Mining Online Sources to Analysis Socio-economic Impacts of Hamun Lake Desiccation

Mahdi Akbari
Water, Energy and Environmental Engineering Research Unit, Faculty of Technology, University of Oulu, Finland

Gituhub address: https://github.com/MahdiZizou/Hamun-Lake-NLP-project.git

Email:
mahdi.akbari@oulu.fi

ORCID iDs:
Mahdi Akbari: https://orcid.org/0000-0002-6598-9994

Abstract— Desiccation of Hamun Lake, in Sistan region, has led into unemployment in the region. Sistan region is very controversial region in border of Iran and Afghanistan affected by terrorism activities in recent years. Economic and medical consequences of Hamun Lake desiccation are discussed a lot in literature and media of Iran. Social media in crisis situations, such as environmental disasters, have been recognized by scholars and practitioners as key communication channels that can complement traditional channels. However, there is limited empirical examination from the user perspective of the functions that social media play and the factors that explain such uses. In this study we examine Twitter use to study how Twitter users have reacted to desiccation of Hamun Lake. We analyzed frequency of words in tweets to gain insight on what aspect of Hamun Lake is more important for users. Also, we applied some other Natural Language Processing techniques such as sentiment analysis, lexical categorization and semantic analysis. The results showed that twitter users reacted to Kamal-Khan dam inauguration majorly because this dam can affect Hamun Lake considerably by cutting inflow to this lake. Also, sentiment analysis of tweets revealed that negative feeling is more probable in retrieved tweets related to Hamun Lake and Sistan Region. Lexical categorization of tweets showed how economic consequences of Hamun lake desiccation is reflected in tweets. Finally, semantic analysis of online sources using Google search engine showed that "Desiccation of Hamun Lake", "Fishing halt" and "Dust in Zabol" has most similarity using WebJaccard similarity index. This similarity shows how Hamun Lake desiccation has led into dust problem in Zabol, the closest city to Hamun Lake, and unemployment in the Sistan region.

Keywords— *Twitter API, Google API, NLTK, WordCloud, SentiStrength, Empath Client, WebJaccard Similarity*

I. INTRODUCTION

Desiccation of water bodies due to drought or anthropogenic activities is a widespread phenomenon such as Caspian Sea Marginal Gulfs [1] and Aral Sea [2] in Eurasia, and Hamun Lake [3] and Urmia Lake [4] in Iran. This phenomenon has environmental and economic consequences.

Hamun Lake is located on the Iran-Afghanistan border in the Sistan region (Figure 1-a). This lake is consisting of 4 water bodies known as Hamun-i Puzak, Hamun-i Sabari, Hamun-i Hirmand and Gaud-i Zirreh (Figure 1-a). They are in Helmand Basin which is a transboundary basin and the most important river flowing to these water bodies is Hirmand (or Helmand) River. Sistan region has four cities and 980 villages, with a population of more than 400,000 [3]. Hamun Lakes are the biggest fresh lake in all over the Iran platue and are identified in the Ramsar Convention [5] which are very important for the economy and environment of the region [3]. The majority of settlements around Hamun Lakes are in Iran [6], majorly in Zahedan and Zabol cities (Figure 1-b). Therefore, desiccation of Hamun Lakes affects this country more than Afghanistan. For example, Kamal-Khan Dam (Figure 1-b) made a conflict between Iran and Afghanistan in recent years. Iran claims that this Dam will affect Hamun Lakes considerably [7]. Annual inflow from Hirmand River to Iran is shown in Figure 1-c. we can observe that recently annual inflow average has decreased from 3600 to 1900 MCM. In 1970s, a treaty on Hirmand River monthly inflow to Iran, known as Water Protocol (Figure 1-d), was signed between Iran and Afghanistan in which 820 Million Cubic Meter (MCM) (Figure 1-d) annually inflow to Iran was guaranteed by Afghanistan government [8]. This protocol is the only official document between two governments on Hirmand River inflow [9].

In the Sistan region, life quality of 400,000 people including economy and medical aspects are depended on Hirmand River inflow and Hamun Lakes [10]. This region has attracted scientific interest because of being a major dust source in southwest Asia [11]. Zabol as one the most important cities in the Sistan region and most close one to Hamun Lakes (Figure 1-b) had the most polluted air of the

world in year 2012 where mean annual PM_{10} = 527 and $PM_{2.5}$ = 217 $\mu g/m^3$ [12]. A study showed that 63% of the people in Zabol suffer from respiratory diseases, whereas the health damage and medical costs for patients exceeded 166.7 million U.S. Dollars during the period 1999–2004. [13]. In Zabol city, from September 2010 to July 2011 (316 days), only 21 days (less than 7 percent) was in healthy condition [3]. In 1977 more than 55 percent of Sistan people in Iran were working in agricultural sector but this ration has reduced to less than 22 percent in 2015 due to drought periods [14]. Drought has negatively impacted fisheries which have been brought to a halt [3] and caused high unemployment in Sistan region. Unemployment is one of the most important reasons of terrorism activities in Middle Eastern and North African known as MENA [15]. For example, Grishk Dam (shown in Figure 1-a) had two damages which would cost 3 million US dollars to repair by a group of twenty Taliban fighters in 2005.

Social media in crisis situations, such as natural disasters, have been recognized by scholars and practitioners as key communication channels that can complement traditional channels [16]. Considering that there is still a lack of clarity about the role of traditional media in crisis and risk communication [17], it comes as no surprise that there is even more limited understanding about the role of social media in these contexts [18]. Studies in this area mostly deal with the use of social media from an organizational perspective [19]. Such studies have documented cases of effective and ineffective uses of social media in crisis, such as a university response during an earthquake [20], or in global issues such as climate change. This stream of research has led to the development of best practices for organizations from a public relations perspective, such as: communicate quickly, be credible, be accurate, be simple, be complete, and communicate broadly [21]. Similarly, governmental organizations have realized the potential of social media in dealing with crises. For example, the Organization for Economic Cooperation and Development (OECD) and the U.S. Congress have recently developed reports from their institutional perspectives outlining benefits and challenges of social media for crisis managers. Studies of social media use by lay populations, however, remain scarce. We argue that there is an equally pressing need to understand social media use during natural disasters from the perspective of ordinary users and public communicators, consistent with how much the audience has changed. An important body of research in information and computational sciences has explored this area [22], while researchers are only starting to investigate the implications of social media in disaster relief situations [16].

In this study we investigated Tweets to see how people react to Hamun Lake desiccation in order to highlight the most important effects of this environmental disaster. In this research we examined text mining techniques such as text categorization, sentiment analysis, semantic similarity, etc. to develop some information of how people feel about Hamun Lake desiccation based online on text resources. Using statistical analysis, we determined most frequent words of tweets and WebJaccard as an index to show similarity to understand what main consequences of Hamun Lake desiccation are. Proposed approach by this research can help to have more insights about what is happening in Sistan society to cover lack of social data in the region.

II. METHODOLOGY

A. Datasets most frequent words

We analyzed content shared on Twitter about Hamun Lake and Sistan region. First, we build a positive dataset (D+) and negative dataset (D-) from tweets by defining queries based on some keywords (Table 1). We used Twitter API module in python known as Tweepy [23] for retrieving tweets. It should be noted that because most tweets related to studied topic are in Persian (popular language in Iran and Afghanistan), so keywords were defined in Persian. Then, we translated all Tweets to English using Google Translate module in python known as googletrans [24]. Based on defined queries by OR operator between keywords of each datasets (Table 1), excluding retweets, D+ and D- were built.

B. Tokenization of tweets

After building D+ and D-, we tokenized them all by tokenizer function from NLTK module in python [25]. In this step, we drew histogram of the most common terms, excluding stop-words, for D+ and D-.

C. Part of speech (POS) tagging

After tokenizing all tweets, we tagged them all by POS tagger in NLTK module in python. Finally, using tokenized and POS tagged words from tweets in D+ and D-, we determined most frequent noun and verbs of each data set.

D. Graphical representation of datasets by WordCloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites [26]. We used this technique in our research to have a graphical representation for D+ and D-.

E. Sentiment analysis of datasets by SentiStrength

SentiStrength estimates the strength of positive and negative sentiment in short texts, even for informal language. It has human-level accuracy for short social web texts in English, except political texts. More negative scores from SentiStrength reports more negative feeling in the text [27].

F. *Lexical categorization by Empath*

Empath is a tool that can generate and validate new lexical categories on demand from a small set of seed terms (like "bleed" and "punch" to generate the category violence). Empath draws connotations between words and phrases by deep learning a neural embedding across more than 1.8 billion words of modern fiction [28]. We applied this tool on D+ and D- to see what main lexical categories of these datasets are.

G. *Semantic analysis by WebJaccard similarity index*

Page counts for the query P AND Q, can be considered as an approximation of co-occurrence of two words P and Q on the Web. We modify popular co-occurrence index of Jaccard to compute semantic similarity using page counts retrieved from Google API in python [29]. For the rest of this paper we use the notation H(P) to denote the page count for the query P in Google search engine. WebJaccard coefficient (WJ) between words P and Q is defined by:

$$WJ(P,Q) = \begin{cases} 0 & if\ H(PQ) < c \\ \frac{H(PQ)}{H(P)+H(Q)-H(PQ)} & otherwise \end{cases} \quad (1)$$

Here, PQ denotes the conjunction query P AND Q. Given the scale and noise in web data, it is possible that two words may appear on some pages purely accidentally. In order to reduce the adverse effects attributable to random co-occurrences, we set the WJ to zero if the page count for the query PQ is less than a threshold c.

All above-mentioned steps are taken using python programming language. All codes are available in Github page of the project: https://github.com/MahdiZizou/Hamun-Lake-NLP-project.git.

III. RESULTS AND DISCUSSIONS

We could find very low numbers of tweets for D+ (less than 10), but D- included more than 300 tweets by defined query. Most common words for D+ are "water", "Alborz" and "upright", and most common words for D- are "dam", "Kamal-Khan", "third phase" and "water" as shown in Figure 2. The low number of positive tweets related to Hamun Lake and Sistan region, shows how negatively this region is affected by Hamun Lake desiccation. Having "water" as most frequent word in both D+ and D- shows that water is very important issue in the region and most people talk about it in their tweet regardless of showing positive or negative feelings. In Iran, dams are one of the most important causes of environmental issues like Urmia Lake desiccation [30]. This is reflected in D- because "dam" word is very popular among tweets. Another important point in D-, is presence of "Kamal-Khan" and "third phase". As shown in Figure 1-a, On Hirmand River we have 4 dams and most close one to Iran border is Kamal-Khan dam. This dam is in third phase of construction and has inaugurated experimentally. This fact is also affected in tweets of D- considerably.

Also, after POS tagging of all word of tweets in each positive and negative datasets, most common nouns and verbs of D+ and D- are plotted in Figure 3. Ashraf Ghani, president of Afghanistan, has recently visited Kamal-Khan Dam and wished it's operation happens soon [31]. The verbs "open" and "inaugurate" and noun "Ghani" are popular in D- because of mentioned fact about Kamal-Khan dam.

Based on graphical representation of most common words from Word Cloud (Figure 4), positive words like "upright", "peace", "beauties", "full", "celebration" and "fairness" are popular in D+. However, words "Kamal-Khan", "Ashraf-Ghani", "inaugurated", "" and "third phase" can reflect negative feeling of people, majorly Iranian, in D- about damming of Hirmand River by Afghanistan (also known as Helmand River).

After applying SentiStrength on each of D+ and D-, overall score of them found to be 0 and -1 respectively. It makes sense based on the defined query for building each of them. Furthermore, applying Empath showed that main lexical categories for D+ are "Beach", "Water", "Swimming", "Sailing", "Cleaning" and "Ocean". However, major categories for D- are "Negative emotion", "Economics", "Poor", "Aggression", "Water" and "Government". The lexical categories for both D- and D+ also includes Water which shows how water resource management problem is crucial for Hamun Lake and Sistan region. Also, negative economic consequences of Hamun Lake are reflected in categories of D-.

The result of WJ similarity index for different queries (shown in Table 2) shows that most similar queries using Google search engine are "Desiccation of Hamun Lake", "Fishing halt" and "Dust in Zabol". This is consistent with our prior knowledge from Sistan region because two important main consequences of Hamun Lake desiccation are destruction of fishing industry of the region and long dust storms of the region especially in Zabol city [3], [10].

Main limitation of this study was the small size of datasets because we had to use Persian language in Twitter API to cover more tweets. Although we also defined some English keywords to make D+ and D- as largest as possible, we could only retrieve 327 tweets (325 for D- and 5 for D+). Therefore, inferences from retrieved datasets can be more robust if size of D+/D- increases. This fact is due to some reasons: 1) the Hamun Lake desiccation is a disaster in Sistan region and inherently it is not likely to have positive tweets about it, 2) Hamun Lake is close to border of Iran and Afghanistan. Sistan Region in both countries is very undeveloped region and affected people from Hamun Lake desiccation do not have access to internet commonly and 3) even in other

developed citied of Iran such as Tehran, Hamun Lake desiccation is not hot topic because of security issues related to Sistan area which is affected by terrorist activities a lot in recent years. However, another desiccated lake in north west of Iran, Urmia Lake, received more attentions in media and academic communities compared to Hamun Lake.

IV. Conclusion

In this research we found that Natural Language Process techniques can help to gain insight about undeveloped regions affected by environmental disasters like water body desiccation. Economy of Sistan region is deeply affected by Hamun Lake desiccation and this fact was also shown by Empath lexical categorization. Some important facts of the region like inauguration of Kamal-Khan dam by president of Afghanistan was also found by analysis of most common words in tweets. Finally, from sentiment analysis we found that "Hamun Lake" and "Sistan region" as main keywords of our queries for retrieving tweets, are more probable to show negative feeling among twitter users than positive sentiment. Finally, the most important consequences of Hamun Lake desiccation based on literature was found by WJ index which are Zabol dust and fashioning industry ruin.

Acknowledgment

The authors are thankful to Eng. Nabil Arhab who provided insights and expertise that greatly assisted the research.

References

[1] M. Akbari *et al.*, "Vulnerability of the Caspian Sea shoreline to changes in hydrology and climate," *Environ. Res. Lett.*, 2020.

[2] A. AghaKouchak *et al.*, "Aral Sea syndrome desiccates Lake Urmia: call for action," *J. Great Lakes Res.*, vol. 41, no. 1, pp. 307–311, 2015.

[3] A. Rashki, D. G. Kaskaoutis, C. J. d. W. Rautenbach, P. G. Eriksson, M. Qiang, and P. Gupta, "Dust storms and their horizontal dust loading in the Sistan region, Iran," *Aeolian Res.*, vol. 5, pp. 51–62, 2012.

[4] M. Akbari, A. Torabi Haghighi, M. M. Aghayi, M. Javadian, M. Tajrishy, and B. Kløve, "Assimilation of satellite-based data for hydrological mapping of precipitation and direct runoff coefficient for the Lake Urmia Basin in Iran," *Water*, vol. 11, no. 8, p. 1624, 2019.

[5] A. Moghaddamnia, M. G. Gousheh, J. Piri, S. Amin, and Dja. Han, "Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques," *Adv. Water Resour.*, vol. 32, no. 1, pp. 88–97, 2009.

[6] M. Pesaresi and S. Freire, "GHS Settlement grid following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015)," *Eur. Comm. Jt. Res. Cent.*, 2016.

[7] BBC, "Afghanistan: Dam construction in Afghanistan should not be a concern for Hassan Rouhani," 2017.

[8] Ministry of Enegry (MoE) and Yekom Consultancy Company, "Update of Sistan River studies (hydrological report)," 2014.

[9] ICANA, "25% migration of Sistan people due to drought," 2015. [Online]. Available: khabaronline.ir/news/587938. [Accessed: 25-Aug-2020].

[10] A. Rashki, D. G. Kaskaoutis, A. S. Goudie, and R. A. Kahn, "Dryness of ephemeral lakes and consequences for dust activity: The case of the Hamoun drainage basin, Southeastern Iran," *Sci. Total Environ.*, vol. 463–464, pp. 552–564, 2013.

[11] A. S. Goudie and N. J. Middleton, *Desert dust in the global system*. Springer Science & Business Media, 2006.

[12] WHO, "Global Health Observatory data repository," *World Health Organization*, 2016. [Online]. Available: https://apps.who.int/gho/data/view.main.AMBIENTCITY2016?lang=en. [Accessed: 08-Nov-2020].

[13] A. Miri, H. Ahmadi, A. Ghanbari, and A. Moghaddamnia, "Dust storms impacts on air pollution and public health under hot and dry climate," *Int J Energy Env.*, vol. 2, no. 1, pp. 101–105, 2007.

[14] Ministry of Cooperatives Labour and Social Welfare Iran, "Employment status of Sistan and Baluchestan province," 2017.

[15] A. Bagchi and J. A. Paul, "Youth unemployment and terrorism in the MENAP (Middle East, North Africa, Afghanistan, and Pakistan) region," *Socioecon. Plann. Sci.*, vol. 64, no. December 2017, pp. 9–20, 2018.

[16] B. Takahashi, E. C. Tandoc, and C. Carmichael, "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines," *Comput. Human Behav.*, vol. 50, pp. 392–398, 2015.

[17] A. Af Wåhlberg and L. Sjöberg, "Risk perception and the media," *J. Risk Res.*, vol. 3, no. 1, pp. 31–50, 2000.

[18] A. R. Binder, "Figuring out #fukushima: An initial look at functions and content of US twitter commentary about nuclear risk," *Environ. Commun.*, vol. 6, no. 2, pp. 268–277, 2012.

[19] S. R. Veil, T. Buehner, and M. J. Palenchar, "A Work-In-Process Literature Review: Incorporating Social Media in Risk and Crisis Communication," *J. Contingencies Cris. Manag.*, vol. 19, no. 2, pp. 110–122, 2011.

[20] N. Dabner, "'Breaking Ground' in the use of social media: A case study of a university earthquake response to inform educational design with Facebook," *Internet High. Educ.*, vol. 15, no. 1, pp. 69–78, 2012.

[21] K. Freberg, K. Saling, K. G. Vidoloff, and G. Eosco,

"Using value modeling to evaluate social media messages: The case of hurricane irene," *Public Relat. Rev.*, vol. 39, no. 3, pp. 185–192, 2013.

[22] A. Acar and Y. Muraki, "Twitter for crisis communication: Lessons learned from Japan's tsunami disaster," *Int. J. Web Based Communities*, vol. 7, no. 3, pp. 392–402, 2011.

[23] J. Roesslein, "Tweepy: Twitter for Python!," *URL https//github.com/tweepy/tweepy*, 2020.

[24] S. Han, "googletrans: googletrans for Python!," *URL https//github.com/ssut/py-googletrans.git*, 2020.

[25] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.," 2009.

[26] S. Kadam, "Generating Word Cloud in Python," 2020. [Online]. Available: https://www.geeksforgeeks.org/generating-word-cloud-python/. [Accessed: 11-Oct-2020].

[27] SentiStrength, "SentiStrength." [Online]. Available: http://sentistrength.wlv.ac.uk/. [Accessed: 11-Oct-2020].

[28] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 4647–4657, 2016.

[29] Google, "Python Quickstart for Google API," 2020. [Online]. Available: https://developers.google.com/docs/api/quickstart/python. [Accessed: 10-Oct-2020].

[30] M. Akbari, A. T. Haghighi, M. M. Aghayi, M. Javadian, M. Tajrishy, and B. Kløve, "Assimilation of satellite-based data for hydrological mapping of precipitation and direct runoff coefficient for the Lake Urmia basin in Iran," *Water (Switzerland)*, vol. 11, no. 8, 2019.

[31] BBC, "Experimental water intake of Afghanistan's Kamal Khan Dam began near the border with Iran," 2020. [Online]. Available: https://www.bbc.com/persian/afghanistan-54854466. [Accessed: 10-Oct-2020].

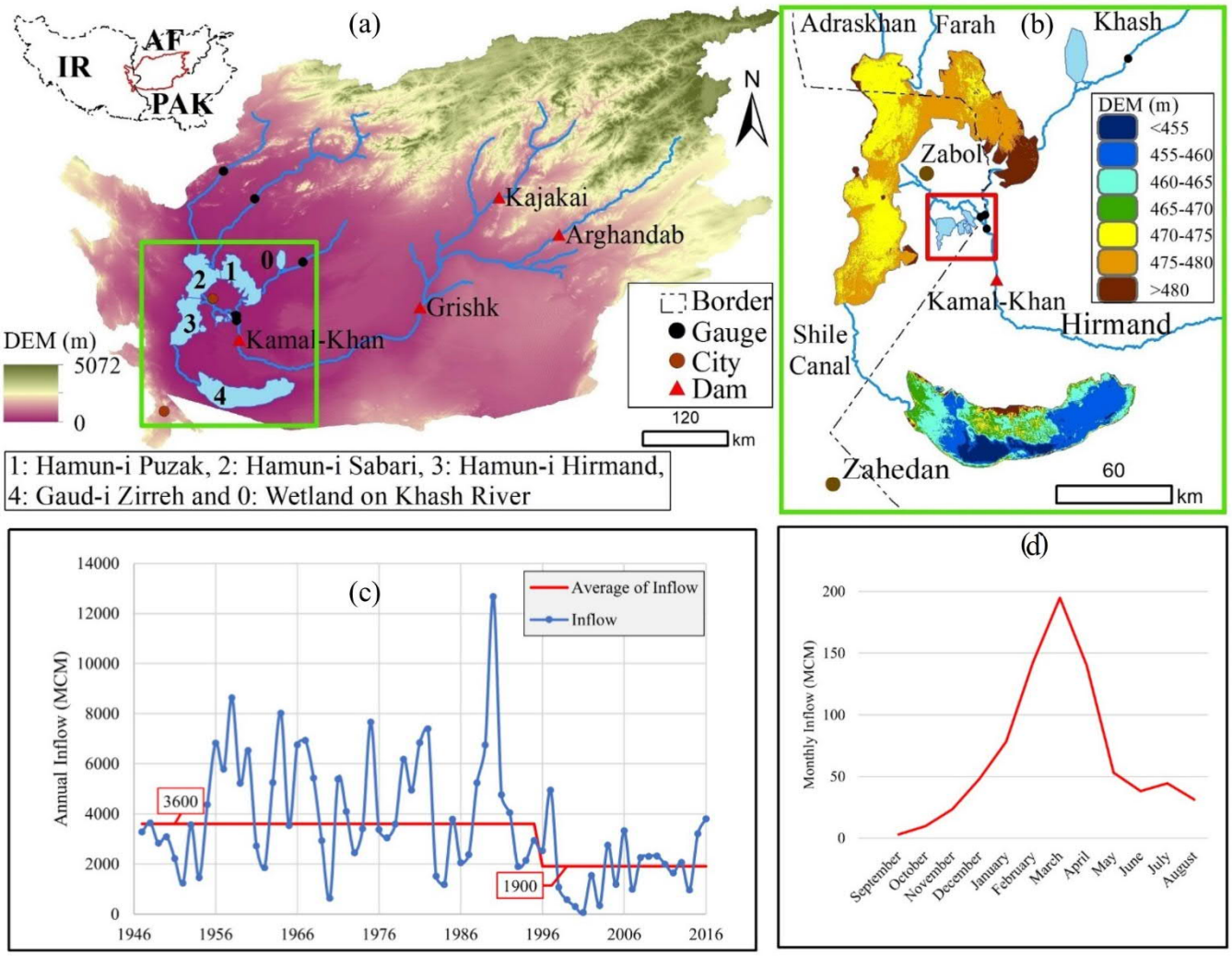

FIGURE 1. STUDIED AREA: A) HAMUN LAKES AND HELMAND BASIN WITH THE LOCATION OF DAMS, INFLOW GAUGES AND CITIES OF THE BASIN, B) DEM OF HAMUN LAKES AND CLOSE WATER BODIES TO HAMUN LAKES, C) ANNUAL INFLOW OF HIRMAND RIVER TO IRAN FROM 1946 TO 2016 AND D) WATER PROTOCOL ON HIRMAND RIVER INFLOW TO IRAN

TABLE 1. KEYWORDS OF QUERIES TO COLLECT POSITIVE AND NEGATIVE TWEETS RELATED TO HAMUN LAKE

Dataset	***Key words***	***Translation***
D+	زیبایی سیستان	The beauty of Sistan
	زیبایی زابل	The beauty of Zabol
	توریسم سیستان	Sistan Tourism
	دریاچه هامون زیبا	Beautiful Hamun Lake
	هامون پر آب	Full-of-water Hamun
D-	تشنگی دریاچه هامون	Thirsty Hamun
	گرد و غبار زابل	Dust in Zabol
	بیکاری سیستان	Sistan unemployment
	خشکسالی سیستان	Sistan drought
	تشنگی سیستان	Thirsty Sistan
	تشنگی زابل	Thirsty Zabol

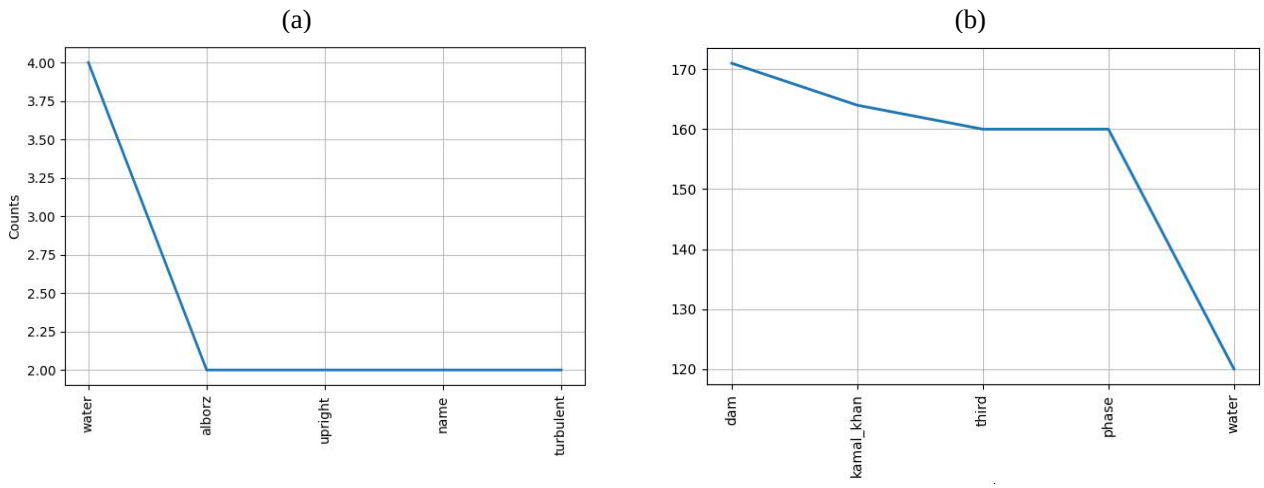

FIGURE 2. MOST COMMON WORDS IN A) D+ AND B) D-

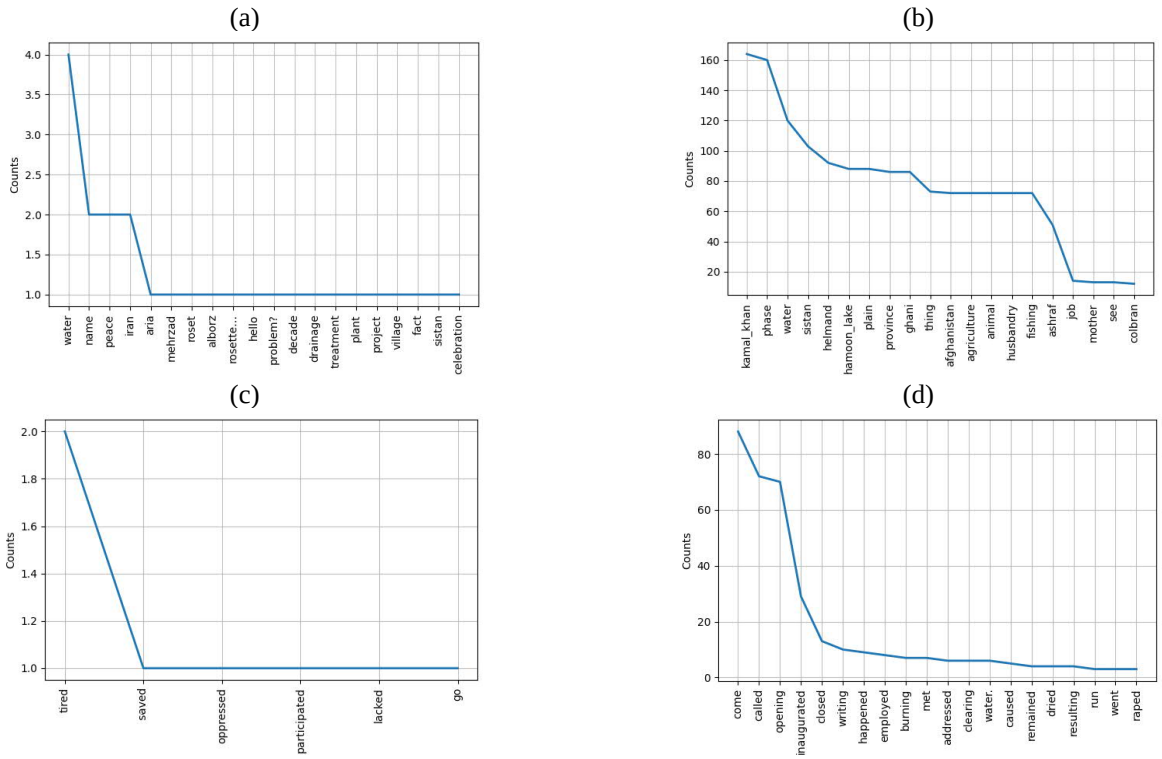

FIGURE 3. MOST COMMON NOUNS IN A) D+ AND B) D- WITH MOST COMMON VERBS IN C) D+ AND D) D-

(a) (b)

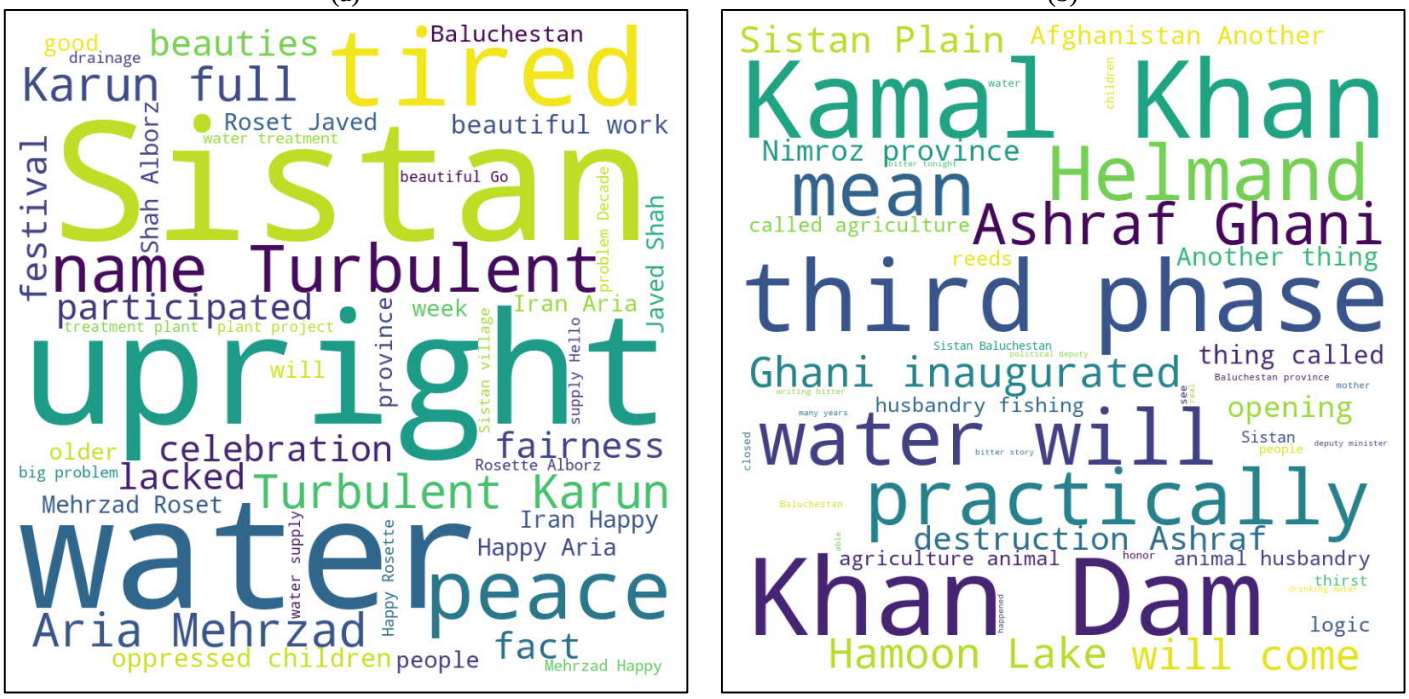

FIGURE 4. OUTPUT OF WORD CLOUD FOR A) D+ AND B) D-

TABLE 2. WEBJACCARD COEFFICIENT (WJ) FOR VARIOUS QUERIES

Query	Translation	WJ (P, Q) (%)
Q=خشکی دریاچه هامون	Desiccation of Hamun Lake	0.53
P=بیکاری سیستان و زابل	Unemployment in Sistan and Zabol	
Q=خشکی دریاچه هامون	Desiccation of Hamun Lake	20.6
P=توقف ماهیگیری سیستان	Fishing halt	
Q=خشکی دریاچه هامون	Desiccation of Hamun Lake	2.9
P=کمبود آب شرب	Domestic water deficit	
Q=خشکی دریاچه هامون	Desiccation of Hamun Lake	20.3
P=گرد و غبار زابل	Dust in Zabol	
Q=خشکی دریاچه هامون	Desiccation of Hamun Lake	5.9
P=کاهش گردشگری سیستان و زابل	Decrease in tourism in Sistan and Zabol	