**Pakistan Flood of August – 2022 Causes and Implications - Disaster Data Mining based on Social Media Texts**

Department of Information Science, University of North Texas

INFO 5810: Data Analysis and Knowledge Discovery

Professor: Ijay Kaz-Onyeakazi

05 Nov 2022

**Hashika Mutyala (11514776) - Team Lead**

**Komal Singh Bhamra (11483011) – Member**

**Likitha Jarugula (11597739) – Member**

**Vasista Acha (11601274) - Member**

**Introduction**

We took topic of disaster data mining based on social media texts and for that we have done some background research and found some disasters happened recently in different countries and states such as floods or cyclones or earthquakes etc., and we are taking Pakistan’s floods as the disaster topic. It was found that the floods have left one-third of the country underwater. The satellite images show the burst of the largest lake in the Pakistan (Shewan Sharif) made people to flee and run from their respective villages. But the main reason for the devastation is due to the extreme hot conditions at start in April and May 2022 because of that they have almost got a rainfall i.e., 190 percent more than normal and people say that they couldn’t predict and that’s the reason why they had to face many issues. There is a high significance taking this topic because it was the first time that this much of floods happened in Pakistan and it affected Pakistan in every kind like impacting Agriculture, living and how it’s still impacting it. Extreme perception is the reason for 2022 Flood and what made it to deal with the climate Changes.

**Research Questions**

As per the Research paper , they wanted to know how the precipitation will be for different parts of Pakistan and what’s the average precipitation during the month of the floods from last 20 years to check the precipitation average and then to see if weather and atmospheric rivers are any reasons for this kind of climate as due to high climates the floods happened and they couldn’t predict the same and they were not ready for the precautionary causes because of that and it made too many life’s to be affected. How does precipitation affect the climate and how is the precipitation during earlier years at the same time and what was the rate of the precipitation at that time.

**Research Purpose**

**The main purpose of the Research is to find the data about the floods from twitter dataset about the floods in the Pakistan in month of Aug 2022 and then find what are the reasons and implications for the floods from the tweets and see different insights and if the considered research can conclude any info about the floods and why did it happen and what made these many people to leave houses as one part of Pakistan is more affect with heavy floods and why is it not predicted early despite of unusual temperatures before month and to see if this is un-natural or if it was an effect of some conditions.**

**Research Methods**

We have done my basic research by referring to different articles from the internet to see the information about floods in Pakistan and retrieved some data in order to see data to see the real issue behind it. And we are dividing the study into five sections. Section one indicates a brief introduction, section two tells about tools and techniques related to the data text mining where the data is found and how the data is taken out and how dataset is formed , section three explains about the analysis part of how the data explains the problem and what necessary insights we can obtain from it and section four indicates the data mining challenges with and section five defines the conclusion of the dataset and to define what all can be the reasons for this level of floods to happen. This research can give proper info about what could be the reason as different people from around the world is tweeting and we must find the tweets and obtain words from it and make sense out of it and to obtain specific information like what are the earlier trends and to find other sights like in the same month how is the average precipitation before 5 years.

**Literature Review**

For thisstudy, the author went through the Forty-One previous research paper for filling all the gaps for his analysis and he has mentioned the citing’s of different research in the paper and he also used datasets to do necessary analysis to understand the insights and we can understand that author found a dataset that consists of data which has August month Precipitation data from last 20 years and even we have analysed around 12 papers to see how the researchers are going for finding the solution.

**Research Design and Strategy**

In this study, the author used some data sets, but he did not employ any sort of machine learning. He is employing a straightforward method to review different papers and he just used some basic statistics and found some trends and insights of the data.

**Design and Strategy for the Research**

Data Collection

Using WEB Scraping API

using

Convert it to CSV format by converting it to data frame

Desired Output

Data Cleaning

Finding Required Data for statistical Analysis

**METHODOLOGY**

All the necessary analysis can only be done when you have specific data, and we need to find the data that could give weather and necessary precipitation data for all years whichever is available so we can do necessary statistics to see what made the floods to happen which never happened in last 70 odd years. And I got to know about different kinds of tools, techniques, and processes for data mining.

**Learnings:** By this we have learnt usage of different tools like jupyter and some packages like tweepy which is a web scraping API that can download tweets directly from the twitter using a specific keyword search and it will display all the tweets and then u can understand the insights of the learning.

**Overall Assessment:** This paper is more interested in knowing the root cause of the floods and rather than knowing the losses that happened to Pakistan whether it may be economically or with the number of populations. And he got a dataset which has 20 years of data about the precipitation, and he has applied statistical analysis on it to get the results

**Future Research:** I am thinking that using this research we can find the number of deaths happened in different parts of Pakistan and the report of their injuries or to know whether people are safe or not. We can apply different NLP Techniques to understand the Situation of the Pakistan by doing a analysis if required to understand the level of issue.

**Qualitative Research:** This research indicates the history of the problem and the case study which focuses on insights of the data.

**Quantitative Research:** This research indicates the process of collecting all the numerical data to find the patterns, make predictions and to generalize result on basis of one sample.

**Data Collection**

We are using data mining to deal with larger datasets and to find proper insights from it. For this we have tried to scrape a large amount of data from the twitter as people going to tweet about the floods in the social media texts using twitter and so I must do web scraping with API using Tweepy package where it will take necessary credentials and the API key for your dataset. And the next thing is more important as we have to find the right key word to search as per our data requirement we are trying to find about the keywords like Pakistan, floods, fled, deaths, month, southern Pakistan and all sides of Pakistan to see where more floods are happening and which has more deaths and precipitation rate. All this info needs to retrieve from the tweets for a specific span of time and getting the whole tweet. Taking right keywords is the key for the data collection and more data you have the better it will be as we will get the required info.

But as per the research they have found a dataset that could get info of the precipitation for last 20 years and if we could find the dataset, we can find more insights and do a better analysis and predict it for next 5 years to see if there is any insight that could tell us this kind of hazards which will be a heavy loss to mankind.

**Data Cleaning**

Data is obtained through web scraping, and they have sent it to a .csv file to do necessary analysis. Now we need to clean the data as we might get too much of unnecessary data, we will have to remove it to find required data and we will have to create new features for creating an understanding as the model needs to find insights from the model. Data cleaning is of different types. In Tweets there are unnecessary symbols and that needs to be removed as it won’t mean anything and by adding those Infront of a word leads to a bad meaning for it. The emoticons and Emoji’s will also make a specific sense sometimes it represents good and sometimes there’s a sarcasm intended too so different set of words along with emoticons makes some sense too at times. We are going to use NLP techniques like Stemming, lemmatization to get features of the tweets as in order to find some data like if we need to know how many people died or how many are injured. Depending on the requirement we will use specific techniques as of now we want to see the average precipitation. As per our motive we need to know why this happened, so need to know if this is unprecedented or if it’s something which wasn’t checked early i.e., if all the life’s that are lost because of the lack of knowledge of the government to take necessary precautions.

**Data Analysis Plan**

we would like to do different analysis by putting visualization of the data by creating baggage of words and by plotting it , we would be able to see the most used words which is required i.e., Cause of the Floods or the kills or the injured ones or the people who fled from their place to other place for living and even you can find it with respect to the economy too if the floods are affecting economy. We are trying to find data in two ways one is by scraping and one from internet directly as the floods dataset will be available and the weather info of every country is also maintained so with that we can create Data Analysis to check the importance of the weather and to forecast it and see if the trends are plotted to see if there is any sudden spike which indicates the issue of the weather as that will indicate the insight why is it happening so that government can get ready for such cases so that the economy won’t get affected and loss of this many people at once can also be avoided with a caution.

**Conclusion**

From the research paper, it was found that after applying necessary techniques and Statistics of the given data, it was found that there is extreme precipitation during august 2022 which made the floods to happen and before 5,6 months there is a change in temperature as there is an unrealistic heat during those days and government weren’t looking after the climate changes that are found to be not usual and hence, they couldn’t protect it. Even with the past data it was found that precipitation was never this high and hence it was found that there are other conditions which made the flood to happen. This research can be useful to analyse different insights like analysing the deaths and injuries in different parts of Pakistan and how the economy got impacted by the same.

**Contributions**

* Hashika Mutyala is Team lead, who worked on Web Scraping of the data and converting into data frames Hashika installed necessary packages and then implemented Tweepy Web Scraping API to obtain all the information using API with the creds.
* Komal Singh Bhamra is a Group Member who worked on finding different Research papers and going through it to see and understand the analysis of it.
* Likitha Jarugula is a group Member who worked on dealing with Search of dataset and finding ways of research methodologies.
* Vasista Acha is a group member who deal with data cleaning which includes different functionalities. And finding the necessary features from it.

**References and Citations**

1. Kushwaha, Anuj & Singh, Rajesh & Malik, Iqra & Solanki, Hiren & Chuphal, & Vegad, Urmin & Dangar, Swarup & Mahto, Shanti Shwarup & Mishra, Vimal & J S, Nanditha & Singh Chuphal, Dipesh. (2022). The Pakistan flood of August 2022 causes and implications. 10.13140/RG.2.2.31705.80480.
2. Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., De Roo, A., Salamon, P., et al. (2017).394  
   Global projections of river flood risk in a warmer world. Earth’s Future, 5(2), 171–182.395 https://doi.org/10.1111/eft2.183396  
   Algarra, I., Nieto, R., Ramos, A. M., Eiras-Barca, J., Trigo, R. M., & Gimeno, L. (2020).397
3. Significant increase of global anomalous moisture uptake feeding landfalling398  
   Atmospheric Rivers. Nature Communications, 11(1), 1-7.399  
   Almazroui, M., Saeed, S., Saeed, F., Islam, M. N., & Ismail, M. (2020). Projections of 400
4. Precipitation and Temperature over the South Asian Countries in CMIP6. Earth Systems401and Environment, 4(2), 297–320. https://doi.org/10.1007/S41748-020-00157-4027/FIGURES/13403  
   Arshad, M., Ma, X., Yin, J., Ullah, W., Ali, G., Ullah, S., et al. (2021). Evaluation of GPM-404  
   IMERG and TRMM-3B42 precipitation products over Pakistan. Atmospheric Research,405  
   249, 105341. https://doi.org/10.1016/J.ATMOSRES.2020.105341406
5. Aziz, F. (2022). Pakistan floods: what role did climate change play? The Conversation.407  
   Retrieved from https://theconversation.com/pakistan-floods-what-role-did-climate-408  
   change-play-189833409
6. Baqir, M., Sobani, Z. A., Bhamani, A., Bham, N. S., Abid, S., Farook, J., & Beg, M. A.410

(2012). Infectious diseases in the aftermath of monsoon flooding in Pakistan. Asian411

Pacific Journal of Tropical Biomedicine, 2(1), 76.

<https://doi.org/10.1016/S2221-4121691(11)60194-9413>

1. Berghuijs, W. R., Woods, R. A., Hutton, C. J., & Sivapalan, M. (2016). Dominant flood414  
   generating mechanisms across the United States. Geophysical Research Letters, 43(9),4154382–4390. https://doi.org/10.1002/2016GL068070416  
   Berthet, L., Andréassian, V., Perrin, C., & Javelle, P. (2009). How crucial is it to account for417 antecedent moisture conditions in flood forecasting? Comparison of event-based on 418 continuous approaches on 178 catchments. Hydrol Earth Syst Sci, 13(6), 819–831.419
2. https://doi.org/10.5194/hess-13-819-2009420  
   Bertola, M., Viglione, A., & Blöschl, G. (2019). Informed attribution of flood changes to421 decadal variation of atmospheric, catchment and river drivers in Upper Austria. Journal422  
   of Hydrology, 577.
3. https://doi.org/10.1016/j.jhydrol.2019.123919423  
   Bhutto. (2022). The west is ignoring Pakistan’s super-floods. Heed this warning: tomorrow it 424 will be you. The Guardian. Retrieved from425  
   https://www.theguardian.com/commentisfree/2022/sep/08/pakistan-floods-climate-crisis426
4. Blöschl, G., Gaál, L., Hall, J., Kiss, A., Komma, J., Nester, T., et al. (2015). Increasing river427  
   floods: fiction or reality? Wiley Interdisciplinary Reviews: Water, 2(4), 329–344.428 <https://doi.org/10.1002/wat2.1079429>
5. Brakenridge, R., & Anderson, E. (2006). MODIS based flood detection, mapping and430  
   measurement: The Potential for operational hydrological applications. Transboundary431  
   Floods: Reducing Risks Through Flood Management, July(1–12).432  
   <https://doi.org/https://doi.org/10.1007/1-4020-4902-1_1433>
6. Cao, H., Zhang, H., Wang, C., & Zhang., B. (2019). Operational Flood Detection Using434 Sentinel-1 SAR Data over Large Areas. Water, 11(786).435

<https://doi.org/https://doi.org/10.3390/w11040786436>

1. Cao, Q., Mehran, A., Ralph, F. M., & Lettenmaier, D. P. (2019). The Role of Hydrological 437

Initial Conditions on Atmospheric River Floods in the Russian River Basin. Journal of438  
Hydrometeorology, 20.https://doi.org/10.1175/JHM-D-19