EGU21-13198
https://doi.org/10.5194/egusphere-egu21-13198
EGU General Assembly 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Natural Language Processing In Integrated Water Resources Management. Case Study: Three Bolivian River Basins

Camilo Andres Gonzalez Ayala[1], Santiago Duarte Prieto[1], Ana Escalera[1], Gerald Corzo Perez[1], Hector Angarita[2], and German Santos Granados[3]

[1]IHE Delft Institute for Water Education, Delft, Netherlands (cgo001@un-ihe.org, s.duarte@un-ihe.org, aes002@un-ihe.org, g.corzo@un-ihe.org)

[2]Stockholm Environment Institute, Bogotá, Colombia (hector.angarita@sei.org)

[3]Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia (german.santos@escuelaing.edu.co)

The socio-economic development of a country depends mainly on adequate integrated water resources management (IWRM). Sectors such as mining and agriculture are two main economic activities in Bolivia, that negatively impact the water resource quality and availability. Also, 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. Nowadays, the media such as newspapers, television, radio, report on these problems, in terms of water resources, which are experienced year after year in the country. Furthermore, due to advances in technology, this information can be found digitally. In the same way, people have made use of social networks, such as twitter, to express their opinion on a specific topic. The type of information found both in the media and in social networks is called qualitative information.

This digital information will be extracted using web crawling and web scrapping techniques that allow the process to be automated. This process is performed by applying keywords in the context of water resources in Bolivia, such as names of different water bodies in a basin. Once the information has been extracted, it will be transformed into a quantitative form, in such a way that it is useful for planning and decision-making processes of IWRM in Bolivia.

The purpose of this research is focused on the application of Natural Language Processing in the digital information found for three hydrological basins located in Bolivia, in order to recognize how Bolivian society relates the management of water resources. These hydrological basins are La Paz - Choqueyapu, Tupiza and Pampa – Huari. Initially, the digital information that will be studied in this research consists of three Bolivian newspapers and the information found on Twitter. The application of a sentiment analysis classification model in Python language programming is developed. In order to preserve the semantic information and the different words in the text, Word2Vec model will be used. The extracted digital information is pre-processed, eliminating empty words that do not add sentiments to a text and punctuation marks. Once the information is pre-processed, it is divided into two types, training and testing. The training data will be used to

train the Word2Vec model. The result of the model consists of a value that determines the positive, neutral or negative sentiment of the text. Once the model is trained, the testing data that has not been used will be applied in order to evaluate the performance of the model.

This research helps to identify key elements, actors, frequent words related to IWRM, factors related to river health and improve the concept of citizen science. The results are mapped by geolocation, as a frequency distribution considering the digital perception (sentiment analysis) found and the frequency in which a topic is mentioned in the analysed digital information.