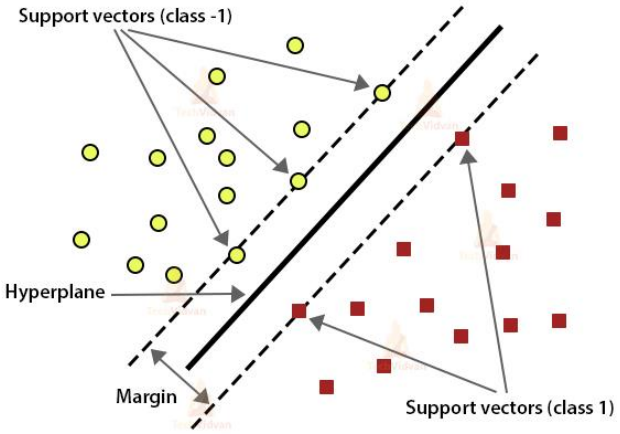



*Figure 5 - Representation of SVM*

Proper selection of parameters 'C' and 'gamma', also known as hyperparameters, is fundamental for improve the performance of SVM model classifier. C controls the wide of the margine and how many samples or observations violate that margin, in other words, the error. James et al. (2013) mentions that when C is low, the margins are narrower adjusting to the data generating less error. Otherwise, with a high C value, the margins are wider allowing for misclassifications, adjusting the data less strictly. Gamma is a parameter that controls the curvature of the hyperplane, the it is not used in linear hyperplanes (linear kernel). High value of gamma means more curvature.

SVM use a set of mathematical functions defined as kernels, that transforms the input data into the required form. Scikit Learn uses the linear, polynomial, radial basis function (rbf) and sigmoid functions. For the proposes of this research only linear and rbf kernels will be applied, then hyperparameter C and gamma will be important in rbf kernel, and C in linear as it does not have curvature. A visualization example of rbf and linear kernels for 3 classes is presented on Figure 6 .

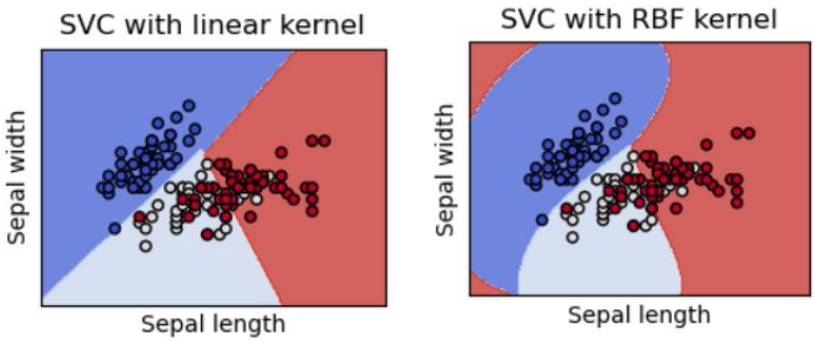



*Figure 6 - SVM linear and rbf Kernels Representation on 3 Classes*

*Source: Scikit Learn webpage: https://scikit-learn.org/stable/modules/svm.html#svm-classification*

To get the best parameters that fits the train data, the GridSearchCV method from Scikit Learn considers all possible parameter combinations specified by the user in the "param_grid" parameter, where the combinations are evaluated and best combination is retained.

In problems where the data is unbalanced the parameter "class_weight" can be used to penalize the class with more values and put more emphasis on the class with lower values. This can be used on Logistic Regression and SVM.

### 3.2.5 Sentiment Analysis Index Scores

The evaluation of the extracted text will be done through a sentiment analysis classification model. For each text extracted a sentiment value between 0 (negative) sentiment and 1 (positive) will be assigned. The criteria of how the sentiment value is assigned is presented on chapter 4.6. The sentiment classification results obtained from the classification algorithms will be compared with the assigned sentiment value in order to evaluate the performance of model.

## 3.3 Analytic Approach

The analythic approach focuses on the following aspects:

- Exploratory analysis of the extracted data in order to identify time series in which there are more or less articles. From the time series (month, year) is important to find the most frequent words by plotting them on wordcloud, associating the words to a solution or a problem present at that moment. Finally, looks to validate the reliability of the

information by comparing the number of articles that words related to the precipitation phenomenon appear in each month with historical precipitation data obtained through Google Earth Engine.

- Identify key actors, variables related to the management of water resources and frequent words in water resources context to improve planning processes according to the community perception. The words related to key actors, variables and frequent words related to water management, also known as water keywords, can be identified from the Named Entity Recognition application on preprocessing.
- Evaluate the sentiment classification model performance with the accuracy, precision, recall and f-score metrics. Kalaivani et al. (2019) mentions that for a given category x the values of this metrics are computed as:

*Equation 4 - Accuracy Metric*

$$Accuracy = \frac{Total\ number\ of\ correct\ predictions}{Total\ number\ of\ samples}$$

*Equation 5 - Precision Metric*

$$Precision = \frac{Samples\ correctly\ classified\ to\ category\ x}{Total\ samples\ classified\ to\ category\ x}$$

*Equation 6 - Recall Metric*

$$Recall = \frac{Samples\ correctly\ classified\ to\ category\ x}{Total\ number\ of\ samples\ in\ category\ x}$$

*Equation 7 - F-score Metric*

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

The accuracy metric is useful when target classes are balanced. Precision – Recall is more useful when the classes are unbalanced. According to Precision and Recall the following 4 cases can presented for each category:

- High precision and high recall: Model handles perfectly the category.

- High precision and low recall: the model doesn't classify the category very well, but when it does it is highly reliable.
- Low precision and high recall: the model classify the category well, but also includes samples of other categories.
- Low precision and low recall: Model don’t classfy the categories properly.

F-score summarized the precision and recall in one metric then, it is also useful in unbalanced data.

This metrics will be obtained applying the Scikit learn metrics classification report and confussion matrix. Model results are compare it with the Spanish Sentiment Analysis (SSA) application performance in order to establish an improvement or not. The SSA was trained using over 800.000 spanish reviews from decathlon, tripadvisor and ebay webpages.

- Associate the concept of basin health with sentiment analysis by grouping the articles as positive or negative according to the sentiment found. The construction of arrays that find in how many articles a pair of words obtained from the Named Entity Recognition process is repeated (eg: a river with a water key word like contamination, or a community, or an authority) problems or solutions that are mentioned more frequently in pulic media can be found and ranked. Through this analysis, factors related to the ecological state of the basin, sanitation, water quality and supply, important in the development of the Bolivia WATCH program, can be identified.

# Chapter 4 Case Study

Bolivia is one of the Countries with the highest reserves of Water Resources worldwide, being the head of the Amazon basin (with a surface area of 718,137 $km^2$) and La Plata basin (with a surface area of 226,268 $km^2$). In addition, it has a closed or endorheic basin in the Altiplanic part with a surface area of 154,176 $km^2$. (Microregiones y Fronteras – ADEMAF, 2015). According to World Water Assessment Programme – UNESCO, Bolivia is the largest reserve of water resources in Latin America (No. 16 worldwide), however, in the classification of the quality of the resource it is ranked No. 67 out of 122 States, which highlight a problem in the resource management.

Agramont et al. (2019), post that Bolivia faces several challenges related to the Integrated Water Resources Management. Hydroclimatic risks like floods and droughts in different seasons of the year, cultural differences, social inequalities, inefficient water services, inequality in resource allocation, claims of limits between communities, water use for energy generation, and the lack of environmental policies are some of the reasons that generate conflicts and complicate the open exchange of information between stakeholders. Additionally, mining and agriculture greatly influence the decisions related to the water resources management, as these are the two main economic activities in the country.

River basin committees and programs in Bolivia have been created, where decisions are made based on a planning system according to the implementation of Bolivia National Watershed Plan. The Bolivia WATCH program is led by the Stockholm Environment Institute (SEI) and the Ministry of the Environment and Water (MMAyA for its Spanish initials) with the purpose of linking the water resource sanitation with the watershed management.

Robust Decision Support (RDS) practice has been implemented in the region as a guiding tool for the water resource planning and decision-making process, obtaining potential results (Purkey et al. 2018). Nevertheless, this process is time-consuming, requires the active and continuous participation of the stakeholders and the strategies that are obtained from this practice are limited by the imagination of the participants.

In that order of ideas, the Bolivia WATCH program, which promotes the implementation of Bolivia National Watershed Plan, seeks for innovative tools and solutions to improve water resources management in the country. The use of Natural Language Processing in unexplored information sources (as news articles) can complement the characterization of a basin and allow its problems to be understood as well as possible. Sentiment Analysis looks for the positive, neutral or negative perception of a text and helps to rank frequent topics.

## 4.1 La Paz – Choqueyapu River Basin

The case study will focus on La Paz – Choqueyapu river basin which is one of the 14 basins under Bolivia WATCH program, and in consequence of Bolivia National Watershed Plan.

The river basin has an area of 489 km$^2$, and the length of its main channel is 44 km. The source of the river is the Pampalarama lagoon, and it crosses the center and the south of the city of La Paz - El Alto until its mouth in the Achocalla River. The river is characterized by being polluted due to waste and pollution generated by being the central receptor of wastewater in the city of La Paz (Ohno et al. 1997). The main municipalities in the basin are La Paz, El Alto, Achocalla and Mecapaca. Each Municipality is composed of a certain number of communities. A schematic of the basin is presented in Figure 7.

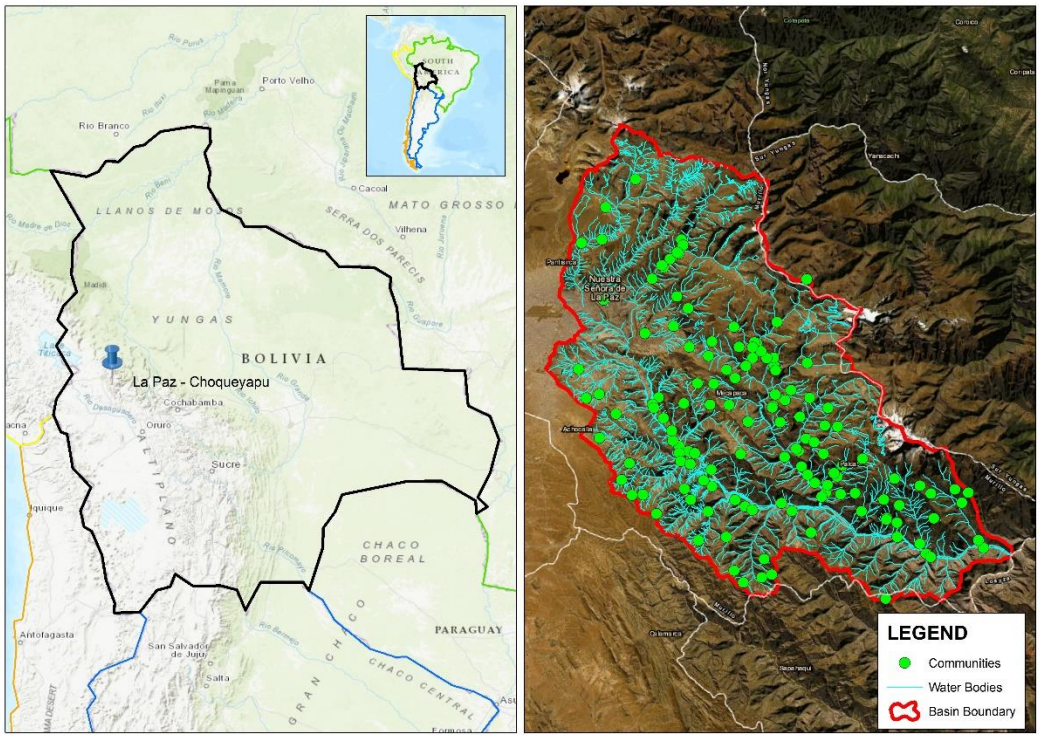



*Figure 7 – La Paz Choqueayapu Basin*

## 4.2 Precipitation in the Last Decade

Rivers in Bolivia are highly vulnerable to climate change, with a high flow in rain seasons causing floods, loss of crops and livestock, damage to hydraulic structures and water supply systems, among others. Otherwise, they have a low flow in drought seasons, which affects the demand for water supply necessary to satisfy the community and local activities.

In order to know the precipitation in the basin in the last decade, the historical data of the product "3B42" provided by the Tropical Rainfall Measurement Mission (TRMM) were obtained using Google Earth Engine, which estimates precipitation in mm/hr with a temporal and spatial resolution of 3 hours and 0.25 degrees (Gavilan et al. 2019), for the period between January 2010 and December 2019. From these historical data, the maximum value was obtained in each month for each year and a multi-year monthly average was made. The results obtained are presented in Figure 8:

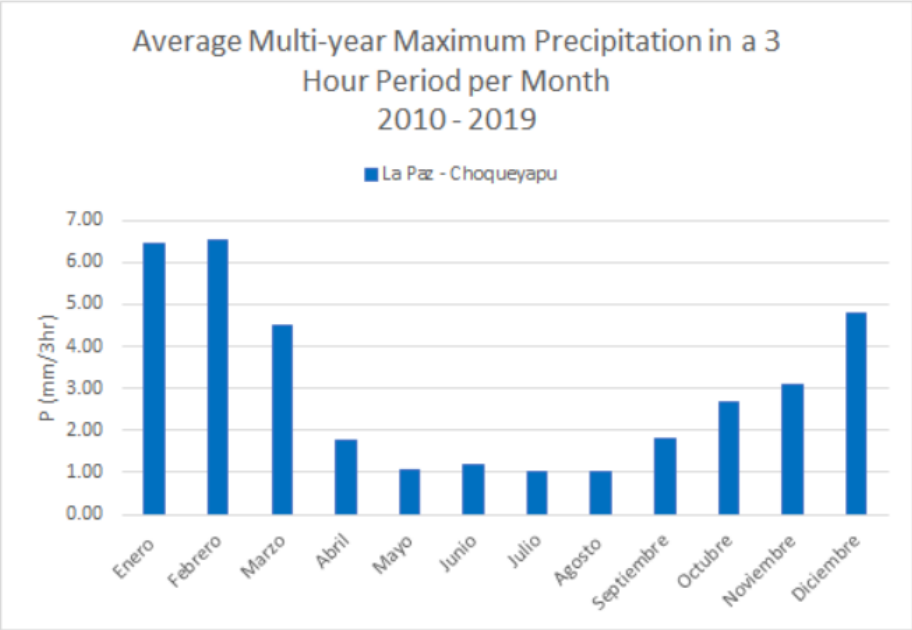



*Figure 8 – Average Precipitacion over the Last Decade. La Paz – Choqueyapu River Basin*
*Source: Product 3B42 of the Tropical Rainfall Measurement Mission. Use of Google Earth Engine*

Figure 8 shows that the hydrologic year starts on october, presenting a rainy period between October and March and a dry period between April and September. The extreme rainy season occurs between January and February.

## 4.3 Sources of Digital Information

The source of digital information applied in this research corresponds to 6 Bolivian newspapers which present information of the different hydroclimatological risks in the country, water management, water supply, as well as articles related to the different projects that have been developed in the country related to water resources. The newspapers and their corresponding webpage are:

- El Pais: https://elpais.bo/
- El Periódico Digital: https://www.elperiodico-digital.com/
- El Potosí: https://elpotosi.net/
- Correo del Sur: https://correodelsur.com/
- La Razón: https://www.la-razon.com/
- Pagina Siete: https://www.paginasiete.bo/

It is important to clarify that the information collected will be applied for the academic purposes of this research and will not compromise confidential information from newspapers and journalists.

## 4.4 Extraction and Storage of Information

The information extraction was performed using the web scrapping technique applying the BeautifulSoup4 and Selenium libraries in the Python programming language on the mentioned newspapers. Search of keywords related to water resources, water bodies in the basin and the name of the municipalities and communities in the basin under study is implemented on each newspaper webpage. The list of keywords is presented in Appendix B. The words that did not yield results in any of the 6 newspaper are not in this list.

The extracted articles were stored in .csv format and consist of 5 characteristics: Article Title, Content, Link, Date and Newspaper from which it was extracted.

Once the information was stored, those duplicate articles were eliminated. Finally, a manual inspection of the data was performed removing those articles that were not related to the context under study (e.g.: Articles related to sports, music, cooking, etc.).

## 4.5 Exploratory data Analysis

The exploratory data analysis consists of analyzing the characteristics found in the database, for example, the total number of samples, how many samples belong to a particular characteristic, etc.

Figure 9 shows how the articles are divided according to the newspaper source. The newspapers with the highest number of articles are La Razon with 1468 corresponding to 43.3% of the total information and Pagina Siete with 34% (1154 articles). The remaining percentage less than 25% correspond to the 4 remaining newspapers. El País with 391 articles (11.5%), El Periódico with 185 (5.5%), Correo Del Sur with 113 (3.3%) and finally with 2.4% is El Potosí with 82 articles. The total number of samples corresponds to 3393 articles.

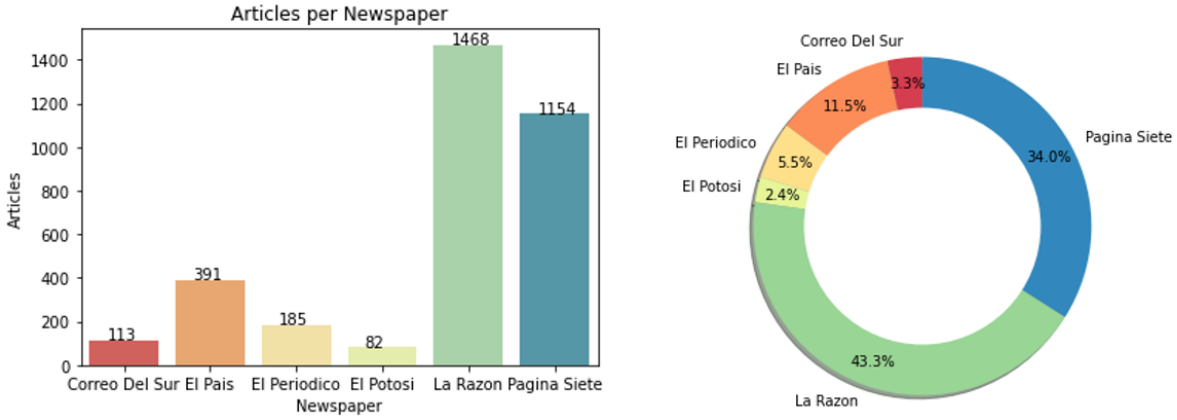



*Figure 9 - Articles Distribution per Newspaper*

If an article count is made by date, we can identify the period of time were the information is presented (Between April of 2010 and February of 2021). It is highlighted that at the end of 2016 there is a day in which more than 25 articles. At the beginning of 2018 there is a date with more than 15 articles. In the second part of the decade 2010 - 2020 it is found that at the beginning and end of each year the number of articles is between 5 and 10, while for the first part of the decade the articles are less than 5. The information described is presented graphically in Figure 10.

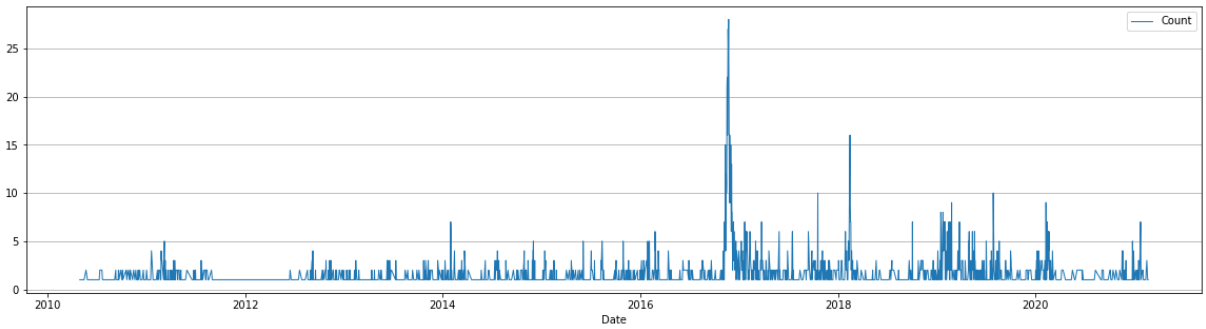



*Figure 10 - Articles Distribution by Date*

Grouping the number of articles per year, it is observed in Figure 11 that the year with the most articles is 2016 with 717 articles, followed by 2019 with 542, and 2017 with 454. The years 2013, 2014, 2015, 2018 and 2020 present between 200 and 300 articles each. 2011 and 2012 present 175 and 130 articles respectively, while 2010 and 2021 have less than 100 as these years don't have information from all the months.

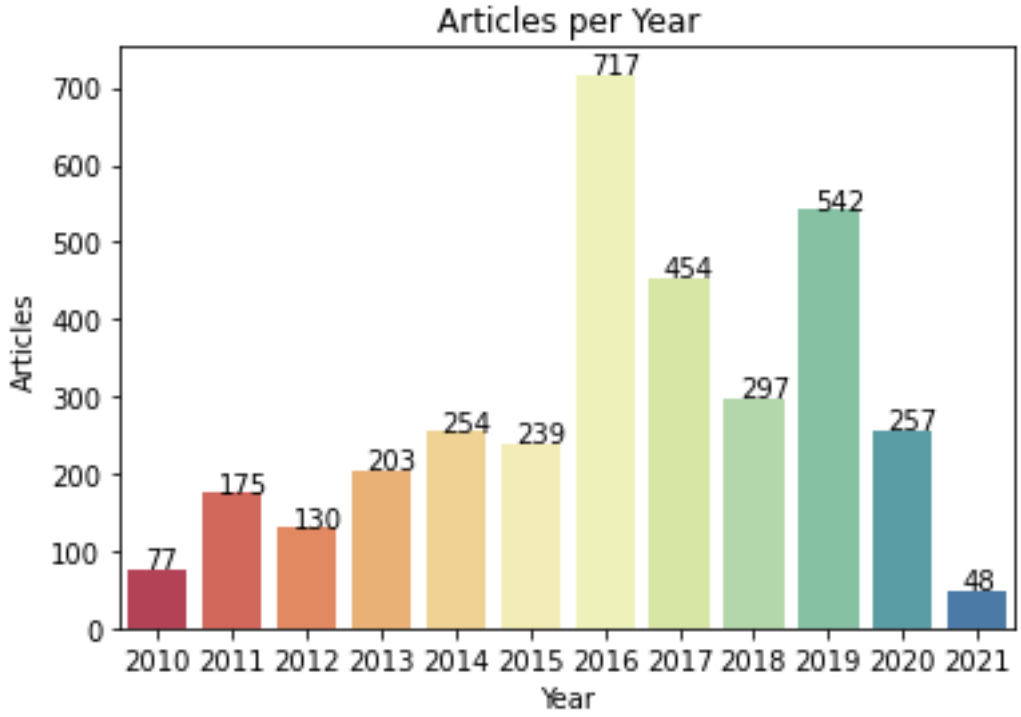



*Figure 11 - Aritcles Distribution per Year*

The month with the highest number of articles is November with 585, continued by February with 447 and January with 435 data according to Figure 12. Of this figure, two periods also stand out, between October and March with more than 200 articles per month, and between April and September with less than 200 articles except for July. This graph is quite similar to

the multiannual monthly precipitation presented in Figure 8, where rainy and dry periods were identified, finding a direct relationship between the information obtained and the precipitation in the basin.

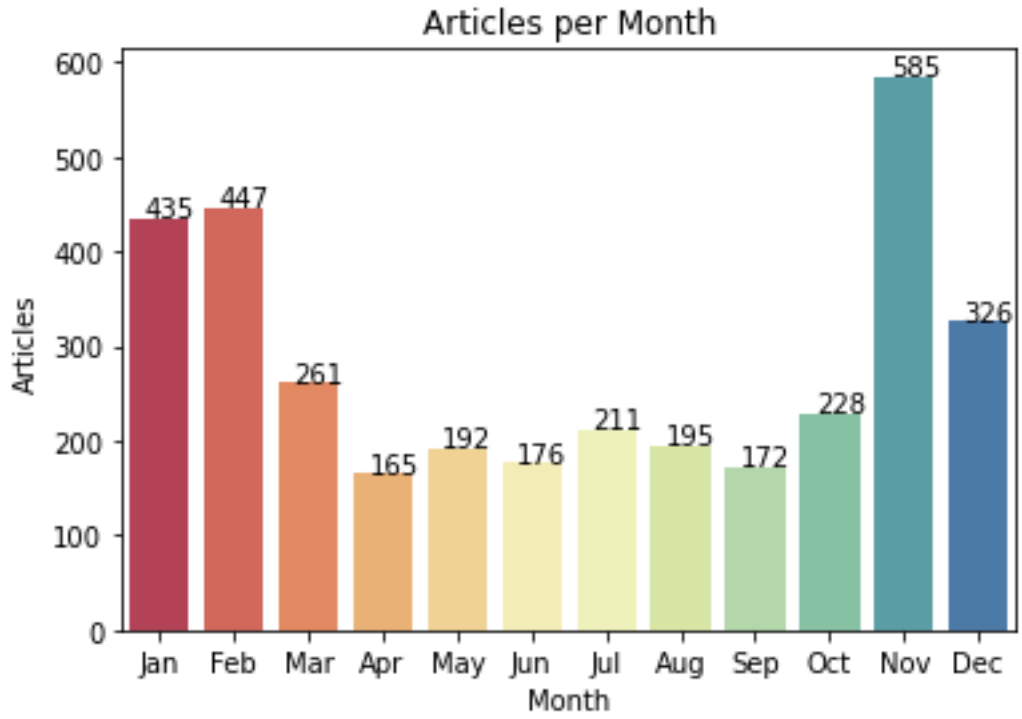



*Figure 12 - Aritcles Distribution per Month*

To reinforce the idea presented in the previous paragraph, a very useful application of exploratory analysis is to associate a topic with a period of time. For example, searching words related to precipitation ("lluvia", "lluvias", "precipitación", "precipitaciones") by counting the number of articles in which any of these words appears and associating it with a period of time such as month, months were precipitation topic is frequently mentioned can be identified. Figure 13 shows the frequency in which the precipitation topic is mentioned per month. Comparing Figure 8 and Figure 12 it is clear that the sources of information explored present reliable information on the description of a hydroclimatological phenomenon in a basin, such as precipitation, clearly identifying the periods of rain (between October and March), drought (between April and September ) and detailing extreme rain events (January and February).

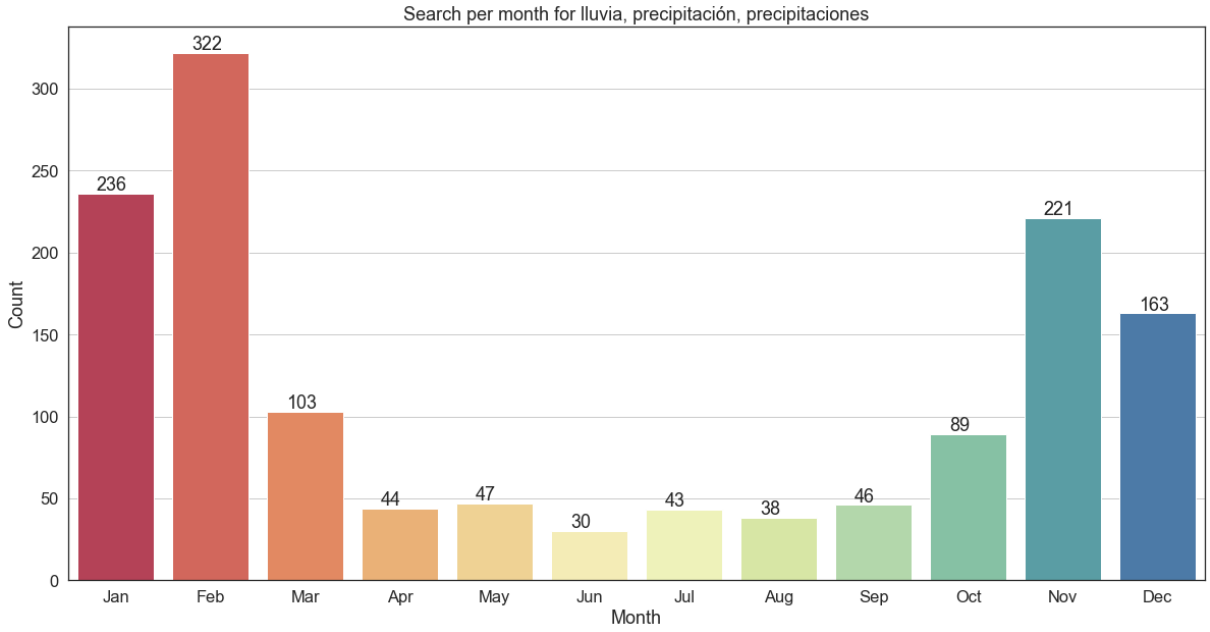



*Figure 13 - Looking for Words Related to Floods in the Whole Dataset. Distribution by Month*

Finally, if the database is associated by year and month as illustrated in Figure 14, it is observed that November 2016 is the period with the most articles with 354, followed by December 2016 with 136 and January 2019 with 101.

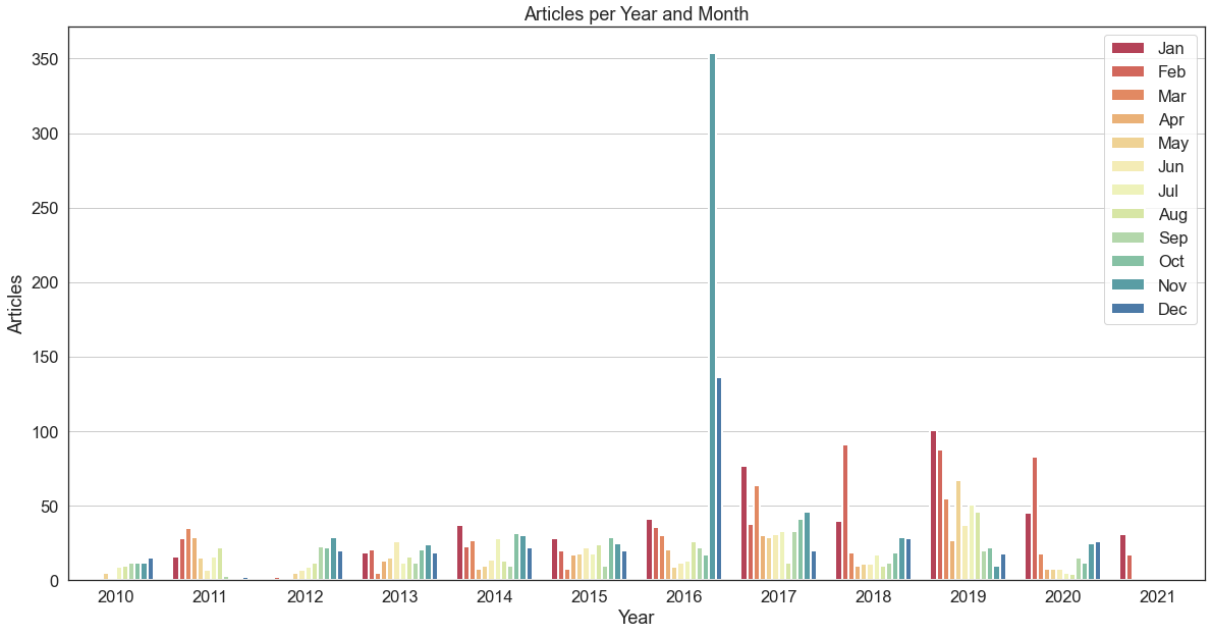



*Figure 14 - Articles per Year and Month*

In order to get the most frequent topics in November 2016 period, a wordcloud will be held, which is presented in Figure 15.


*Figure 15 - Wordcloud of Period November of 2016*

From the wordcloud presented for period November 2016, keywords such as agua (water), EPSAS corresponding to the authority in charge of water service in the region, the communities of La Paz and El Alto, represa (dam), racionamiento (rationing), crisis, servicio (service), agua potable (drinking water), emergencia (emergency), problema (problem), Lluvia (rain) and sequía (drought) stand out. From these words it can be inferred that for this period the most frequent topic was water shortages due to lack of rain (or drought), which led to a water crisis and emergency causing water rationing in the municipalities of La Paz and El Alto. Using this method, frequent words can be identified over a period of time. According to this, for each month and each year a wordcloud was made which is present in Appendix C.

## 4.6 Add of Sentiment Value of each Text

In order to perform the sentiment classification model, it is important to assign manually the target value to each sample, in this case, the sentiment. For this research, the following criteria is considered to classify an article as positive (value = 1) or negative (value = 0):

- **Positive sentiment (1):** Articles focused on: Development and investment in projects related to water resources management and the environment. Reduction of contamination, threat and risk indexes. Monitoring of water sources and hydroclimatological phenomena. Precipitation events after a long drought or that pose

no threat to communities. Water distribution network coverage expansion. Increasing hours of water distribution. End of rationing. Increase in the level of reservoirs, dams and water bodies without risk of overflows. Education, water care and water security projects and policies. Citizen participation in decision-making. Innovation. Effective solution to a problem such as repairing damaged pipes. High water quality index.

- **Negative Sentiment (0):** Articles focused on: Project overcosts. Mismanagement of the authorities in charge of the water resource. Increase in the levels of contamination, risk and levels of water bodies with risk of overflows. Floods that generate personal, material and economic damages. Declaration of yellow, orange or red alert or state of emergency. Overflows, landslides, electrical storms and hydroclimatological phenomena that represent a risk in the community. Rationing, shortage of drinking water, water scarcity. Community disagreement in the management of the resource. Decrease in the hours of water supply. Invasion of mining in water sources. Ineffective solutions. Broken pipes, ducts. Low resilience to an event. Low water quality or with presence of bacteria.

From the manual sentiment classification process, it is obtained that 68.2% of the articles have a negative classification (2315 articles), while 31.8% (1078 articles) present a positive classification (Figure 16).

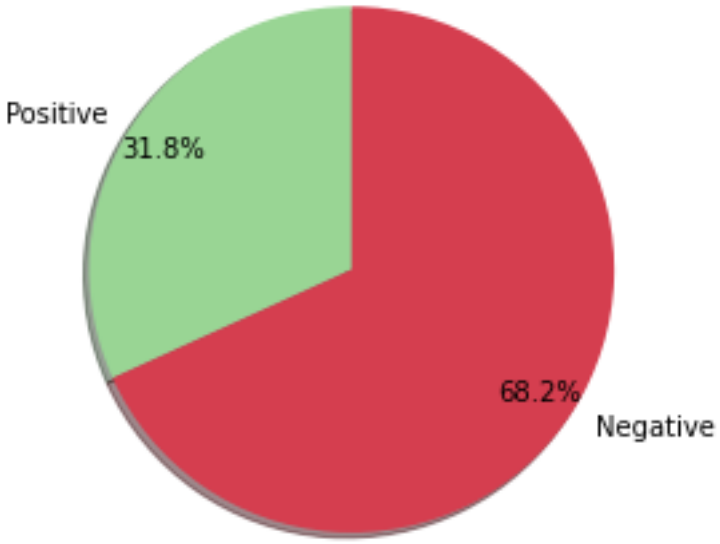



*Figure 16 - Distribution of Sentiment Analysis Classification over the Dataset*

## 4.7 Preprocessing of Information

### 4.7.1 Application of Named Entity Recognition

Identifying key words and actors, water bodies, water systems such as dams or reservoirs, names of municipalities and communities that belong to the basin as well as authorities in charge of water resources management in La Paz - Choqueyapu basin is one of the main tasks of this investigation. The result of applying Named Entuty Recognition (NER) in the articles database is a set of words that can be categorized into each of the entities described above.

One of the advantages of this method is that it allows to identify similitudes of how an entity is named (e.g .: Ajuankhota reservoir can be found also as Ajankota, Ajhuan Cota, Ajuan Khota, Ajuancota, Ajuankkota, or Ajuankota). However, many of the entities obtained through NER may not be applicable to the case study. The identification of similarities between entities and those useful for the case study must be performed manually according to the knowledge of the researcher. Those entities composed of two or more words (e.g .: La Paz, El Alto, Río Choqueyapu), were joined as a unigram replacing the spaces with the underscore symbol (La_Paz, El_Alto, Río_Choqueyapu). The synonyms and compound words were transformed to their original name or unigram as appropriate in the content of the articles.

From NER application on the data set, 13 authorities, 14 municipalities, 64 communities, 15 water systems related to dams and reservoirs, 30 water bodies (rivers, lagoons) and 262 water keywords were identified.

The list of keywords, key authorities, water bodies (rivers and lagoons), water systems (reservoirs, dams), municipalities and communities identified by this method and applicable to the case study for the different analyisis is presented in Appendix D.

### 4.7.2 Non – Alphabetic Characters Removal

Non - Alphabetic characters such as punctuation symbols, emojis, numbering symbols, digits, among others, do not add any value to the content of the text for analysis and are therefore removed. However, the underscore symbol (‘_’) must be preserved as it is used to unify compound words in unigrams. The non – alphabetic characters removal was performed with the regular expressions (REGEX) python library.

### 4.7.3 Tokenization

Words were transformed into tokens spliting them by spaces on each article content. The results is a list of tokens.

### 4.7.4 Stopwords Removal

NLTK python library contains a list of stopwords applied in Spanish language, however, this list contains some negation and affirmation words (‘no’, ‘si’) that change the meaning of a text. In consequence, these words were removed from the stopwords spanish list. The result of this process is a list of tokens without stopwords.

### 4.7.5 Word Embedding

Word Embedding is the last step of preprocessing and it consists of transform the tokens into something understandable for a machine in this case a vector represented by numbers. It was performed by Term Frequency – Inverse Data Frequency and Word2Vec (Continuous Bag Of Words and Skipgram). The parameters applied to each model are presented on Table 1. Detailed information of the patameters was presented on chapter 3.2.2.

*Table 1 - Parameters of Word Embedding*

| **Parameter** | **Model** | | |
|---|---|---|---|
| | **TF-IDF** | **Word2Vec CBOW** | **Word2Vec Skipgram** |
| analyzer | word | - | - |
| min_df | 2 | - | - |
| use_idf | True | - | - |
| smooth_idf | True | - | - |
| min_count | - | 2 | 2 |
| window | - | 5 | 5 |
| size | - | 100 | 100 |
| negative | - | 5 | 5 |
| workers | - | 12 | 12 |
| cbow_mean | - | 1 | 0 |
| sg | - | 0 | 1 |

The obtained feature vector from the TF-IDF embedding is a vector of dimensions 3393 x 21373, where 3393 represent the total amount of articles and 21373 the different unique words found according to the stablished parameters. The content of the feature vectors are numbers representing the TF-IDF word score and will be only present if the word is present in the article and in the TF-IDF Vocabulary, otherwise, a value of 0 is assigned. The TF-IDF score is a float positive number.

The Word2Vec Embedding results consists of a vocabulary with all the words presented according to the parameters, for this case, the word must appear at least 2 times in the whole articles dataset. For each word a vector of dimensions equals to the size parameter (100) is

created. As Word2vec is a method that maintains the semantics of the text contrary of TF-IDF, words with similar meanings or contexts will have similar values in the 100-dimensional vector. An example with the word 'Tubería' (pipe) is presented on Table 2 where the 10 most similar words regarding to it and its similarity score.

*Table 2 - Similarity of word Tuberia by Word2Vec CBOW and SG*

| **Similarity for word 'Tubería' (Pipe)** | | | |
|---|---|---|---|
| **Skipgram** | | **CBOW** | |
| **Similar Word** | **Score** | **Similar Word** | **Score** |
| 'acero' | 0.762134433 | 'ducto' | 0.743422627 |
| 'tendido' | 0.75368166 | 'cañería' | 0.730845928 |
| 'ducto' | 0.719509125 | 'tendido' | 0.699697971 |
| 'pulgadas' | 0.700049043 | 'acero' | 0.685598612 |
| 'soldadura' | 0.698281765 | 'soldada' | 0.650083303 |
| 'dn' | 0.697483122 | 'tubo' | 0.643033624 |
| 'pvc' | 0.693754673 | 'pulgadas' | 0.633140564 |
| 'soldado' | 0.674282491 | 'tubos' | 0.625089347 |
| 'soldada' | 0.674243093 | 'uniones' | 0.621035218 |
| 'fierro' | 0.668003082 | 'fierro' | 0.599342048 |

From results of Table 2 it can be said that all the words that appear there in both skipgram and CBOW, are related to, or are part of the context of the word "Tubería", for which both models maintain the semantics of the words and keep relations between words that belong to the same context.

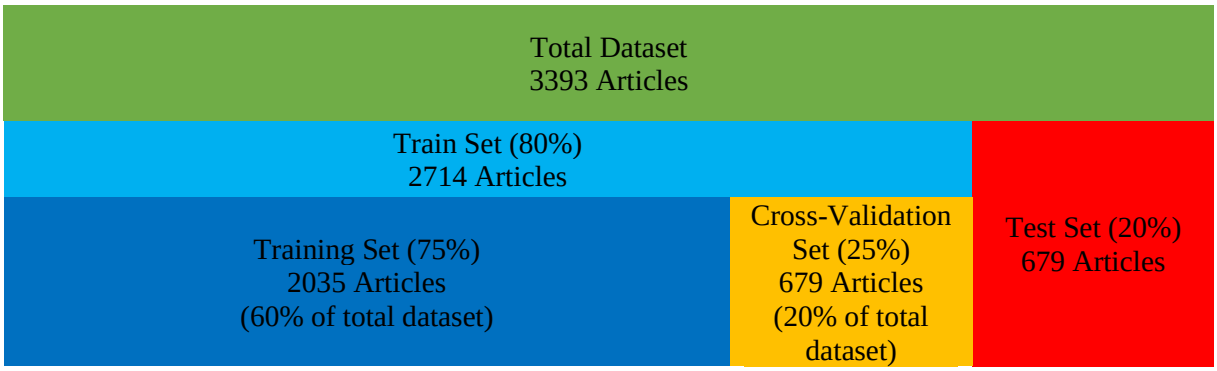
Word2Vec feature vector consists of 3393 rows x 100 columns. Rows represent articles as in TF-IDF embedding, while the columns represent the mean vector of the article and is obtained as the mean value of each word vector that is present on the article.

## 4.8 Train, Cross – Validation and Test Set Division

Dataset articles were initially divided into random train and test subsets applying the scikit learn model selection 'train_test_split'. The train set corresponds to 80% of the information while the test set 20%. After this, the train set was divided into two sets, Training and Cross - Validation. Training set is employed to train the machine learning algorithm in the classification model and corresponds to 75% of the train set. The remaining 25% of the train set corresponds to the Cross - Validation set which will be applied to measure the accuracy of the model. If results are not accurate, the calibration of parameters is required until a desired

model is obtained. Once the results of the model are satisfactory, the test set is applied to evaluate the model. Table 3 presents how the data was splited.

*Table 3 - Divide of Train, Cross-Validation and Test Sets*




## 4.9 Sentiment Classification Model Setup

The classification algorithms that will be used to train the sentiment analysis classification model are Logistic Regression, Multinominal Naive Bayes (Only applies to the TF-IDF Word embedding since it does not accept negative input values) and Support Vector Machines (SVM ) evaluating linear and Radial Basis Function kernels.

The input data are the feature vectors obtained using TF-IDF and Word2Vec - CBOW and Word2Vec - Skipgram word embeddings with their corresponding sentiment target value. The input format corresponds to arrays. An example of the input data is presented in Table 4 and Table 5.

*Table 4 - Representation of TF-IDF Input Vector*

| **TF - IDF Word Embedding X data** | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | …. | 8.03115 | 0 | 0 |
| 0 | 6.59576 | 0 | …. | 0 | 0 | 0 |
| 0 | 0 | 0 | … | 0 | 0 | 5.03542 |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| 3.29788 | 0 | 0 | … | 8.03115 | 0 | 0 |
| 0 | 0 | 0 | … | 0 | 2.38764 | 0 |
| 0 | 6.59576 | 0 | … | 0 | 0 | 5.03542 |
| 3.29788 | 0 | 0 | … | 8.03115 | 2.38764 | 0 |
| **dim (2035 x 21373)** | | | | | | |

| **Target Value Y data** |
|---|
| 0 |
| 0 |
| 0 |
| : |
| : |
| : |
| 0 |
| 1 |
| 0 |
| 0 |
| **dim (2035 x 1)** |

*Table 5 - Representation of Word2Vec Input Vector*

| **Word2Vec Word Embedding X data** | | | | | | | **Target Value Y data** |
|---|---|---|---|---|---|---|---|
| -0.03285806 | 0.08743405 | -0.02374505 | …. | -0.02811376 | -0.08435487 | -0.12757765 | 0 |
| -0.09437709 | 0.0562713 | -0.02377511 | …. | 0.011602 | -0.11107008 | -0.09833226 | 0 |
| -0.03649909 | 0.11317812 | 0.01186433 | … | -0.01121286 | -0.06595349 | -0.07811417 | 0 |
| : | : | : | : | : | : | : | : |
| : | : | : | : | : | : | : | : |
| : | : | : | : | : | : | : | : |
| -0.01998022 | 0.08355556 | -0.0284026 | … | 0.0285013 | -0.08046184 | -0.12759428 | 0 |
| -0.01665204 | 0.06114158 | -0.01822457 | … | 0.02724764 | -0.07967509 | -0.12460267 | 1 |
| -0.05255925 | 0.0679699 | -0.04439014 | … | -0.0159146 | -0.05502792 | -0.07844597 | 0 |
| -0.05150954 | 0.07702911 | -0.02758045 | … | 0.02763231 | -0.05003128 | -0.10796028 | 0 |
| **dim (2035 x 100)** | | | | | | | **dim (2035 x 1)** |

Initially, the models will be trained with the default values in order to obtain an initial accuracy value. Subsequently, a balance will be applied to the data (class_weight parameter) taking into account the wide difference that exists between negative and positive sentiment values in the dataset. Finally, the selection model of Scikit learn 'GridSearchCV' will be applied evaluating different parameters (e.g .: C and gamma on SVM) in order to obtain those that best fit the training set. Once the final parameters of each algorithm have been defined after an evaluation with the cross-validation set, they will be evaluated with the test set. Finally, the results will be compared with those obtained by the Spanish Sentiment Analysis library.

# Chapter 5 Results and Discussion

## 5.1 Sentiment Classification Model

Firstly, evaluation of the model with cross validation set intended to perform calibration and select the best model. Then, set data is used.

### 5.1.1 Selection of Model – Use of CV Set

Training set was used to build the model. Cross-Validation set is used to evaluate the model on this stage. Evaluation metrics results of the performed sentiment classification models build from the machine learning classification algorithms with the three types of embeddings are presented on Table 6. For the precision, recall and F-score metrics the macro average value is applied, which considers equal weight to positive and negative category.

*Table 6 - Sentiment Classification Model Metrics for CV-Set*

| Algorithm | Embedding | Accuracy | Precission | Recall | F - Score |
|---|---|---|---|---|---|
| **Logistic Regression** | TFIDF | 0.79 | 0.77 | 0.75 | 0.76 |
| | Word2Vec - CBOW | 0.78 | 0.78 | 0.69 | 0.71 |
| | Word2Vec - SG | 0.76 | 0.78 | 0.66 | 0.67 |
| | TFIDF Balanced | 0.79 | 0.76 | 0.76 | 0.76 |
| | Word2Vec - CBOW Balanced | 0.75 | 0.73 | 0.75 | 0.73 |
| | Word2Vec - SG Balanced | 0.75 | 0.73 | 0.76 | 0.74 |
| **Multinomial Naive Bayes** | TFIDF | 0.77 | 0.74 | 0.74 | 0.74 |
| **SVM- Kernel 'rbf' (C = 1 gamma = auto)** | TFIDF | 0.77 | 0.79 | 0.68 | 0.7 |
| | Word2Vec - CBOW | 0.8 | 0.79 | 0.74 | 0.76 |
| | Word2Vec - SG | 0.81 | 0.79 | 0.75 | 0.76 |
| | TFIDF Balanced | 0.78 | 0.75 | 0.74 | 0.74 |
| | Word2Vec - CBOW Balanced | 0.78 | 0.76 | 0.77 | 0.76 |
| | Word2Vec - SG Balanced | 0.78 | 0.75 | 0.77 | 0.76 |
| **SVM- Kernel 'linear' (C = 1)** | TFIDF | 0.79 | 0.76 | 0.75 | 0.76 |
| | Word2Vec - CBOW | 0.79 | 0.79 | 0.71 | 0.72 |
| | Word2Vec - SG | 0.78 | 0.79 | 0.69 | 0.7 |
| | TFIDF Balanced | 0.79 | 0.76 | 0.75 | 0.76 |
| | Word2Vec - CBOW Balanced | 0.75 | 0.73 | 0.75 | 0.73 |
| | Word2Vec - SG Balanced | 0.78 | 0.75 | 0.77 | 0.76 |

From the results obtained in Table 6 it can be seen that the model that best fits in 3 of the 4 metrics corresponds to the SVM with rbf kernel with an accuracy of 0.81, a precision of 0.79, an F-Score of 0.76 and a value recall of 0.75 that is not very far from the maximum value obtained of 0.77.

In the Logistic regression algorithm, it can be noted that the models with the best score metrics are those with TF-IDF word embedding. Accuracy is higher when balanced data is not applied and decreases when it is applied. The opposite is the case with recall, which increases its value when class weight balance is applied to the data and is very low when it is not. This situation occurs in all classification algorithms. Finally, from the Linear Regression algorithm, it can be seen that Word embedding using Word2vec with both Skipgram and CBOW architectures, does not have the best performance.

The result obtained through the Naive Bayes Multinomial algorithm presents a balanced result in all aspects, without being the highest in any metric, but not the lowest either.

The accuracy and precision of the models build with the SVM algorithm with linear kernels and rbf decrease considerably when class weight balanced is working. The recall value increases and the F-score remains.

Considering that the categories to be predicted are not balanced as shown in Figure 16, where the negative predictions represent 68.2% of the dataset, the F-Score value will be set as the first decision metric (maximise this value), as it keeps a balance between precision and recall. If there is equality between several models, the other metrics will be considered to decide the model with best performance.

In an attempt to calibrate the model, the objective function is to maximise the F-Score value. The GridSearchCV function of Scikit-learn is performed in the Logistic Regression and SVM algorithms, evaluating the hyperparameters presented in Table 7. The values obtained that present the best estimation parameters on each algorithm, for each word embedding are presented in Table 8. Finally, Table 9 presents the evaluation metrics results after calibration with the cross-validation set.

*Table 7 - Hyperparameters Evaluated on GridSearch CV*

| Algorithm | Parameters | Evaluated Values |
|---|---|---|
| **Logistic Regression** | **C** | 0.1, 1, 10, 100, 1000 |
| | **class_weight** | 'balanced', None |
| | **solver** | 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga' |
| **SVM Linear and rbf kernels** | **C** | 0.1, 1, 10, 100, 1000 |
| | **class_weight** | 'balanced', None |
| | **gamma (only for rbf)** | 1, 0.1, 0.01, 0.001, 0.0001 |

*Table 8 – Best Estimated Parameters Obtained from GridSearchCV*

| Algorithm | Embedding | Best Parameters |
|---|---|---|
| **Logistic Regression** | **TFIDF** | C = 1, class_weight = balanced, solver = sag |
| | **Word2Vec - CBOW** | C = 1000, class_weight = None, solver = newton-cg |
| | **Word2Vec - SG** | C = 1000, class_weight = None, solver = sag |
| **SVM - linear kernel** | **TFIDF** | C = 0.1, class_weight = balanced |
| | **Word2Vec - CBOW** | C = 1000, class_weight = None |
| | **Word2Vec - SG** | C = 100, class_weight = None |
| **SVM - rbf kernel** | **TFIDF** | C = 100, class_weight = balanced, gamma = 0.0001 |
| | **Word2Vec - CBOW** | C = 10, class_weight = balanced, gamma = 1 |
| | **Word2Vec - SG** | C = 100, class_weight = balanced, gamma = 1 |

*Table 9 - Sentiment Classification Model Metrics After GridSearchCV Application*

| Algorithm | Embedding | Accuracy | Precission | Recall | F - Score |
|---|---|---|---|---|---|
| **Logistic Regression** | **TFIDF** | 0.8 | 0.78 | 0.77 | 0.77 |
| | **Word2Vec - CBOW** | 0.81 | 0.78 | 0.77 | 0.77 |
| | **Word2Vec – SG** | 0.8 | 0.78 | 0.76 | 0.77 |
| **SVM - linear kernel** | **TFIDF** | 0.79 | 0.76 | 0.75 | 0.76 |
| | **Word2Vec - CBOW** | 0.81 | 0.79 | 0.76 | 0.77 |
| | **Word2Vec – SG** | 0.79 | 0.77 | 0.75 | 0.76 |
| **SVM - rbf kernel** | **TFIDF** | 0.8 | 0.78 | 0.75 | 0.76 |
| | **Word2Vec - CBOW** | 0.79 | 0.76 | 0.78 | 0.77 |
| | **Word2Vec – SG** | 0.78 | 0.75 | 0.77 | 0.76 |

From Table 9 results, five models have an F-Score value equal to 0.77: Logistic Regression with their 3 types of word embedding and SVM linear and kernel model with Word2vec CBOW word embedding. Then, the sentiment analysis classification model will be chosen considering

the other evaluation metrics scores. The SVM linear kernel Word2vec CBOW, presents an accuracy of 0.81 and the highest precision of all the models made with a value of 0.79. The results are similar to those initially obtained using SVM rbf kernel and Word2vec SG (Table 6), however, the recall and the F-Score are slightly higher. Therefore, the representative sentiment classification model for this case study corresponds to the SVM algorithm with linear kernel and word embedding with Word2Vec CBOW architecture.

### 5.1.2 Evaluation of Sentiment Classification Model with Test Set

The results obtained from evaluating the test set in the chosen sentiment classification model are an accuracy of 0.82, and macro average metrics of precision equals to 0.79, recall of 0.76 and F-score of 0.77, similar values than the ones obtained from the cross validation set. Confusion matrix for test set is presented on Table 10 and a summary of the achieved evaluation metrics for each category and the model is displayed on Figure 17.

*Table 10 - Confusion Matrix for Test Set*

| | 0 | 1 | |
|---|---|---|---|
| 0 | 426 | 46 | 472 |
| 1 | 79 | 128 | 207 |
| | 505 | 174 | |

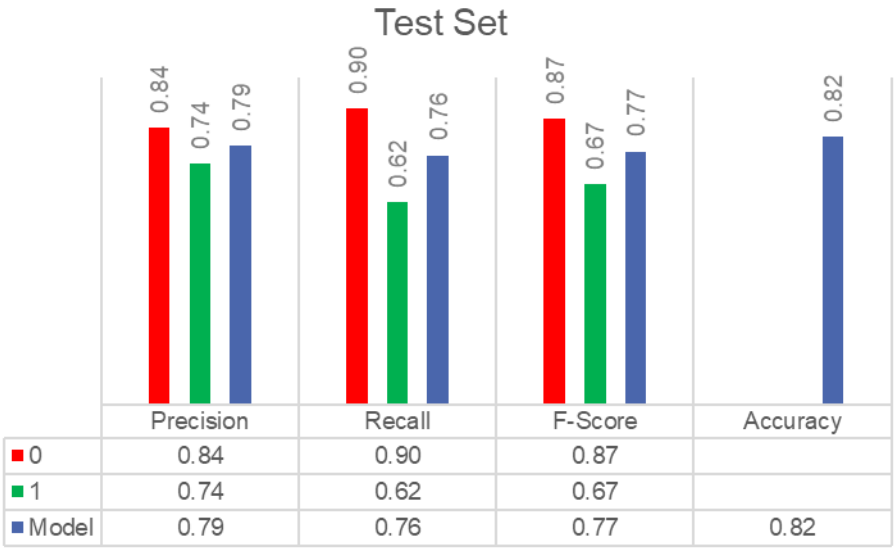



*Figure 17 - Summary of Evaluation Metrics on Test Set*

## 5.2 Comparison: Use of Spanish Sentiment Analysis Library

The Spanish Sentiment Analysis will be applied in the whole data set. The values obtained from the model are between 0 and 1, where a value between 0 and 0.5 represents a negative sentiment, and values higher than 0.5 a positive sentiment. The model chosen from chapter 5.1 corresponding to SVM linear kernel Word2vec CBOW embedding will also be applied to the entire data set. The accuracy metric will be the one that will be compared between both models. The results obtained are presented in Table 11. The results sow that for the whole data set our model has an overall accuracy of 0.84 which is a huge improvement compare to the accuracy obtained by Spanish Sentiment Analysis library.

*Table 11 – Confusion Matrix for Spanish Sentiment Analysis and the Sentiment Classification Model using whole Dataset*

**Spanish Sentiment Analysis**

| | 0 | 1 | |
|---|---|---|---|
| 0 | 2300 | 15 | 2315 |
| 1 | 1066 | 12 | 1078 |
| | 3366 | 27 | |

**Sentiment Classification Model**

| | 0 | 1 | |
|---|---|---|---|
| 0 | 2128 | 187 | 2315 |
| 1 | 360 | 718 | 1078 |
| | 2488 | 905 | |

Under the accuracy results found, it can be deduced that the context in which a model is built greatly affects its performance. Spanish Sentiment Analysis was trained with reviews related to marketing, while the built model contains articles clearly related to water resources.

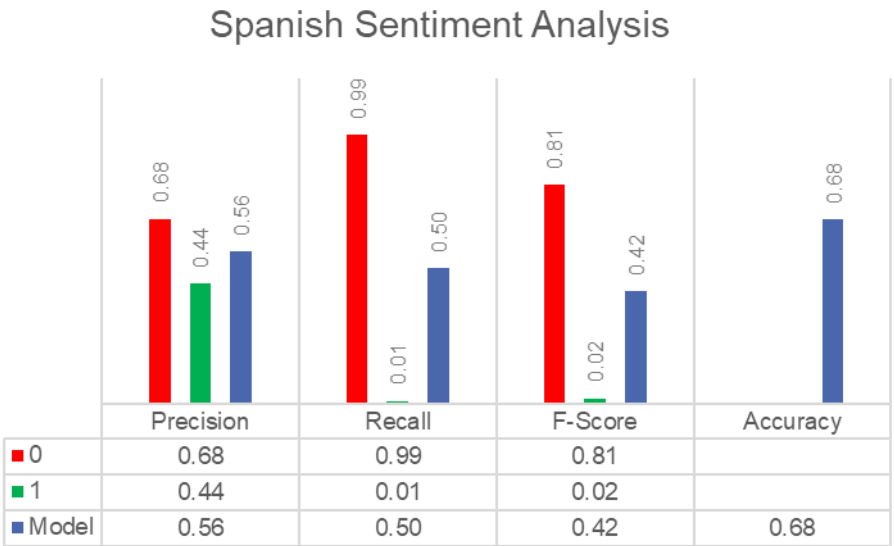



*Figure 18 - Evaluation Metrics for Spanish Sentiment Analysis on Whole Dataset*

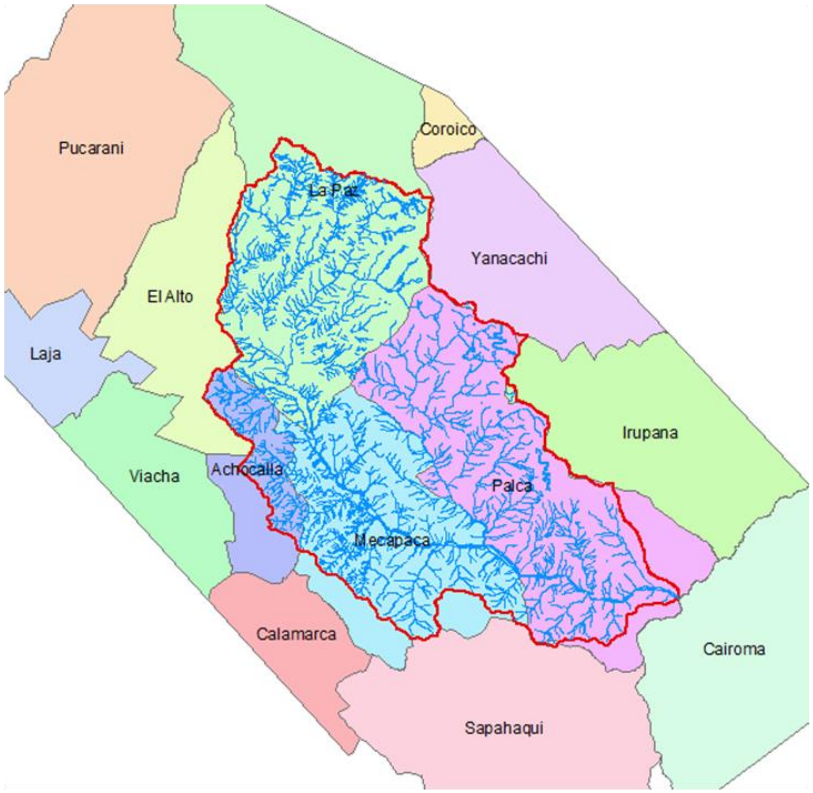

| | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|
| 0 | 0.86 | 0.92 | 0.89 | |
| 1 | 0.79 | 0.67 | 0.72 | |
| Model | 0.82 | 0.79 | 0.81 | 0.84 |

*Figure 19 - Evaluation Metrics for Sentiment Classification Model on Whole Dataset*

## 5.3 Grouping Data According to the Sentiment Value

Finally, by grouping the texts as positive or negative and constructing an array that relates the different groups of entities between them (eg: keywords and communities, keywords and water bodies) depending on whether both words are found in the content of the article, the problems can be identified of greater importance since the articles are classified as negative. Furthermore, the most frequently commented entities can be identified as positive or negative, finding the number of articles in which that entity is found. The results of this section will be presented in the research defense considering the time it takes to run the code that performs these two procedures described above.

| Word | Number of Articles |
|---|---|
| la_paz | 1674 |
| epsas | 676 |
| senamhi | 648 |
| el_alto | 601 |
| emergencia | 568 |
| riesgo | 545 |
| precipitación | 430 |
| agua_potable | 424 |
| inundaciones | 366 |
| desborde | 341 |
| deslizamiento | 330 |
| racionamiento | 297 |
| alerta_naranja | 289 |
| cultivo | 282 |
| desastre | 281 |
| saneamiento | 266 |
| basura | 260 |
| caudal | 257 |
| sequía | 250 |
| crecida | 244 |
| alcantarillado | 238 |




| Word | Number of Articles |
|---|---|
| la_paz | 689 |
| epsas | 355 |
| agua_potable | 310 |
| el_alto | 302 |
| inversión | 244 |
| mmaya | 200 |
| emergencia | 191 |
| racionamiento | 179 |
| saneamiento | 174 |
| hampaturi | 140 |
| planta_de_tratamiento | 134 |
| aaps | 134 |
| tubería | 130 |
| sequía | 127 |
| alcantarillado | 122 |
| riego | 116 |
| riesgo | 114 |
| abastecimiento | 105 |
| senamhi | 102 |
| dotación | 98 |
| precipitación | 92 |

*Figure 20 - Group Data by Sentiment. Look Most Common Words. Watershed Level*

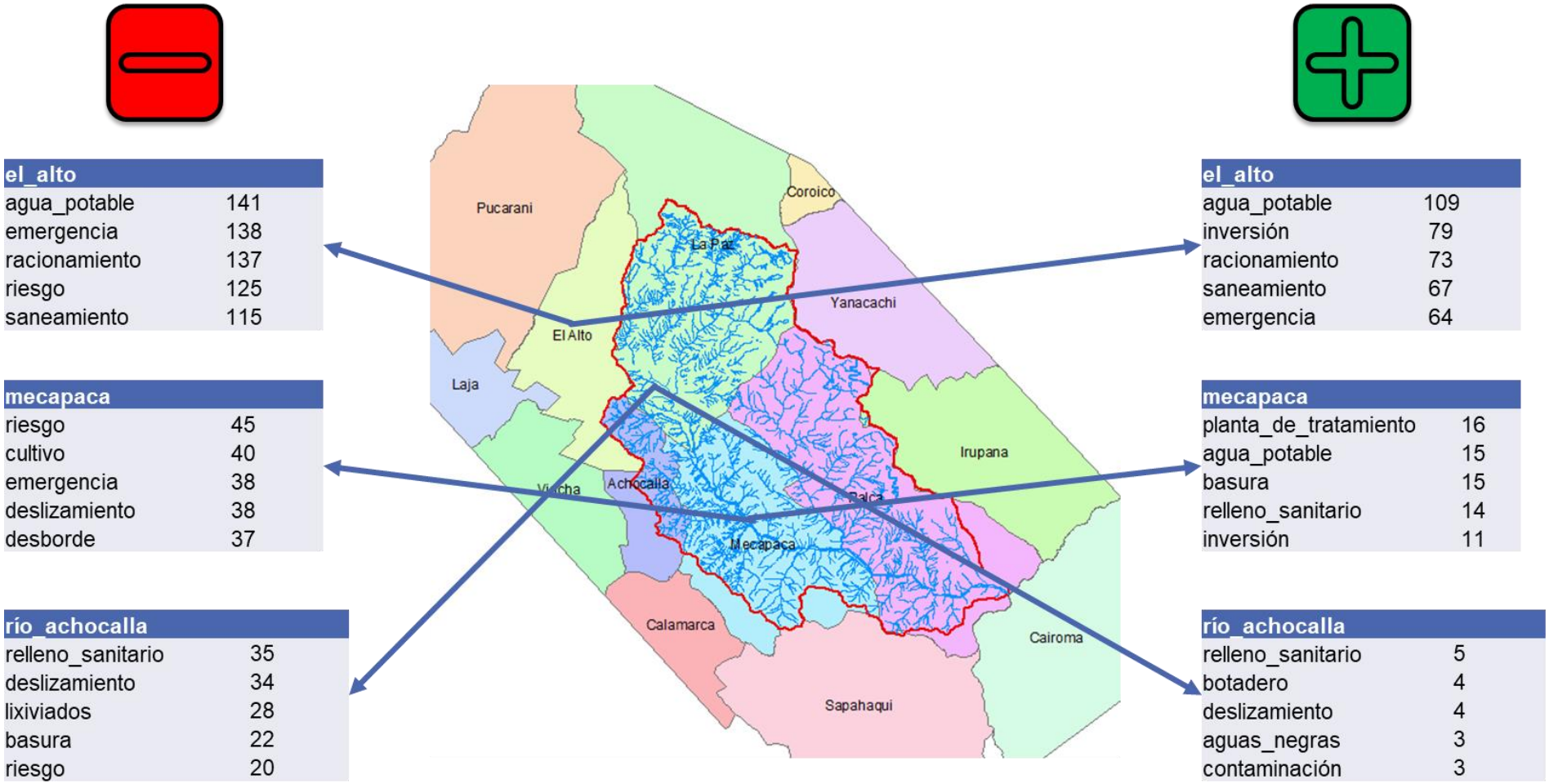



*Figure 21 - Group Data by Sentiment. Look Most Common Words. Detailed Level*

# Chapter 6 Conclusions and Recommendations

This research demonstrate that the information obtained from the digital public media under analysis presents a good level of reliability by associating the content of the articles with the different hydroclimatological events and positive or negative events related to the management of water resources. From this information, temporary patterns of rains, floods, droughts, water shortages, among other aspects can be found.

Supervised model was chosen in this research, therefore, the classification of the texts to a positive or negative sentiment was done manually. The criteria for evaluating whether a text is positive or negative were proposed by this research according to the researcher's experience in the context of water resources. Consequently, it is not strict to follow these parameters, but they do serve as a guide for future researchers.

The sentiment analysis classification model with the best performance obtained from this reseach corresponds to the SVM algorithm with the linear kernel and the Word2vec CBOW word embeding with a value of accuracy in the whole dataset of 84%. This value is much higher than the one found by the Spanish Sentiment Analysis library of 63%, so the build sentiment classification model could have more efficiency applied for articles related to water resources in different Spanish speaking countries. This must be proved in fututе work.

Future work will also be focused on improving the preprocessing of information. Applying techniques such as lemmatization that reduce the dimensionality of the feature vector and improve the performance of the model, as well as the evaluation of the model with articles from different Spanish-speaking countries such as Colombia. Similarly, the use of other classification algorithms such as Deep learning models, Recurrent Neural Networks as Long Short Term Memory, can be evaluated and compared with the values obtained by this research.

Some recommendations are mainly focused on perform a good exploratory analysis with data visualization techinques, on the information helps to understand the context and check if the information is reliable and well distributed. In the particular case of this research, the

negative data almost doubled the positive data, so developing a balance of data before building the model can improve the results of the evaluated metrics.

# References


Agramont, A., Craps, M., Balderrama, M., & Huysmans, M. (2019). Transdisciplinary learning communities to involve vulnerable social groups in solving complex water-related problems in Bolivia. *Water (Switzerland), 11*(2) doi:10.3390/w11020385

Ali Fauzi, M. (2019). Word2Vec model for sentiment analysis of product reviews in indonesian language. *International Journal of Electrical and Computer Engineering, 9*(1), 525-530. doi:10.11591/ijece.v9i1.pp.525-530

Al-Saqqa, S., & Awajan, A. (2019). The use of Word2vec model in sentiment analysis: A survey. Paper presented at the *ACM International Conference Proceeding Series*, 39-43. doi:10.1145/3388218.3388229

Carrera, J. S., Key, K., Bailey, S., Hamm, J. A., Cuthbertson, C. A., Lewis, E. Y., . . . Calhoun, K. (2019). Community science as a pathway for resilience in response to a public health crisis in flint, michigan. Social Sciences, 8(3) doi:10.3390/socsci8030094

Ekenga, C. C., McElwain, C. -., & Sprague, N. (2018). Examining public perceptions about lead in school drinking water: A mixed-methods analysis of twitter response to an environmental health hazard. International Journal of Environmental Research and Public Health, 15(1) doi:10.3390/ijerph15010162

Galvez, V. & Rojas, R. (2019) Collaboration and Integrated Water Resources Management: A Literature Review. *World Water Policy; 5 (179– 191)*. doi:10.1002/wwp2.12013

Gavilan, S., Pastore, J., Uranga, J., Ferral, A., Lighezzolo, R., & Aceñolaza, P. (2019). Metodología operativa para la obtención de datos históricos de precipitación a partir de la misión satelital Tropical Rainfall Measuring Mission. Validación de resultados con datos de pluviómetros. *Revista de la Facultad de Agronomía. 118,* 115-125. doi:10.24215/16699513e011.

Ilyas, S. H. W., Soomro, Z. T., Anwar, A., Shahzad, H., & Yaqub, U. (2020). Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion. Paper presented at the ACM *International Conference Proceeding Series*, 1-6. doi:10.1145/3396956.3396973

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R.

Japhne, A., & Murugeswari, R. (2020). Opinion mining based complex polarity shift pattern handling for improved sentiment classification. Paper presented at the *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020*, 323-329. doi:10.1109/ICICT48043.2020.9112565

Kalaivani, K. S., Kuppuswami, S., & Kanimozhiselvi, C. S. (2019). Use of NLP based combined features for sentiment classification. *International Journal of Engineering and Advanced Technology*, 9(1), 621-626. doi:10.35940/ijeat.F8290.109119

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Paper presented at the *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.*

Murakami, A., Nasukawa, T., Watanabe, K., & Hatayama, M. (2020). Understanding requirements and issues in disaster area using geotemporal visualization of twitter analysis. *IBM Journal of Research and Development, 64*(1-2) doi:10.1147/JRD.2019.2962491

Murphy, J. T., Ozik, J., Collier, N. T., Altaweel, M., Lammers, R. B., Kliskey, A., Alessa, L., Cason, D., & Williams, P. (2014). Water relationships in the U.S. southwest: Characterizing water management networks using natural language processing. *Water (Switzerland), 6*(6), 1601-1641. doi:10.3390/w6061601

Nalini, C., Kharabe, S., & Sangeetha, S. (2019). Efficient notes generation through information extraction. *International Journal of Engineering and Advanced Technology*, 8(6 Special Issue 2), 160-162. doi:10.35940/ijeat.F1041.0886S219

Noga, J., & Wolbring, G. (2013). Perceptions of water ownership, water management, and the responsibility of providing clean water. Water (Switzerland), 5(4), 1865-1889. doi:10.3390/w5041865

Parlar, T., & Sarac, E. (2019). IWD based feature selection algorithm for sentiment analysis. Elektronika Ir Elektrotechnika, 25(1), 54-58. doi:10.5755/j01.eie.25.1.22736

Purkey, D. R., Arias, M. I. E., Mehta, V. K., Forni, L., Depsky, N. J., Yates, D. N., & Stevenson, W. N. (2018). A philosophical justification for a novel analysis-supported,

stakeholder-driven participatory process for water resources planning and decision making. *Water (Switzerland), 10*(8) doi:10.3390/w10081009

Reyes-Menendez, A., Saura, J. R., & Alvarez-Alonso, C. (2018). Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach. *International Journal of Environmental Research and Public Health, 15*(11) doi:10.3390/ijerph15112537

Sharma, S., & Bansal, M. (2020). Real-time sentiment analysis towards machine learning. *International Journal of Scientific and Technology Research*, 9(2), 987-989.

Singh, S., Ahmad, M., Bhattacharya, A., & Azhagiri, M. (2019). Predicting stock market trends using hybrid SVM model and LSTM with sentiment determination using natural language processing. International Journal of Engineering and Advanced Technology, 9(1), 2870-2875. doi:10.35940/ijeat.A1106.109119

Subirats, L., Conesa, J., & Armayones, M. (2020). Biomedical holistic ontology for people with rare diseases. *International Journal of Environmental Research and Public Health*, 17(17), 1-11. doi:10.3390/ijerph17176038

Van Cauwenbergh, N., Ballester Ciuró, A., & Ahlers, R. (2018). Participatory processes and support tools for planning in complex dynamic environments: A case study on web-GIS based participatory water resources planning in Almeria, Spain. *Ecology and Society, 23*(2) doi:10.5751/ES-09987-230202

Wang, Z., Ke, L., Cui, X., Yin, Q., Liao, L., Gao, L., & Wang, Z. (2017). Monitoring environmental quality by sniffing social media. Sustainability (Switzerland), 9(2) doi:10.3390/su9020085

Xiong, J., Hswen, Y., & Naslund, J. A. (2020). Digital surveillance for monitoring environmental health threats: A case study capturing public opinion from twitter about the 2019 Chennai water crisis. *International Journal of Environmental Research and Public Health, 17*(14), 1-15. doi:10.3390/ijerph17145077

Xu, S., Li, Y., & Wang, Z. (2018). Bayesian naïve bayes classifiers to text classification. *Journal of Information Science, 44*(1), 48-59. doi:10.1177/0165551516677946

Zhang, D., Qiang, M., Jiang, H., Wen, Q., An, N., & Xia, B. (2018). Social sensing system for water conservation project: A case study of the south-to-north water transfer project in china. Water Policy, 20(4), 667-691. doi:10.2166/wp.2018.141

Zheng, F., Simpson, A. R., & Zecchin, A. C. (2014). An efficient hybrid approach for multiobjective optimization of water distribution systems. Water Resources Research, 50(5), 3650-3671. doi:10.1002/2013WR014143

# APPENDICES

# Appendix A. - Ethics Approval Letter

# Appendix B. - List of Keywords for Data Extraction

| Keyword |
| --- |
| aaps |
| achachicala |
| achocalla |
| ajuankhota |
| amachuma |
| ananta |
| apaña |
| aramani |
| autoridad jurisdiccional de administración minera |
| avircato |
| ayma |
| cachapa alto |
| cachapaya |
| cahuayuma |
| cairoma |
| calamarca |
| carpani |
| carreras |
| cayimbaya |
| cebollullo |
| chacaltaya |
| challasirca |
| chanca |
| chañurani |
| choquechihuani |
| choquecota |
| chullo oxani |
| cohani |
| cohoni |
| collana |
| comibol |
| cooperativa de servicios de agua |
| coroico |
| cotaña |
| el alto |
| emagua |
| epsas |
| estrellani |
| hampaturi |
| huajchilla |
| huancarani |
| huaricana |
| huayhuasi |
| irupana |
| jalancha |

| Keyword |
| --- |
| khapi |
| kunkahuikhara |
| la enconada |
| la glorieta |
| la granja |
| la paz |
| laguna achocalla |
| laguna ajuankhota |
| laguna chojña khota |
| laguna estrellani |
| laguna milluni |
| laja |
| lipari |
| lluto |
| lorocota |
| lurata |
| mamaniri huerta |
| mecapaca |
| millocato |
| ministerio de medio ambiente y agua |
| palca |
| palcoma |
| palomar |
| pantini |
| peñol |
| pinaya |
| plan director de la cuenca |
| plan nacional de cuencas |
| presa khota khota |
| programa mi agua |
| pucarani |
| quebrada alpacoma |
| represa chacaltaya |
| represa chuquiaguillo |
| represa condoriri |
| represa hampaturi |
| represa incachaca |
| represa la paz |
| represa milluni |
| represa pampahasi |
| represa pampalarama |
| represa tuni |
| retamani |
| río achachicala |
| río achocalla |

| Keyword |
| --- |
| río achumani |
| río alpacoma |
| río aruntaya |
| río choqueyapu |
| río chuquiaguillo |
| río huayñajahuira |
| río irpavi |
| río jilusaya |
| río kaluyo |
| río kantutani |
| río kollpajahuira |
| río la paz |
| río luribay |
| río orkojahuira |
| río palca |
| río palcoma |
| río pararani |
| río viscachani |
| sacani |
| sapahaqui |
| seguencani |
| senamhi |
| serkhekhota |
| tahuapalca |
| taypichaca |
| taypichulo |
| tirata |
| totorani |
| trancajahuira |
| uma palca |
| uni |
| urujara |
| viacha |
| villa asunción |
| villa concepción |
| wilacota |
| yanacachi |
| yanari |
| yupampa |

# Appendix C. - WordClouds By Year and Month


2010
agua
paz
zona
río
año
lluvia
2011
agua
paz
zona
vecino
río
epsa
2012
agua
paz
zona
proyecto
alto
2013
agua
paz
alto
zona
río
2014
agua
paz
río
lluvia
2015
agua
paz
lluvia
río
2016
agua
paz
alto
epsa
ciudad
2017
agua
paz
río
lluvia
2018
agua
paz
río
2019
paz
agua
basura
alpacoma
2020
agua
paz
río
lluvia
2021
río
municipio
agua
lluvia
Jan
agua
Paz
río
lluvia
Feb
agua
Paz
río
lluvia
Mar
agua
zona
Paz
proyecto
Apr
agua
zona
Paz
Alto
May
agua
zona
Paz
vivienda
Jun
agua
año
Paz
Jul
agua
zona
Paz
Aug
agua
Alto
Paz
Sep
agua
año
río
Paz
Oct
agua
zona
Paz
Nov
agua
EPSA
Paz
Alto
Dec
agua
lluvia
río
zona

# Appendix D. - NER Entities Result

The categories are aut: Authority, com: Community, kw: water keyword, mun: Municipality, wb: Water body as a river or a lagoon and ws: Water system as a dam or reservoir.

| ENTITY | CATEGORY |
|---|---|
| aaps | aut |
| abastecimiento | kw |
| abastecimiento_de_agua | kw |
| acceso_al_agua | kw |
| achachicala | com |
| achocalla | com / mun |
| acueducto | kw |
| aducción | kw |
| agua_dulce | kw |
| agua_potable | kw |
| agua_sostenible | kw |
| agua_subterránea | kw |
| agua_sucia | kw |
| agua_sustentable | kw |
| agua_turbia | kw |
| aguas_abajo | kw |
| aguas_arriba | kw |
| aguas_contaminadas | kw |
| aguas_del_illimani | aut |
| aguas_negras | kw |
| aguas_residuales | kw |
| aguas_servidas | kw |
| ajuankhota | ws |
| alcantarilla | kw |
| alcantarillado | kw |
| alcantarillado_sanitario | kw |
| alerta | kw |
| alerta_agropecuaria | kw |
| alerta_amarilla | kw |
| alerta_máxima | kw |
| alerta_meteorológica | kw |
| alerta_naranja | kw |
| alerta_roja | kw |
| alerta_temprana | kw |
| almacena | kw |
| almacenamiento | kw |
| aluminio | kw |
| amachuma | com |
| análisis_de_riesgos | kw |
| ananta | com |
| apaña | com |
| aramani | com |

| ENTITY | CATEGORY |
|---|---|
| ascenso_de_nivel | kw |
| atención_ciudadana | kw |
| atención_de_desastre | kw |
| atención_de_emergencias | kw |
| atención_de_riesgo | kw |
| atención_inmediata | kw |
| autoridad_jurisdiccional_de_administración_minera | aut |
| avircato | com |
| ayma | com |
| azufre | kw |
| bacteria | kw |
| baños_ecológicos | kw |
| basura | kw |
| bombardeo_de_nubes | kw |
| botadero | kw |
| cachapa_alto | com |
| cachapaya | com |
| cahuayuma | com |
| cairoma | mun |
| calamarca | mun |
| calentamiento_global | kw |
| calidad_ambiental | kw |
| calidad_del_agua | kw |
| calidad_del_aire | kw |
| calidad_hídrica | kw |
| cambio_climático | kw |
| canal_de_riego | kw |
| cañería | kw |
| carencia_de_agua | kw |
| carpani | com |
| carreras | com |
| carros_aguateros | kw |
| caudal | kw |
| cayimbaya | com |
| cebollullo | com |
| chacaltaya | com |
| challasirca | com |
| chanca | com |
| chañurani | com |
| choquechihuani | com |
| choquecota | com |
| chullo_oxani | com |
| cloro | kw |
| cohani | com |
| cohoni | com |

| ENTITY | CATEGORY |
|---|---|
| coliformes | kw |
| collana | com |
| comibol | aut |
| condensación | kw |
| conexión_de_agua | kw |
| conexiones_clandestinas | kw |
| conexiones_ilegales | kw |
| contaminación | kw |
| contaminación_ambiental | kw |
| contaminación_del_agua | kw |
| contaminación_hídrica | kw |
| cooperativa_de_servicios_de_agua | aut |
| coroico | mun |
| corrosión | kw |
| cortes_de_agua | kw |
| cosecha_de_agua | kw |
| cotaña | com |
| crecida | kw |
| crisis_del_agua | kw |
| cronograma_de_distribución | kw |
| cuerpo_de_agua | kw |
| cuidado_del_agua | kw |
| cultivo | kw |
| damnificados | kw |
| daño | kw |
| dbo | kw |
| decreto_de_emergencia | kw |
| déficit_de_agua | kw |
| déficit_de_precipitación | kw |
| déficit_del_líquido | kw |
| déficit_hídrico | kw |
| demanda_bioquímica_de_oxígeno | kw |
| demanda_de_agua | kw |
| demanda_química_de_oxígeno | kw |
| depósito_de_agua | kw |
| derrame | kw |
| derroche | kw |
| derrumbe | kw |
| desabastecimiento | kw |
| desaguadero | kw |
| desarrollo_agropecuario | kw |
| desarrollo_comunitario | kw |
| desarrollo_de_proyectos | kw |
| desarrollo_económico | kw |
| desarrollo_forestal | kw |

| ENTITY | CATEGORY |
|---|---|
| desarrollo_integral | kw |
| desarrollo_productivo | kw |
| desarrollo_rural | kw |
| desarrollo_social | kw |
| desarrollo_sostenible | kw |
| desastre | kw |
| desastre_ambiental | kw |
| desborde | kw |
| descenso_de_niveles | kw |
| desechos | kw |
| desechos_biológicos | kw |
| desechos_mineros | kw |
| desechos_químicos | kw |
| desechos_residuales | kw |
| desechos_tóxicos | kw |
| desertificación | kw |
| deshielo | kw |
| desinfección | kw |
| deslizamiento | kw |
| desplome | kw |
| desvío | kw |
| deterioro | kw |
| dióxido_de_carbono | kw |
| dique | kw |
| dirección_especial_de_gestión_integral_de_riesgo | aut |
| disponibilidad | kw |
| disposición_final | kw |
| distribución_de_agua | kw |
| distribución_de_agua_potable | kw |
| dotación | kw |
| dqo | kw |
| drenaje | kw |
| drenaje_pluvial | kw |
| ducto | kw |
| e_coli | kw |
| eficiencia_hídrica | kw |
| el_alto | mun |
| el_niño | kw |
| emagua | aut |
| embalse | kw |
| embovedado | kw |
| emergencia | kw |
| emergencia_hídrica | kw |
| energía | kw |
| enfermedades | kw |

| ENTITY | CATEGORY |
|---|---|
| enfermedades_diarreicas | kw |
| epidemia | kw |
| epsas | aut |
| escasez_de_agua | kw |
| escasez_de_lluvia | kw |
| esfera_del_agua | kw |
| estanque_pacajes | wb |
| estaño | kw |
| estrellani | ws |
| evacuación | kw |
| evaporación | kw |
| exceso_de_agua | kw |
| exceso_de_lluvia | kw |
| explotación | kw |
| falta_de_agua | kw |
| falta_de_agua_potable | kw |
| falta_de_lluvia | kw |
| forestación | kw |
| fuga | kw |
| gestión_de_riesgo | kw |
| grado_de_contaminación | kw |
| granizada | kw |
| granizo | kw |
| guerra_del_agua | kw |
| hampaturi | com |
| helada | kw |
| hidroeléctrica | kw |
| higiene | kw |
| huajchilla | com |
| huancarani | com |
| huaricana | com |
| huayhuasi | com |
| humedad | kw |
| impacto_ambiental | kw |
| incendios | kw |
| incendios_forestales | kw |
| innovación | kw |
| instalación_de_agua | kw |
| inundaciones | kw |
| inversión | kw |
| irupana | mun |
| jalancha | com |
| khapi | com |
| kunkahuikhara | ws |
| la_enconada | com |

| ENTITY | CATEGORY |
|---|---|
| la_glorieta | com |
| la_granja | com |
| la_niña | kw |
| la_paz | com / mun |
| laguna_achocalla | wb |
| laguna_ajuankhota | wb |
| laguna_chojña_khota | wb |
| laguna_estrellani | wb |
| laguna_milluni | wb |
| laja | mun |
| lipari | com |
| litio | kw |
| lixiviados | kw |
| lluto | com |
| lorocota | com |
| lurata | com |
| magnesio | kw |
| mamaniri_huerta | com |
| manejo_de_cuencas | kw |
| mazamorra | kw |
| mecapaca | com / mun |
| medidas_de_contingencia | kw |
| megadeslizamiento | kw |
| mercurio | kw |
| metales | kw |
| millocato | com |
| mineral | kw |
| ministerio_de_medio_ambiente_y_agua | aut |
| mitigación_de_riesgo | kw |
| monitoreo | kw |
| naturaleza | kw |
| nevada | kw |
| nieve | kw |
| nitratos | kw |
| nivel_de_reserva | kw |
| obra_de_toma | kw |
| olor | kw |
| oro | kw |
| oxígeno_disuelto | kw |
| palca | com / mun |
| palcoma | com |
| palomar | com |
| pantini | com |
| parásito | kw |
| peligro | kw |

| ENTITY | CATEGORY |
|---|---|
| peñol | com |
| pérdidas_de_agua | kw |
| periodo_de_lluvia | kw |
| periodo_seco | kw |
| pileta | kw |
| pinaya | com |
| plan_director_de_la_cuenca | aut |
| plan_nacional_de_cuencas | aut |
| planta_de_tratamiento | kw |
| planta_potabilizadora | kw |
| plomo | kw |
| potabilización | kw |
| precipitación | kw |
| presa_khota_khota | ws |
| preservación | kw |
| presión_del_agua | kw |
| prevención | kw |
| prevención_de_riesgo | kw |
| programa_mi_agua | aut |
| provisión | kw |
| provisión_de_agua | kw |
| proyecto_multipropósito | kw |
| ptar | kw |
| pucarani | mun |
| quebrada_alpacoma | wb |
| racionamiento | kw |
| rebalse | kw |
| reciclaje | kw |
| recuperación | kw |
| recursos_naturales | kw |
| red_de_agua | kw |
| red_de_distribución | kw |
| reducción_de_riesgo | kw |
| reforestación | kw |
| rehabilitación | kw |
| relleno_sanitario | kw |
| remanente | kw |
| represa_chacaltaya | ws |
| represa_chuquiaguillo | ws |
| represa_condoriri | ws |
| represa_hampaturi | ws |
| represa_incachaca | ws |
| represa_la_paz | ws |
| represa_milluni | ws |
| represa_pampahasi | ws |

| ENTITY | CATEGORY |
|---|---|
| represa_pampalarama | ws |
| represa_tuni | ws |
| reserva_de_agua | kw |
| reservorio | kw |
| resiliencia | kw |
| retamani | com |
| reutilizar | kw |
| revestimiento | kw |
| riada | kw |
| riego | kw |
| riesgo | kw |
| río_achachicala | wb |
| río_achocalla | wb |
| río_achumani | wb |
| río_alpacoma | wb |
| río_aruntaya | wb |
| río_choqueyapu | wb |
| río_chuquiaguillo | wb |
| río_huayñajahuira | wb |
| río_irpavi | wb |
| río_jilusaya | wb |
| río_kaluyo | wb |
| río_kantutani | wb |
| río_kollpajahuira | wb |
| río_la_paz | wb |
| río_luribay | wb |
| río_orkojahuira | wb |
| río_palca | wb |
| río_palcoma | wb |
| río_pararani | wb |
| río_viscachani | wb |
| rotura | kw |
| sacani | com |
| salmonella | kw |
| salud_ambiental | kw |
| salud_pública | kw |
| saneamiento | kw |
| sapahaqui | mun |
| seguencani | com |
| seguridad_hídrica | kw |
| senamhi | aut |
| sequía | kw |
| serkhekhota | wb |
| sifonamiento | kw |
| sin_servicio | kw |

| ENTITY | CATEGORY |
|---|---|
| sistema_de_agua | kw |
| sistema_de_agua_potable | kw |
| sostenibilidad | kw |
| sumidero | kw |
| suministro_de_agua | kw |
| suministro_de_agua_potable | kw |
| tahuapalca | com |
| tanque | kw |
| tanque_de_agua | kw |
| taypichaca | ws |
| taypichulo | com |
| tecnología | kw |
| terraceo | kw |
| terraza | kw |
| tirata | com |
| tormenta_eléctrica | kw |
| torrente | kw |
| torrentera | kw |
| totorani | wb |
| trancajahuira | wb |
| trasvase | kw |
| tubería | kw |
| uma_palca | com |
| uni | com |
| urujara | com |
| vertedero | kw |
| vertiente | kw |
| viacha | mun |
| villa_asunción | com |
| villa_concepción | com |
| wilacota | com |
| yacimiento | kw |
| yanacachi | mun |
| yanari | com |
| yoduro_de_plata | kw |
| yupampa | com |
| zanja | kw |