UNIVERSITE JOSEPH KI-ZERBO

ECOLE DOCTORALE INFORMATIQUE ET CHANGEMENTS CLIMATIQUES

BURKINA FASO

Unité-Progès-Justice

MASTER RESEARCH PROGRAM

SPECIALITY: INFORMATICS FOR CLIMATE CHANGE (ICC) MASTER THESIS

Subject:

THE CONTRIBUTION OF CROWDSOURCING TO URBAN FLOOD DATA COLLECTION, MONITORING, AND COMMUNICATION: CASE STUDY IN THE CITY OF OUAGADOUGOU, BURKINA FASO

Presented on the 4th July 2022 by:

Haddy Jawla

Supervisors

Dr. Seyni SALACK, Regional Thematic Coordinator – Risks and Vulnerability to Climate Extremes, WASCAL

Prof Sidat Yaffa, UTG/WASCAL Director of Doctoral Research Program on Climate Change and Education, University of The Gambia.

Academic year 2020-2021

Dedication

This master thesis is dedicated to my beloved parents, especially my dad, who has supported me financially, emotionally, and psychologically over the years. To him, I owe my academic success.

Acknowledgment

I would like to express my appreciation to WASCAL for awarding me the master's scholarship to pursue and embark on my educational career and make a difference in my life.

Special thanks to Prof. Tanga Pierre ZOUNGRANA, the director of Ecole Doctorale Informatique et Changement Climatique (ED-ICC), the deputy director of EDICC, Dr. Ousmane COULIBALY, and the Scientific coordinator of EDICC, Dr. Benewinde Jean-Bosco ZOUNGRANA for their support and encouragement during the course.

I sincerely thank my supervisors, Dr. Seyni Salack and Prof. Sidat Yaffa, for their supervision, encouragement, and guidance during my work. I am grateful for the continued advice, enthusiastic support, invaluable insights, recommendations, positive critics, and timely feedback on my work.

To my dearest husband, I owe you my deepest gratitude for the love, support, source of confidence, inspiration, encouragement, and, most importantly, the time you bear with me during the period of my study.

To my parents and Siblings, all this wouldn't have been possible without your endless prayers and support. Special thanks to my dad. All of this is the fruit of your labor; thanks for making me who I am today. Thanks to my in-laws for the understanding and the prayers I am grateful for.

I thank Dr. Bako Ferdinand for making available all the shapefiles for this project. Dr. Momodou Lamin Tarro and Alpha Jallow, thanks for the continuous support and encouragement. Special thanks to a classmate and a friend Mariatou Faal for the proofreading.

I appreciate the WASCAL Ph.D. students, Valentin Ouedraogo, Sadraki Yabre, Issoufou Gnibga Yangouliba, Bognini Karafa, and Madou Sougue, for the help and exchange of materials needed to write this project. Finally, I would thank all my colleagues, ED-IC staff, and the WASCAL Competence Center staff for all kinds of support rendered during my stay. Special thanks to 2ie Ph.D. students for reviewing my document.

I am forever grateful to Allah the almighty for giving me the strength and good health that enabled me to complete this work.

Abstract

According to observations and the nearest future, flooding is among the most frequent climate hazards; the frequency and amplitude are expected to increase according to climate change projections. During such events, Facebook, Twitter, Snapchat, LinkedIn, and other social media are used to air firsthand information. Crowdsourcing uses electronic tools to improve data collection, monitoring, and communications including web scraping to collect data and information during and after disasters to gain situational awareness. It also uses an open data kit (ODK) to collect other *in-situ* datasets. In this study, we used web scraping and ODK techniques to find the link between the triggers and consequences of urban floods.

The web scraping was combined with machine learning algorithms and exploratory analysis tools to collect flood-related messages from the Twitter website for the City of Ouagadougou (Burkina Faso). Using natural language processing techniques, the data set was normalized by *tokenization* and *lemmatization* to clean and extract emotions for the textual data of each tweet. Further, we used machine learning approaches such as NLTK and TextBlob to analyze the sentiments and polarity in the data sets and supervised machine learning (logistic regression) for text classification. The accuracy of the data was assessed using the rain gauge dataset and the reports of the disaster management agency. The tweet word cloud generated exhibited the onset of urban flood events, the causes, the effects, and the spatial extent of flood events in Ouagadougou. The text classification yielded a 93% accuracy of the Twitter data sets compared to instrumental measurements and officially reported observations. The ODK questionnaire was deployed among multiple stakeholders across the City of Ouagadougou to identify landfill triggers of floods in urban areas such as both official and unofficial dump sites.

Hence, crowdsourcing is now designated as an efficient method for data collection, improving the quality and visualization to contribute to event databases, effect-based monitoring, and communications of urban floods.

Keywords: Burkina Faso; City of Ouagadougou; Crowdsourcing; ODK Questionnaire; Urban Floods; Web Scraping.

Résumé

Selon les observations, les inondations font partie des aléas climatiques les plus fréquents et les projections à court terme indiquent que leurs fréquence et amplitude devraient augmenter sous l’effet des changements climatiques. Lorsqu’elles surviennent, Facebook, Twitter, Snapchat, LinkedIn et autres médias sociaux sont utilisés pour diffuser des informations de première main. Le *crowdsourcing* utilise des outils électroniques pour améliorer la collecte et le suivi de données et des communications, y compris la récupération automatique des données du web ou « *web scraping* » pour collecter les données et les informations sur le web afin informer sur la situation pendant et après les catastrophes. Il utilise également un kit de données ouvert (ODK) pour collecter d'autres types de données *in-situ*. Dans cette étude, nous avons utilisé les techniques de web scraping et d'ODK pour trouver le lien entre les causes et les conséquences des inondations urbaines.

Le web scraping a été combiné avec des algorithmes d'apprentissage automatique et des outils d'analyse exploratoire pour collecter sur Twitter, les messages relatifs aux inondations dans la ville de Ouagadougou (Burkina Faso). À l'aide de techniques de traitement du langage naturel, l'ensemble des données a été normalisé par *tokenisation* et *lemmatisation* pour nettoyer et extraire les émotions des données textuelles de chaque tweet. Les outils d'apprentissage automatique tels que Natural Language Toolkit (NLTK) et TextBlob ont permis d’analyser les sentiments et la polarité dans les ensembles de données et la régression logistique, la classification de texte. La validation des données a été faite à l'aide des données collectées au pluviomètre et des rapports officiels sur les inondations. Le nuage de mots généré a permis de connaitre le début, les causes, les effets et l'étendue spatiale des inondations à Ouagadougou. La classification textuelle des données Twitter a donné un résultat satisfaisant avec une précision de 93% par rapport au pluviomètre et aux rapports officiels. Le questionnaire ODK déployé auprès des parties prenantes dans la ville de Ouagadougou a permis d’apprécier la répartition des décharges (officielles et sauvages) et d'identifier celles qui peuvent être facteurs des inondations urbaines,

Par conséquent, le crowdsourcing peut être considéré comme une méthode efficace pour la collecte, l'amélioration de la qualité et la visualisation de données pouvant contribuer à la constitution des bases de données, à la veille basée sur les effets et à la communication sur les inondations urbaines.

Mots clés: Burkina Faso; Crowdsourcing; Twitter; Inondations urbaines; Questionnaire ODK ; Ville de Ouagadougou ;

Table of Contents

Acronyms and Abbreviations

API	Application Programming Interface
ASCII	*American Standard Code for Information Interchange*
CONASUR	Conseil National de Secours d'Urgence et de Réhabilitation
CSV	Comma Separated Values
DGTTM	Direction Générale des Transports Terrestres et Maritimes.
ED-ICC	Ecole Doctorale Informatique et Changement Climatique
GPS	Global Positioning System
GUS	Green Urban Spaces
GWP	*Global Water Partnership*
HRE	Heavy Rain Events
MLA	Machine Learning Algorithm
MSW	Municipal Solid Waste
MSWM	Municipal Solid Waste Management
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
ODK	Open Data Kit
TF-IDF	Term Frequency–Inverse Document Frequency
UNDP	United Nations Development Programme
UNDRR	United Nations Office for Disaster Risk Reduction
UPA	Peri-Urban Agriculture

USAID	: United States Agency for International Development
WASCAL	: West African Science Service Centre on Climate Change and Adapted Land Use.
WB	: World Bank
WMO	: World Meteorological Organization.
WWW	: World Wide Web.

List of Figures

[redacted]INTRODUCTION

Floods are one of the major climate-related disasters occurring in various parts of West Africa and around the world. A flood is overflowing the normal confines of a stream or other body of water or the accumulation of water over areas that are not normally submerged. Floods include river (fluvial) floods, flash floods, urban floods, pluvial floods, sewer floods, coastal floods, and glacial lake outburst floods (Dokken, 2012.)They are triggered by processes such as heavy rainfall, poor drainage, dumpsites/landfills, and groundwater saturation and may cause serious losses and damages to agricultural lands, residential settlements, and cities with high living costs to the economy of a country. Climate change may enhance flood hazards, according to Rentschler et.al, (2020) as global warming which has led to an increase in temperature has affected the pattern and intensity of the precipitation leading to heavy rainfall (Park et al., 2021). Flooding also poses numerous hazards to healthcare facilities, including damage to equipment and building structures, electrical hazards, loss of power, and poor indoor scent after flood water recedes (Hazard Control, 2006). Natural disasters are reaching catastrophic proportions as the world population grows and more people migrate to hazard-prone areas (Brown et.al, 1994). From 2000 to 2019, floods accounted for 44% of all disasters, affecting 1.6 billion people globally (UNDRR, 2019). According to OCHA (2021), flood has affected about 669,000 people in the West and Central African Countries of The Democratic Republic of Congo, Gambia, Niger, Chad, Nigeria, Togo, The Republic of Congo, The Central African Republic, Burkina Faso, and Ghana. The adverse effects of floods come along with their risks and benefits are distributed very unevenly across societies (Kundzewicz et al., 2014). Future changes in flood risk are expected to be driven by a combination of potential changes in climate (especially precipitation), catchment conditions, and loss exposure.

Many approaches have been used to assess the risk of floods using quantitative, qualitative, statistical, and machine learning approaches.

The advances in technological development and the world wide web (www) have involved citizen participation in providing flood information using crowdsourcing of social media outlets, journal articles, and news reports. Crowdsourcing is described as the act of obtaining information, ideas, and services from large sources and people. The notion of crowdsourcing is now closely linked to the potential developments of information and communication technology, particularly the internet and social media, for facilitating crisis mapping (WMO, 2017). Social media refers to the platforms that allow this mode of communication. Social media sites such as Facebook, Twitter, WhatsApp, Instagram, etc. aid in the dissemination of real-time information about any given situation globally. For the past years, researchers have used public social media data to explore a variety of human activities and other phenomena. Despite the distance, the age of technology has dramatically advanced in virtualization and digitalization. Social networking sites are predicted to have 3.96 billion users by 2022, and these numbers are expected to continue to rise as mobile device usage and mobile social networks gain momentum in previously underserved markets (Statista, 2022). There were 2.00 million social media users in Burkina Faso in January 2021 (Datareportal, 2022).

Researchers and business specialists all over the world have been scraping data from social media for more than a decade now, using these data to understand better individuals, groups, and society, as well as discovering brand new opportunities concealed in the data. In Burkina Faso the use of social media has advanced from 2021 to 2022 with 98.1% using Facebook, YouTube having 0.71%, Pinterest having 0.58% *and Twitter* having 0.44%, Instagram having 0.11%, and LinkedIn with 0.02% (Statcounter Global Stats, 2022). Although there are many useful social media platforms, Twitter is among the most reliable platform for social networking, micro-blogging, and sentiment analysis. Despite the high number of Facebook users, Facebook prohibits all automated data pulling and web scraping. The Application Programming Interface (API) shutdown and severe data access limitations imposed by Facebook to protect user data are debatable (octoparse,2020).

To advance in our research work, we used Twitter as a scraper, Despite the high number of Facebook users, we decided to use Twitter as Facebook prohibits all automated scrapers. That is, no part of the website should be visited by an automated crawler. Today, Twitter is one of the most increasingly popular media and widely used social networking and microblogging platforms. Because of its vast amount of content and relative ease of access, for understanding and evaluating human socio-behavioral dynamics, information diffusion, and public emotion. Text analytics, sentiment and opinion mining, topic modeling, text classification and summarization, and other applications have effectively exploited it as a data source. Due to concerns with data quality and dependability of the submitted content, using Twitter as a resource for extracting meaningful information during hazard situations is a difficult process but easier and faster than other sources of information.

1.1 Problem Statement

Some of the consequences of floods are easily reported because they are evident due to the number of casualties, fallen bridges, etc. others are not easily reported due to their cascading nature, teleconnections (e.g., Traffic jams, car accidents, stranded passengers). An increase in frequency and intensity of HRE and the ongoing urbanization may further increase the risk of flash flooding. Under climate change, future global warming scenarios reveal that HRE will be more intense, and flooding may likely be exacerbated in flood-prone areas of the city and beyond as the inadequate disposal and household waste gathering in the streets block drainage systems, streams, and landfills and threaten health in residential areas. There is a need to improve the disaster risk management schemes and early warning systems and guide decision-making in implementing flood-control infrastructure projects at national and regional scales. Currently, warnings for pluvial floods are mostly limited to information on rainfall events over larger areas (i.e., no given rain intensities and durations), which is often not detailed enough to protect people and goods effectively. Many sources are available to report on special events such as floods including National Disaster reports, radio stations, TVs, Written Newspapers, and Social media outlets. A fundamental assumption when dealing with data obtained from social media outlets is that it cannot be considered as reliable as reports from professional observers. Social media channels are mostly misused to publish and share fake content that could lead to insecurity and anxiety among citizens if wrongly endorsed and propagated. In this perspective, there is a growing need for automatic validation techniques. Previously studies have used innovative methodology using social media data along with supervised machine learning techniques and the results generated were highly accurate in predicting disaster mapping (Anbalagan & Valliyammai, 2017). "*The wisdom of the crowd*" principle using the filtering and geo statistical methods results show a good correlation between some social media data and observation data (Eilander et al., 2016). In West Africa, fewer studies were conducted on crowdsourcing through social media outlets in support of decision-making and disaster management. This study endeavors to investigate the reliability of social media data in flood risk analysis and create a citizen awareness of social media, especially Twitter. This would be achieved by using two different crowdsourcing techniques to see how urban floods can be monitored and communicated in the city of Ouagadougou.

1.2 Research Questions and Hypotheses

Can crowdsourcing qualitatively contribute to urban flood risk analysis? To investigate this general question, we specifically treat these four questions:

1. Is web scraping useful in collecting Twitter data on floods?
2. How to improve the quality of scraped tweets to high-quality and reliable datasets for analysis of floods in the City of Ouagadougou?
3. Can Twitter data contribute to urban flood data collection, monitoring, and communication?
4. How do mobile utility applications inform on landfills as triggers of urban floods in the city of Ouagadougou?

We hypothesized that social media outlets and open data forms can improve crowdsourcing in providing reliable data, information, and communication on floods in Ouagadougou. As a case study applied to the City of Ouagadougou, our working specific hypotheses are the following:

1. Web scraping applications can enhance data collection on urban floods.
2. Embedded Machine learning algorithms can improve the reliability of scraped data for the City of Ouagadougou.
3. Twitter data is a valuable resource for data collection, effect-based monitoring, and communication of urban floods.
4. The deployment of an ODK questionnaire improves the data collection on landfills.

In other to achieve these hypotheses we have set the following objectives as indicated below.

1.3 Research Objectives

The main objective of this thesis is to demonstrate the valuable contribution of crowdsourcing based on Twitter messages and the deployment of ODK questionnaires that can improve the data collection, monitoring, and communication on urban floods. To get more appropriate results, the following specific objectives are sought:

1. Design and develop an API to scrape tweets over the past five years from Twitter.
2. Use natural language processing and other machine learning techniques to extract the most valuable data from the scraped messages on floods in the City of Ouagadougou.
3. Assess the performance of the collected data and information relative to rain gauge and official reports
4. Geo-localize the official and unofficial urban landfills which trigger urban floods in the City of Ouagadougou.

Apart from contributing to the science of flood risk analysis, this work is expected i) to support decision-making toward flood risk mitigation and adaptation, and ii) to sensitize the public on the use and the usefulness of citizen science and social media information such as Twitter.

[redacted]CHAPTER 1: Literature Review

2.1 Floods

2.1.1 Some Basic concepts on floods

According to WMO (2013), there is no universal definition of what constitutes a flood. However, it is defined as the overflowing of the usual confines of a stream or other body of water or the accumulation of water over areas that are not usually submerged. Any land usually above water level is said to be flooded if submerged for one or two hours arbitrarily.

The causes of floods can be broadly divided into physical, climatological forces, and human influences, such as vegetation clearing and urban development. The most common causes of floods are climate-related, most notably rainfall. Prolonged rainfall events are the most common cause of flooding worldwide, and the greater the rainfall intensity, the greater the potential for runoff. It may also be caused by artificial factors and can cause massive damage to life and property. Figure 1 illustrates the aspects that concern to cause of floods.

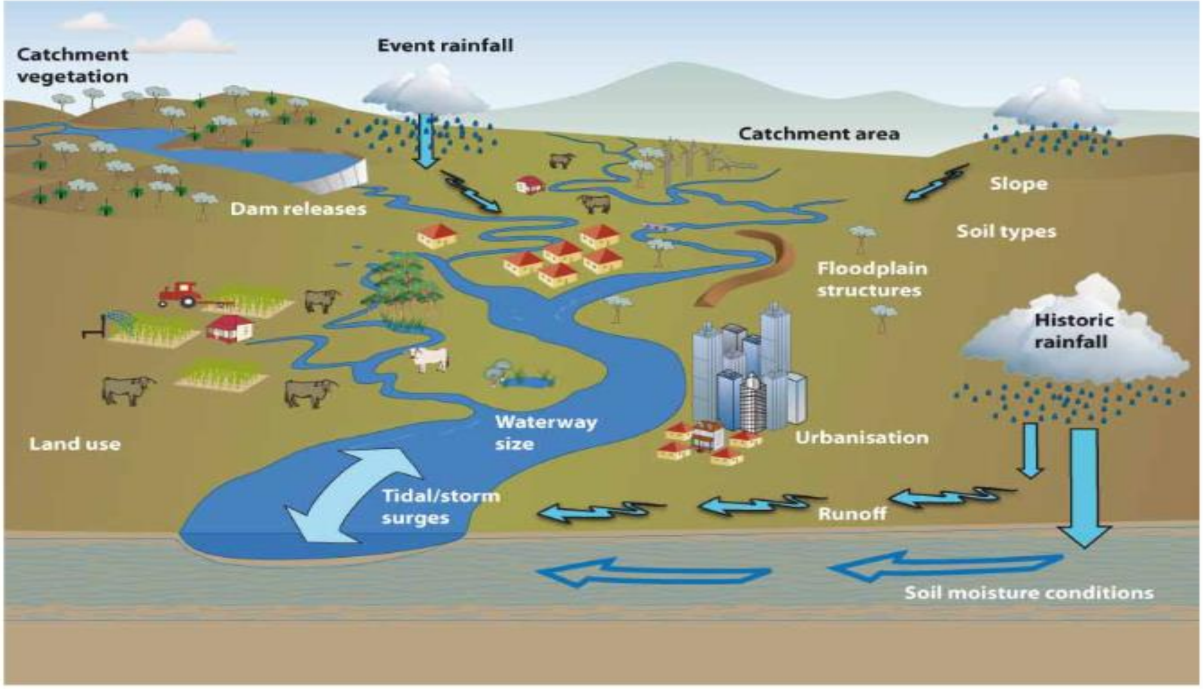


Figure 1 : Factors that lead to or trigger flooding. (source: https://www.chiefscientist.qld.gov.au/publications/understanding-floods/what-factors-contribute)

The most typical categorization includes *Pluvial floods or flash floods. The latter is* described as flood events that occur during or within a few hours of the rainfall that causes the rise in water. Terrain gradients, soil type, plant cover, human habitation, antecedent rainfall, and other hydrological factors all play a role in the occurrence of a flash flood. Thunderstorms, heavy rain events, or deep, moist convection are responsible for most flash floods connected with rainfall. The second category is *Fluvial floods (riverine floods).*

In contrast, to flash floods, fluvial floods typically unfold over days or even months and are usually the result of many individual rainfall episodes spread out over many days. In fact, within a river flood event, several flash flood events can occur. Again, hydrological factors often contribute to a river flood, but river floods are not as sensitive to them as flash floods. The rain causes the river to be filled with too much water, much more than the capacity of the river channel. *Coastal Flooding occurs* along fragile beaches. Storm surges – often from tropical cyclones or extratropical hurricanes – and waves, combined with riverine floods at various tide stages, result in significant loss of life regularly. Coastal floods' intensity is affected by several elements, such as the windstorm's strength, magnitude, speed, and direction. The *Urban Flood* is significantly different from rural flooding as urbanization leads to developed catchments, which increases the flood peaks from 1.8 to 8 times and flood volumes by up to 6 times.

Consequently, flooding occurs very quickly due to faster flow times (in minutes). Urban areas are densely populated, and people living in vulnerable areas suffer due to flooding, sometimes resulting in loss of life. It is not only the event of flooding, but the secondary effect of exposure to infection also has its toll in terms of human suffering, loss of livelihood, and, in extreme cases, loss of life.

2.1.2 Triggers of Urban Floods

Urban flooding occurs when water flows into an urban region faster than it can be absorbed into the soil or moved to and stored in a lake or reservoir. It can be caused by flash flooding, coastal flooding, river floods, or rapid snow melt. According to the WMO report in 2011 Increased vulnerability is due to population growth, economic development, Urbanization, and poverty overcrowding leading to increased landfills such as wastes (solid and liquid wastes). In 2021, Ijaz et al. (2021) conducted a study to determine the connection between Gujrat's urban flooding issue and ineffective solid waste management. The findings indicated that following rain, locations with garbage dumping facilities experience increased challenges.

In the past, wastes, rainwater, and flood management were considered and treated separately in the City of Ouagadougou. Nowadays, the urban people are witnessing a more complex but interlinked approach to managing these three key components[1]. Household waste gathering in the streets blocks drainage systems, streams, and landfills. It threatens the health of residential areas and causes water stagnation and flash floods. A better collection, recycling, and repurposing of solid and liquid waste benefit urban peri-urban agriculture (UPA) and green urban spaces (GUS). In urban areas, besides reducing the quantity of waste in drainage canals to prevent flash flooding, waste recycling also produces clean water and compost as soil nutrients (Sanfo et al., 2022).

[1] https://ichange-project.eu/living-lab-of-west-africa-2/

2.1.3 Flood Risk

Flood risk is a combination of the chance of a flood occurring and the consequences of the flood for people, property, and infrastructure. Flood risks are a function of exposure of the people and the economic activities along with the vulnerability of social and economic fabric. As a result, the impact of such floods on people's lives and livelihoods must be understood as a function of their vulnerability (WMO/GWP, 2008).

However, it is impossible to forecast the exact region of increased risk due to climate change because flood risk dynamics are influenced by various social, technical, and environmental drivers (Komolafe et al., 2015).

To completely comprehend the various components that make up urban flood threats, it is critical to be aware. Risk is sometimes equated with an extreme event or hazard (flood, drought, earthquake, storm, landslide, etc.) Caused by natural forces or a combination of natural forces and human activities.

The consequences of a flood depend upon how exposed the community is to flooding and how vulnerable its people, property, and infrastructure are to the flood’s impacts. Managing risks from floods may involve altering the chance of flooding affecting a community and reducing the effects by reducing the community’s vulnerability and exposure to flooding. The methods that are effective in reducing flood risk are very location-specific. There is no one-size-fits-all solution, and various measures are generally necessary to reduce risk.

2.2 Crowdsourcing in Disaster Management

“*Crowdsourcing*” or “*participatory sensing*” are terms used to describe citizen participation in technology-mediated methods such as social media and mobile applications (Kankanamge et al., 2019). It is currently gaining much attention as a way to improve the efficiency of disaster management (Tavra et al., 2021) because it is a method of communication that can be used before, during, and after a disaster (Harrison & Johnson, 2016). The characteristics of crowdsourcing are openness, dynamics, autonomy, and extensiveness. Hence, crowdsourcing is a decentralization concept that provides a low-cost and quick way to obtain information without interfering with crisis management efforts(Frigerio et al., 2018). This is important as it increases the efficiency of resource matching and promotes communication and coordination between response subjects, as well as between them and affected populations. However, despite the advantages of citizens' participation in collecting information, there are many challenges to face. As Jeff Howe(Howe, 2006) has put it, "*sometimes crowds can be wise, but sometimes they can also be stupid.*" Most crowdsourced information is unstructured and unorganized (Moreira et al.,2015). and needs to be collated and analyzed before it can be used. Data quality is a significant concern with omissions, exaggerations, and errors. The use of crowdsourced data has yet to be systematized, as it cannot stand alone; it needs to be validated by other datasets (Puttinaovarat & Horkaew, 2020).

2.3 Crowdsourcing and Social Media Outlets

Today, social media plays a critical part in most disasters, from gathering vital indicators from victims to interacting with first responders. Crowdsourcing applications based on social media applications such as Twitter, Facebook, and other media outlets offer a powerful capability to collect information from disaster scenes and visualize data for relief decision-making. Social media has recently played a critical role in natural disasters as an information propagator that can be leveraged for disaster relief. During disasters, the speed with which social media spreads and the time it takes for people to react have put traditional communication modes to the test (Kankanamge et al., 2019). Without a doubt, necessary, high-throughput data is generated on social media seconds after a catastrophe arises.

Processing and inferring helpful knowledge from such data, on the other hand, is difficult for a variety of reasons. Traditionally, information, communication, volunteerism, and technology were the medium used by governments to interact with people in a disaster. However, due to volunteer crowdsourcing, new technological foundations such as digital crisis information, mobile communication, digital volunteerism, and geo-technology are replacing these channels.

2.4 Web Scraping of Social Media Outlets

Web-scraping is the automated pulling of textual information from the internet. The Web Scraping technique is a set of computer programs capable of **crowdsourcing** social media platforms (e.g., Twitter & Facebook) and employing machine learning algorithms (MLA) and text mining techniques to structure the textual information collected from crowdsourcing and analyze it to define baseline indicators thresholds values, triggers, and effects of floods. The main purpose of web scraping is to pull out unstructured data from websites and convert it to structured data. For data pre-processing and analytics, the extracted data is stored as a spreadsheet in a local file on the computer or a database.

Suarez presented a new methodology for querying Twitter Search Endpoints while getting around their API limitations. This approach helped to create and advance a web interface for starting, managing, and retrieving data from web crawlers This was achieved using Python and libraries like request and beautiful soup. The results of their study show that the majority of current works use Twitter API, which allows users to search through tweets that are no older than three weeks. It also proves how efficient, cost-effective, and reliable web scraping is in accessing historical data on Twitter with fewer limitations (Hernandez-Suarez et al., 2018).

A similar study which study illustrates how web scraping and Twitter differ in terms of accuracy and effectiveness. This was accomplished by collecting data using web scraping and Twitter API to determine the benefits and drawbacks of both extraction techniques, and qualitative and quantitative evaluations are carried out. The results prove that web scraping is faster than Twitter API and it is more versatile in terms of acquiring data (Dongo et al., 2020).

2.5 Other Data Collection Tools: Open Data Kit

ODK is a free and open-source set of tools that help organizations or individuals collect, manage, and use data in resource-constrained environments. It is a set of tools geared towards simple, effective, and efficient data collection on mobile devices. One of the most well-known software suites in the sector of data collection is called Open Data Kit (ODK), and several researchers have tried to make it more advanced and highly effective to better fulfill the various demands of customers (Anokwa et.al, 2009). It is a system that streamlines data collecting, instantly digitizes data for analysis and enables remote collection progress monitoring. To avoid the usage of paper surveys and considerably shorten survey timeframes, data is collected using smartphones running Google's Android operating system. The application also enables the use of GPS coordinates, images, videos, bar codes, and sound clips as survey attachments or as the foundation for questionnaire responses (Jeffrey-Coker et al., 2010). It comprises the ODK Build, a web application for creating custom forms for data entry based on the xlsform standard, ODK Aggregate, a Java server application to store, analyze, and export form data, and ODK Collect, an Android application that allows for the entry of data directly into mobile devices. these components working together help teams of researchers can collect data quickly and simultaneously in remote areas, and all of their data can be compiled together on a single server(Griscom, 2020).

2.6 Machine learning and Text Mining

Every user is a sensor and contributor to social media and can generate valuable and immediate real-time data to develop better situational awareness. However, social media users generate a massive amount of data that needs to be summarized to provide a big picture in a disastrous situation (Karami et al. 2019). In 2012, Terpstra et al. (2012) conducted a study to investigate the possibilities of real-time and automated analysis of Twitter messages during crises. Twitcident, a recently developed framework and Web-based system that automatically filters, searches, and analyzes tweets regarding incidents, was used to implement this study. Results indicated that automated information filtering provides valuable information for operational1q response.

In a different study, Ripberger et.al, (2014) used binomial regression to assess the reliability of tweets. The paper proposes, develops, and validates a new indicator of public attention to severe weather communication. Twitter datasets were to determine the validity of the metric by systematically comparing fluctuations in Twitter activity to the issuance of tornado watches and warnings, which represent primary essential forms of communication designed to elicit, and correlate with, public attention. The assessment finds that the measure demonstrates a high degree of convergent validity, suggesting that social media data can be used to advance our understanding of the relationship between risk communication, attention, and public reactions to severe weather. a method to filter relevant Tweets and detect events using spatiotemporal clustering and a supervised machine learning approach trained language model to extract rainfall events relevant Tweets. This method shows a higher precision for filtering rainfall event-relevant Tweets, which improves the data quality for the other spatiotemporal event detection (Myneni et.al, 2017).

Hence, machine learning algorithms require training a model to predict the polarity of the text. The model is trained with text messages, labeled for their sentiment, and represented as feature vectors.

The latter conventionally requires text reader preprocessing using language processing tools like NLTK. Text preprocessing mainly involves tokenization, stemming, tagging, and possibly parsing of the text. The selection of the appropriate features from data is crucial, has proven to be a significant issue, and is always a key objective for researchers. Previous work on sentiment analysis has exploited well-known supervised machine learning (Korovesis, 2018)

Text classification by using machine learning technique several models Perceptron, Naïve Bayes and Logistic regression used to compare the model. Among the different classification algorithms using the logistic regression method accuracy level improved. Sentiment analysis is performed by using two different text feature selection methods and three classification methods. The problem statement here is analyzing the sentiment analysis over large datasets (Nath et al., 2017).

CHAPTER 2: DATA AND METHODS

To gather data, many choices must be taken including what data and where it should be collected from. This chapter will look at data collection, storage, processing, and classification. We keep a dataset for training, testing, and sentiment analysis on Twitter in support of our argument. As our goal is to achieve sentiment analysis and reliability for data provided by Twitter. A text classifier was constructed using a supervised machine learning classifier Logistic Regression. Figure 2 below shows the proposed methodology used in this research.

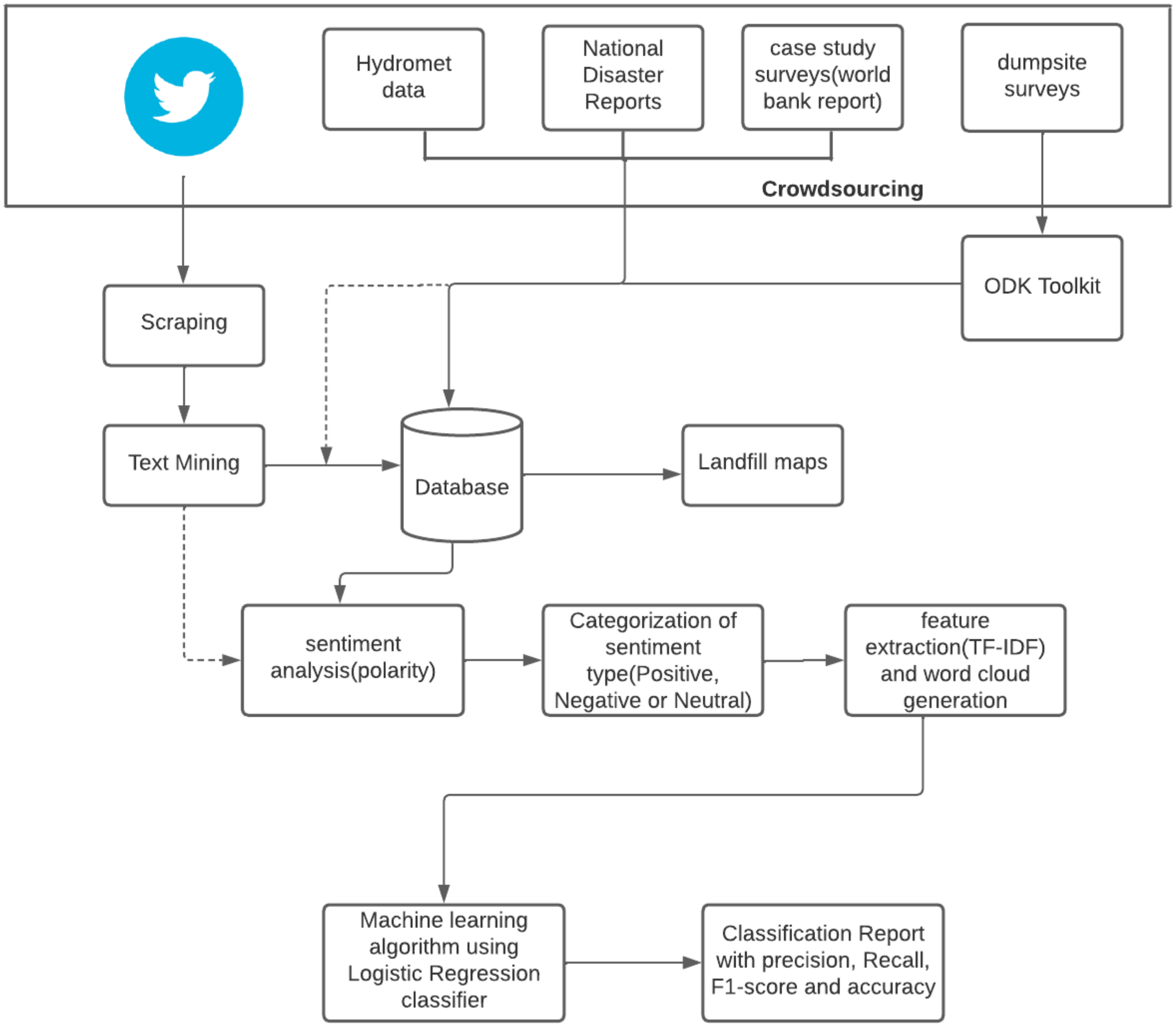

Hydromet data
National Disaster Reports
case study surveys(world bank report)
dumpsite surveys
Crowdsourcing
ODK Toolkit
Scraping
Text Mining
Database
Landfill maps
sentiment analysis(polarity)
Categorization of sentiment type(Positive, Negative or Neutral)
feature extraction(TF-IDF) and word cloud generation
Machine learning algorithm using Logistic Regression classifier
Classification Report with precision, Recall, F1-score and accuracy

Figure 2: The proposed methodology for the thesis.

3.1 Description of the study area

3.1.1 The climate of Burkina Faso

Burkina Faso is a West African country located in the mid-west of the Soudan/Sahel region. It covers an area of about 274,200 km^2 and lies between 9^{o}N-15.5^{o}N and 6^{o}W-3^{o}E[2]. The country is mainly flat, with a mean altitude of approximately 300 m above sea level. Its climate is characterized by a rainy season that lasts 3-4 months and a longer dry season. Rainfall distribution

[2] https://en.wikipedia.org/wiki/Geography_of_Burkina_Faso

across the country follows predominantly a southward gradient leading to three agroecological zones: the Sahelian zone in the North, the Sudano-Sahelian zone in the Center, the Northwest, the East, and in the South, and the Sudanian zone in the West (Figure 3). In these three zones, the average temperature increases from the West and South (Bobo Dioulasso, Banfora, Niangoloko, and Gaoua) to the North (Dori and Ouahigouya). Rainfall is highly variable, and the annual average temperature varies between 17 and 37°C (21 and 34°C) during the dry season (rainy season) across Burkina Faso (Waongo *et al.*, 2015).

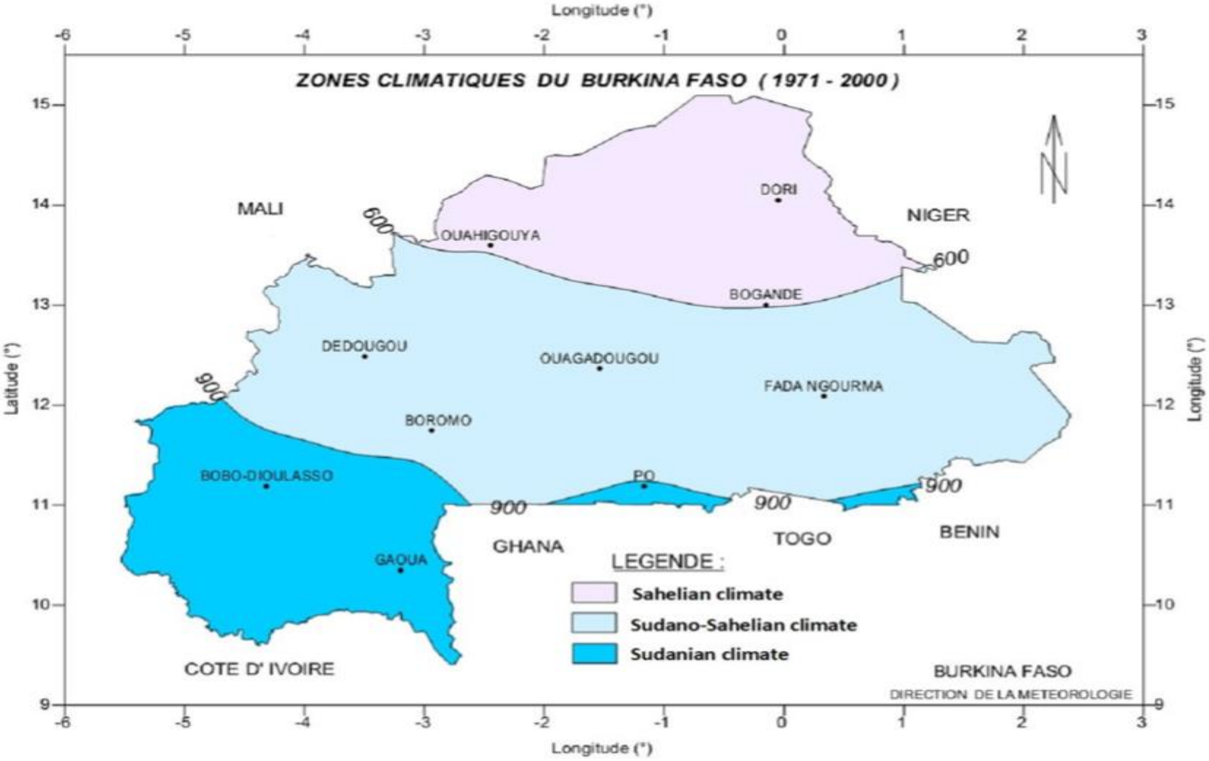


Figure 3: The three Agroecological zones of Burkina Faso. Source: ANAM, 2020

Reports suggest a warming of 0.26°C per decade over the last 30 years (USAID 2017, USAID 2022). By 2050, a 1.4 - 1.6°C rise in temperatures is expected in Burkina Faso (UNDP, 2021). Temperature is projected to increase by 3-4°C by 200-2099; this is substantially higher than the global average (World Bank, 2021). Temperature increases vary across the country, with higher temperatures expected in the north, the southwest, and dry seasons (World Bank, 2021). Flooding events are increasing. Between 1991 and 2020, Burkina Faso was hit by major floods that impacted people and properties (Salack et al., 2021). Despite a high level of uncertainty among models regarding projections on precipitation over the country (Figure 4), future dry and wet periods are likely to become more frequent in this region (Salack et al., 2016; Potsdam Institute, 2020).

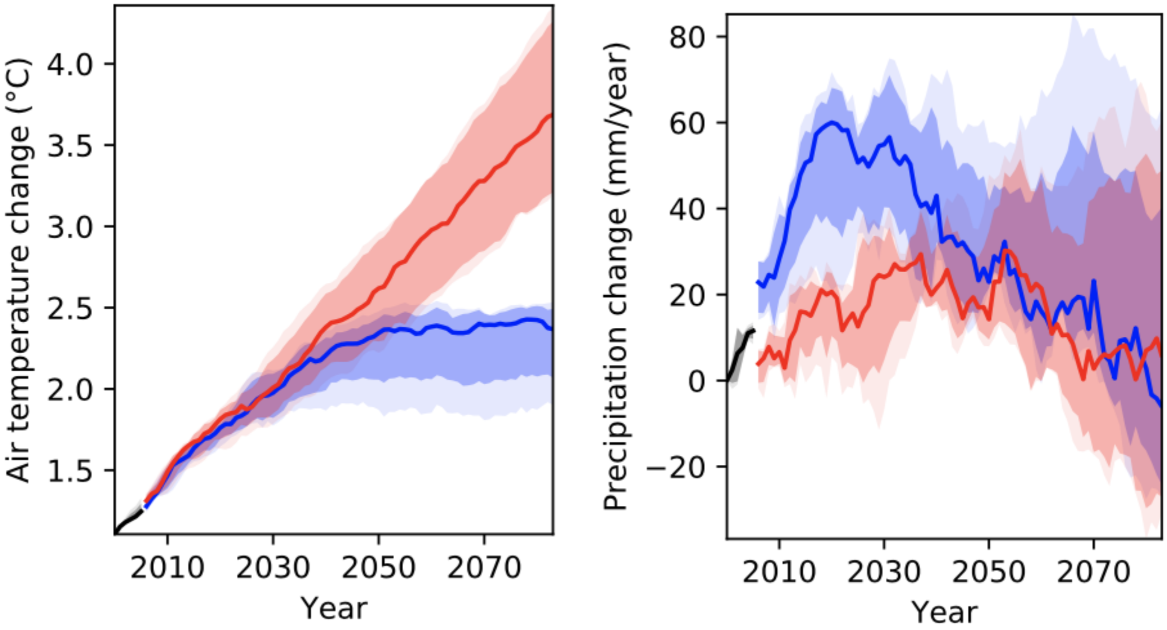


Figure 4: Annual mean air temperature and precipitation projections for Burkina Faso from different GHG emissions scenarios relative to the year 2000 (Adopted from Potsdam Institute 2020).

3.1.2 Challenges of Heavy Rain, Urban Floods, and Waste Management in Ouagadougou

The city of Ouagadougou is the capital of the Province of Kadiogo. Since 2009, the city has had 12 districts divided into 55 areas (Figure 5), with residential areas in formal and informal settlements. Heavy Rain Events (HRE) and uncollected waste in the City of Ouagadougou favor urban flash floods, cause water pollution, and increase the risks of cholera, diarrhea, and other water-borne diseases (Sanfo *et al.*, 2022). Ouagadougou has been facing rapid urban growth for the last two decades, with a constantly increasing rate of urbanization. Consequently, the urban environment has been rapidly shaped, combining heterogeneous features, from planned neighborhoods in the core areas to an unregulated expansion in the peri-urban zones. Vegetation is tree and shrub savanna type with an herbaceous layer, abundant in the rainy season, dominated by species such as Pennisetum, Cenchrus, Aristida, and Brachiaria, and a ligneous layer dominated by Combretum micranthum, Lannea microcarpa, Vitellaria paradoxa, and Parkia biglobosa. (K. Tindano). The morphological elements that characterize the Urban territory are the three (3) dams located in the north of the city center and the green belt delimited across a horizontal plane, including the Bangr-Weogo urban park coupled with other hydrological networks.

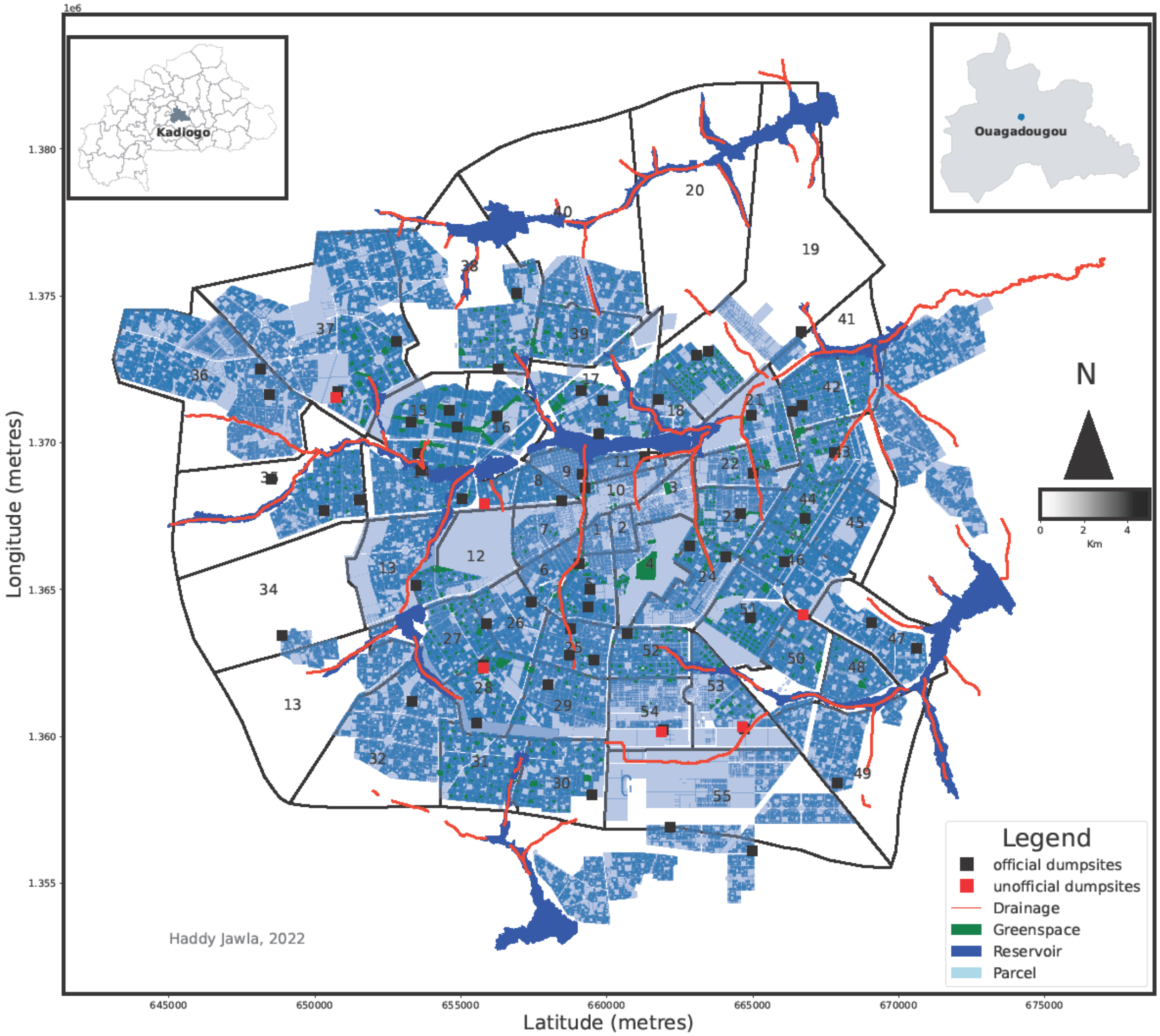


Figure 5: Map of the City of Ouagadougou with its districts (polygons), sectors (numbers), green spaces, drainage system, reservoirs, elements, and waste landfills (official & unofficial).

The annual rainfall has increased significantly over the past decade (Figure 6). The increasing frequency and intensity of precipitation events during the June-September season mainly contribute to an increase in floods and water stagnation in the city. The flash floods are caused by local heavy rain events (HRE), which lead to the inundation of streets and buildings, sometimes reservoir spillage before the stormwater reaches a watercourse. An increased frequency and intensity of HRE and the ongoing urbanization may further increase the risk of flash flooding.

Under climate change, future global warming scenarios reveal that HRE will be more intense, and flooding may likely be exacerbated in flood-prone areas of the city. There is a need to improve the disaster risk management schemes and early warning systems and guide decision-making in implementing flood-control infrastructure projects at the city level and beyond. Currently, warnings for pluvial floods are mostly limited to information on rainfall events over larger areas (i.e., no given rain intensities and durations), which is often not detailed enough to protect people and goods effectively. A proof-of-concept of effect-based monitoring of urban flood events using social media outlets is necessary for Ouagadougou.

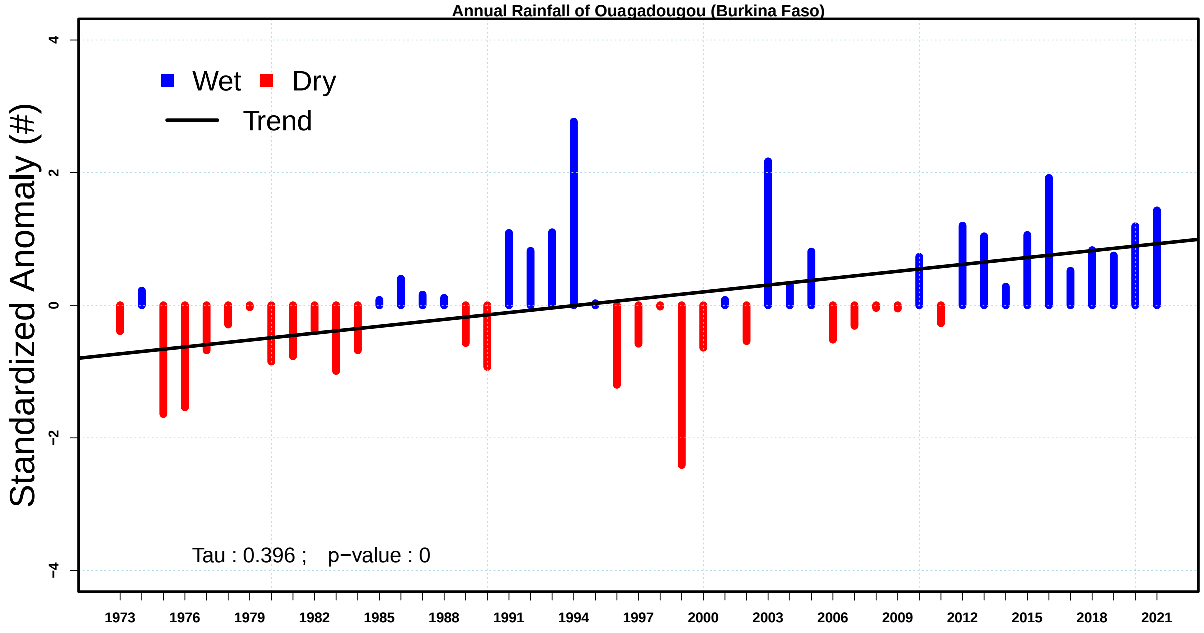


Figure 6: Interannual variability and rainfall trend observed over the City of Ouagadougou (1973-2021). The red (blue) bars denote dry below (above) average years. The tau parameter and p-value indicate the statistical significance level of the trend.

3.2 Twitter Data Extraction

Twitter is one of the most popular online social networking platforms, with many active users who freely post their thoughts, expressions, ideas, or beliefs on a particular topic. These posts can be viewed only by their followers or by querying on the search bar. Twitter data has attracted significant academic attention as a source for developing empirical insights into collective human behavior.

Accordingly, there are multiple ways to retrieve data from Twitter, with no consensus among researchers regarding standard data collection procedures (Rachunok et al., 2022). A critical limitation of this application program interface (API) is that the acquisition is restricted to up to 1% of the entire public stream of tweets, and the outcome is random samples (Oliveira et.al, 2016). Although large amounts of Twitter datasets are freely available online from various sources, it was observed that the datasets on crises and disaster situations were very few.

The API can crawl tweets, but the limitation is not having access to historical data. Therefore, instead of using the Twitter API, we developed web scraping codes based on the python programming language and its libraries with the Mozilla Firefox web browser to enable us to scrape the historical twitter data from January 2015 to December 2020. This was done using our authentic username and person to access the Twitter website. We used hashtags to search for tweets related to flood and inundation in Ouagadougou. The search was done in English and French and later saved in a database. Each Tweet contains multiple fields, but for this research, only a subset of the areas was extracted, namely

- the “date,” “time,”
- the user’s name, and
- the tweet, message itself.

These twitter alphanumeric datasets are saved in an ASCII format with a “.csv” extension and stored in a local database. The web scraping code is available in **Annex 1**.

3.3 Rainfall data

The daily rainfall dataset was collected from “*Agence Nationale de la Météorologie*” of Burkina Faso from January 2015 to December 2020. The dataset was quality controlled and formatted following the quality control procedure described by Salack et al. (2018). This quality control consists of checking erroneous measurement values (e.g., harmful precipitation, temporal sequences with the same measurement value), dates, and coordinates. The generation of multiple data plots enabled a comprehensive visual inspection of each time series. In combination with local meteorological knowledge and experience, the outliers, caused mainly by data entry typesetting errors, are identified and deleted. The Daily rainfall records are not interpolated when documents are missing. The quality-controlled daily rainfall data is used to assess the Twitter data performance and the relation between the number of tweets, the onset of floods, and the occurrence of heavy rain events.

3.4 Official Disaster Reports

For the cross-validation of the Twitter data the world bank report on '*FLOOD-RESILIENT MASS TRANSIT PLANNING IN OUAGADOUGOU*' was obtained from the internet in pdf formatted. Then the essential dates of the flood were extracted the amount of rainfall was saved in the database created CSV format. The data provided in the report was from 2015 to 2019. This data was plotted with the Twitter data and that of the rainfall data to assess data performance between the two datasets.

3.5 Landfill Data Collection

We also made use of the ODK client which is a set of tools geared towards simple, effective, and efficient data collection on mobile devices (Annex 2). The ODK helped to evaluate the relationship between flood and landfills (dumpsites). Google Drive was used as our server to store the official and unofficial geolocation data and areas of all the landfills in Ouagadougou and retrieve it in a real-time server to collect and store.

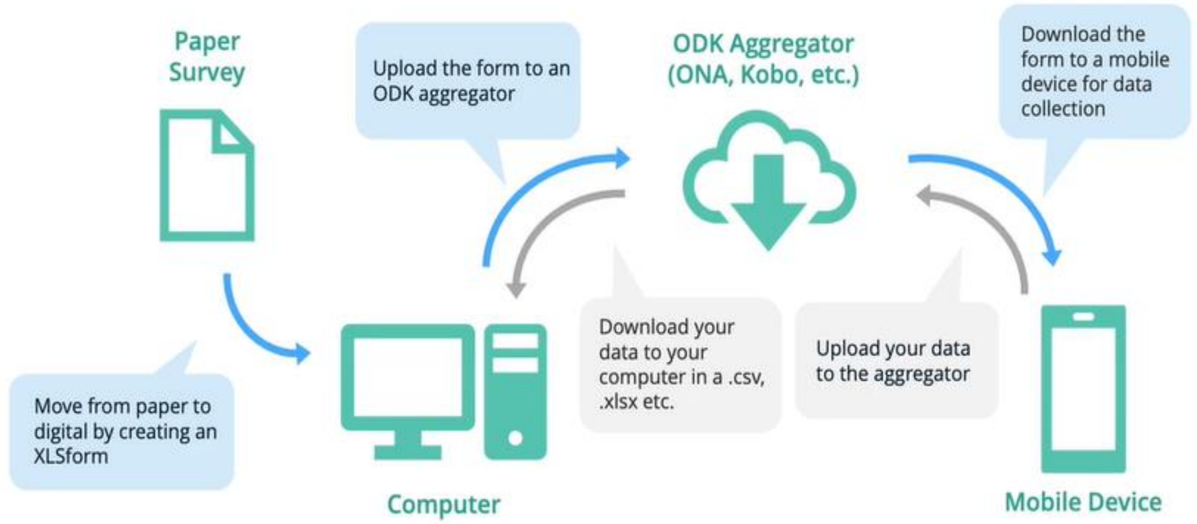

Paper Survey
Upload the form to an ODK aggregator
ODK Aggregator (ONA, Kobo, etc.)
Download the form to a mobile device for data collection
Move from paper to digital by creating an XLSform
Computer
Download your data to your computer in a .csv, .xlsx etc.
Upload your data to the aggregator
Mobile Device

Figure7: Open data kit development process .source: *https://m.facebook.com/surveycto.odk.kobo/photos/a.107793574465278/1518513300595*

3.6 Text Mining and Analysis

3.6.1 Data Pre-processing

Step1: Stop words a, links, slang words, and punctuation removal

Pre-processing is an important phase in text processing as the original tweet extracted may contain all sorts of symbols, slang words, improper grammar, etc. After scraping, the data preparation, in this case, includes removing punctuations, stop-words, numerics, logos, hashtags, and usernames from the tweet. Duplicated tweets are also dropped to avoid redundancy. Converting all the characters to lowercase and removing URLs and links was done using a regular expression module. This is done to make the input data easier to decode and interpret by the machine learning algorithm. All the punctuation such as full stop, comma, brackets, and stop words are removed using the NLTK library, an easy-to-use platform for Python programs to work with the human language.

Step 2: Tokenization

This is just a process of breaking up a piece of text into small pieces (tokens), such as sentences and words using the NLTK. NLTK Tokenization is used for parsing a large amount of textual data into parts to perform an analysis of the character of the text It works by separating words using spaces and punctuation. In this project, we used the “nltk. word tokenize ()” function will be used because it yields a list of every single word in the text

Step 3: Lemmatization

We used one of the natural language processing methods called lemmatization. which is used mainly for vocabulary and morphological analysis of words normally aiming at removing the inflectional words and returning the actual word or dictionary form of the word. After the tokenization is done using NLTK. The words are lemmatized to represent a correct meaning, this would help reduce our data and size and produce positive results and make the program run fast and smoothly. For this project, the WordNet Lemmatizer from the NLTK toolkit gives meaning to the tokenized data.

3.6.2 Sentiment analysis

sentiment analysis is done to Identify the mood or opinion of a speaker/writer feels on a particular topic. In this project, we, want to assess the sentiments attached to tweets about flood floods in Ouagadougou. In this approach, natural language processing (NLP) has been designed and implemented, and sentiment polarity is assigned to the tweets using the TextBlob module. The module is used for sentiment analysis and for calculating the polarity of tweets. The positive value of the other module is Matplotlib, which is stilted for calculating the percentage and drawing the bar chart graph in three different colors which represented the polarity of each tweet and calculates all tweets (positive, negative, and neutral) to make word clouds.

***Word cloud** is a technique for visualizing frequent words in a text where the size of the words represents their frequency.* A word cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance. We used a word generator module in this approach to display the positive, negative, and neutral tweets separately.

3.6.3 Feature Extraction

The feature selection method performed a key role in sentiment analysis for classification algorithms' performance based on the feature selection method. This is where the Twitter dataset is vectorized and trained to extract the test data needed for the text classification. In this paper, we used the TF-IDF (Term Frequency-Inverse Document Frequency), a combination of two individual metrics, TF and IDF, respectively (Equations 1 & 2). TF-IDF is used when we have multiple documents. It is based on the idea that rare words contain more information about the content of a document than words that are used many times throughout all the documents.

The selection of the appropriate features from data is crucial, has proven to be a significant issue, and is always a key objective for researchers. Parts of the raw texts were generated using the widespread approach TF-IDF weighting (Equation 3) with the help of the Scikit-learn library. This weight indicates the importance of a word to a document in a corpus. By this means, the whole training dataset can be represented as a matrix, where the columns correspond to the unique words in the entire corpus and the rows correspond to the documents. From our dataset of 945 tweets, we will create a train/test split of our dataset, where 25% of the dataset will be used for testing based on our evaluation strategy, and the remaining will be used for training the classifier. The **TfidfVectorizer** was used to tokenize documents, learn the text, and inverse document frequency weightings, and encode new documents.

TF and IDF are calculated with the following formulas:

$$TF(t,d) = \frac{\text{number of times t appears in d}}{\text{total number of terms in d}} \tag{1}$$

$$IDF(t) = log\frac{N}{1+df} \tag{2}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t \tag{3}$$

Where *d* refers to a document, *N* is the total number of copies; df is the number of records with the term t. TF-IDF is a word frequency score that tries to highlight more interesting words res has the effect of highlighting words that are distinct in a given document. The bag of Words model is useful because it converts the text into a vector, which helps keep a count of the total frequency of most commonly occurring used words. A bag of words represents text by giving us information about the most frequent word in the document. It involves two things. 1. A vocabulary of known words. 2. A measure of the presence of known words. The model is only interested in the frequency with which the text appears in the document, and not the location. 'TfidfVectorizer' is a function, which is a part of the 'sklearn' library that is used in creating the TF-IDF model

3.6.4 Logistic Regression

One of the most critical components in developing a supervised text classifier is the ability to evaluate it. We need to understand if the model has learned sufficiently based on the examples that it saw to make correct predictions. Logistic regression is a statistical method used for building machine learning models where the dependent variable is dichotomous: i.e., binary. Logistic regression is a simple and easy-to-understand classification algorithm, and Logistic regression can be easily generalized to multiple classes.

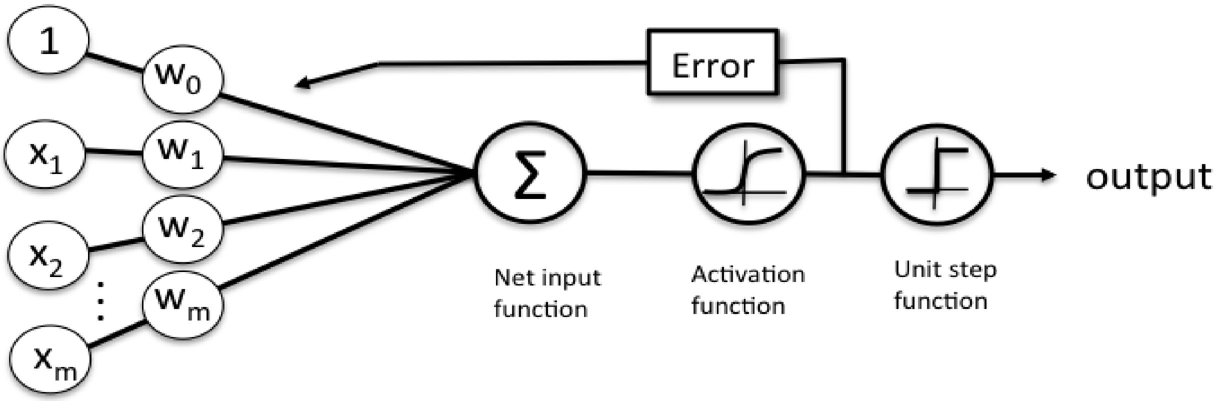


Figure 8:Schematic of a Logistic regression classifier. Source: https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide

The logistic function, commonly known as the sigmoid function, is the foundation of logistic regression. It takes any real-valued integer and translates it to a value between 0 and 1. A linear equation is used as input, and the logistic function and log-odds are used to complete a binary classification task. We will divide the dataset into two subsets: train and test. To perform the train-test split, we'll use Scikit-learn machine learning.

- **Train subset** – we will use this subset to fit/train the model.
- **Test subset** – we will use this subset to evaluate our model.
- After diving into the dataset, let's move on to the next phase of feature scaling.
- **Feature scaling** is required to put all features into the same range, regardless of their relevance.
- We bring all the features into the same range using feature scaling. There are many ways to do part scaling like normalization, standardization, robust scaling, min-max scaling, etc. But here, we will discuss the Standardization technique that we will apply to our features.
- In standardization, features will be scaled to have a mean of 0 and a standard deviation of 1. It does not rise to a preset range.

To implement Logistic Regression, we will use the Scikit-Learn library. We'll start by building a base model with default parameters, then look at how to improve it with Hyperparameter Tuning.

After training our model on the training dataset, we used our model to predict values for the test dataset and recorded them in the "*y_pred_basemodel*" variable.

To evaluate our model's performance, we will use the "f1 score" (Equation 4) as this is a class imbalance problem. Using accuracy as a performance metric is not good. Also, we can say that the f1 score is the go-to metric when we have a class imbalance problem. The formula for calculating the F1 score is as follows:

F1 Score = 2*(Recall * Precision) / (Recall + Precision) ………. (4)

Precision is the ratio of accurately predicted positive observations to the total indicated positive comments.

Precision = TP/TP+FP

The recall is the ratio of accurately predicted positive observations to all observations in actual class – yes.

Recall = TP/TP+FN

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

3.7 Deployment of the Open Data Kit Questionnaire

The objectives of deploying the ODK questionnaire were i) to geolocate the landfills available, and ii) to estimate the surface areas of these landfills in the City of Ouagadougou. This was achieved by creating an xlsform that entails all the questions required for the survey. The server used for this project was google drive. The data collected for the landfills were three:

- Landfills geolocation
- Area covered by the landfills
- Picture of the landfills.

We downloaded the ODK collect application to enable it to work on our android phones and download the xlsform in other to fill and send the data in real-time.

CHAPTER 3: RESULTS AND DISCUSSION

Urban Flood is usually based on many factors, However, time and again most flood risk assessment focus on just a few elements of exposure and disregards other elements that form critical components in flood risk communication. The objective of this paper is to assess the reliability of crowdsourcing in Urban floods. The analysis is based on historical flood data from Twitter, landfill information, and national disaster reports on flood inundation to develop a monitoring and impact-based communication systems.

4.1 Official data on floods and Communications

In the period January 2015 to December 2020, **official** datasets on the events of the significant flood were released by the National Council for Emergency Relief and Rehabilitation (CONASUR) and the World Bank (World Bank, 2021) for the City of Ouagadougou:

- On **24 June 2015:** Heavy rain fell on Ouagadougou and some localities in Burkina Faso. With a volume of between 67 mm and 79.8 mm, this rainfall caused four deaths, including three children, and extensive material damage.
- On **10 July 2016:** Heavy rain fell on Ouagadougou and several regions of Burkina Faso. At the Yalgado Ouedraogo University Hospital and the 14 General Directorate of Land and Maritime Transport (DGTTM), this rain caused enormous material damage. In several other areas of the capital, 1,488 people were forced to abandon their homes invaded by rainwater.
- On **19-20 July 2016:** Heavy rain fell on Ouagadougou and some localities in Burkina Faso, causing extensive damage. 51.4 mm was recorded in Ouagadougou Airport, 55.3 mm in Somgandé and 97.6 mm in Pô (~150 km away from Ouagadougou).
- On **9 August 2016:** 88 mm in less than 12 hours; most affected secteurs of Ouagadougou were 1, 2, 3, 4, 6, 9, 12, and Rimkieta neighborhood.

  On **18 May 2017:** 96.7 mm in less than 12 hours; affected districts 4 and 12 and the Kouritenga, Bissighin, and Rimkieta neighborhoods.
- On **25 July 2018:** Around 100 mm of rainfall fell on Ouagadougou in 24 hours; eight secteurs (1, 2, 3, 6, 7, 10, 11, and 12) were affected by flooding.

- On **July 4, 2019:** 89 mm were recorded by rain gauge located at Ouagadougou Airport, 75 mm in Somgandé, a 2-hour episode.

West Africa experienced the worst floods over the past 30 years, with three deaths in Burkina Faso, 23 in North Togo, and 46,000 displaced people, including 26,000 in Burkina Faso and 14,000 in Togo. In the same year, 17,689 ha of flooded crops and a production loss of about 13,500 tonnes were recorded in Burkina Faso" (Eilander et al., 2016). This information provides enough data to characterize the causes (e.g., heavy rain events), the magnitude and extent of the flood events, and the effects. These events often cause severe stress and trauma sentiments mainly felt by the people living in poor communities with limited resources. However, they are usually accounted for in data analysis and official communication reports. Due to Twitter API limitations, Web scraping was used to extract from the Web pages of Twitter all these attributes. Web scraping technique has enabled us to have access to the historic Twitter data without any limitation or interference. The dataset searched for was given which was also verified manually. This shows that web scraping is a very useful and reliable toolkit in flood data collection. Results from the previous research suggest that Web scraping is 40 times faster than Twitter API to obtain the user_id and tweet_id, Web scraping is also used; thus, a Twitter API call consists of local processing (Web scraping) and an API request (Dongo et.al, 2020). In other words, the tweets contained valuable information and data like those officially reported by CONASUR and the world bank report over the City of Ouagadougou with a higher level of similarity. Table 1 shows the number of tweets in the context of the flood events reports in the official reports of the World Bank (World Bank, 2021)

Table 1: Number of tweets in the context of the flood events reported in the official documents

Flood event date	***Rainfall amount***	***Number of Tweets***
24 -26-26June 2015	*67 mm-79.8 mm*	*25*
10-11 July 2016	*Heavy rain*	*26*
19-20-21 July 2016	*51.4 mm*	*78*
9 August 2016	*88 mm*	*34*
18-19 May 2017	*96.7 mm*	*54*
25-26 July 2018	*100 mm*	*32*
July 4, 2019	*89 mm*	*18*

From this table above, we can further elaborate and compare the two different crowdsourced data. the results show that the majority of flood events correspond directly to the highest tweets posted either. However, the overall effect of multiple terms combined provided more significant and sensible results in evaluating the association between flood tweets and disaster reports

4.2 Sentiment Analysis

Sentiment analysis analyzes attitudes, emotions, or opinions that could be extracted from tweets about flooding. In the context of urban flooding in the City of Ouagadougou, we analyzed Twitter posts (tweets) and classified them into three categories: positive, neutral, and negative. The sentiment values will also define the polarity of the tweets. Making use of supervised machine learning approaches such as the logistic regression to give the accuracy and precision of the Twitter dataset, the analysis results are for floods in Ouagadougou and visualized as a bar chart (Figure 78). Over the five years of tweets, the results show that 95.68% of the words were classified as neutral, 3.28% were positive, and 1.04% were negative sentiments.

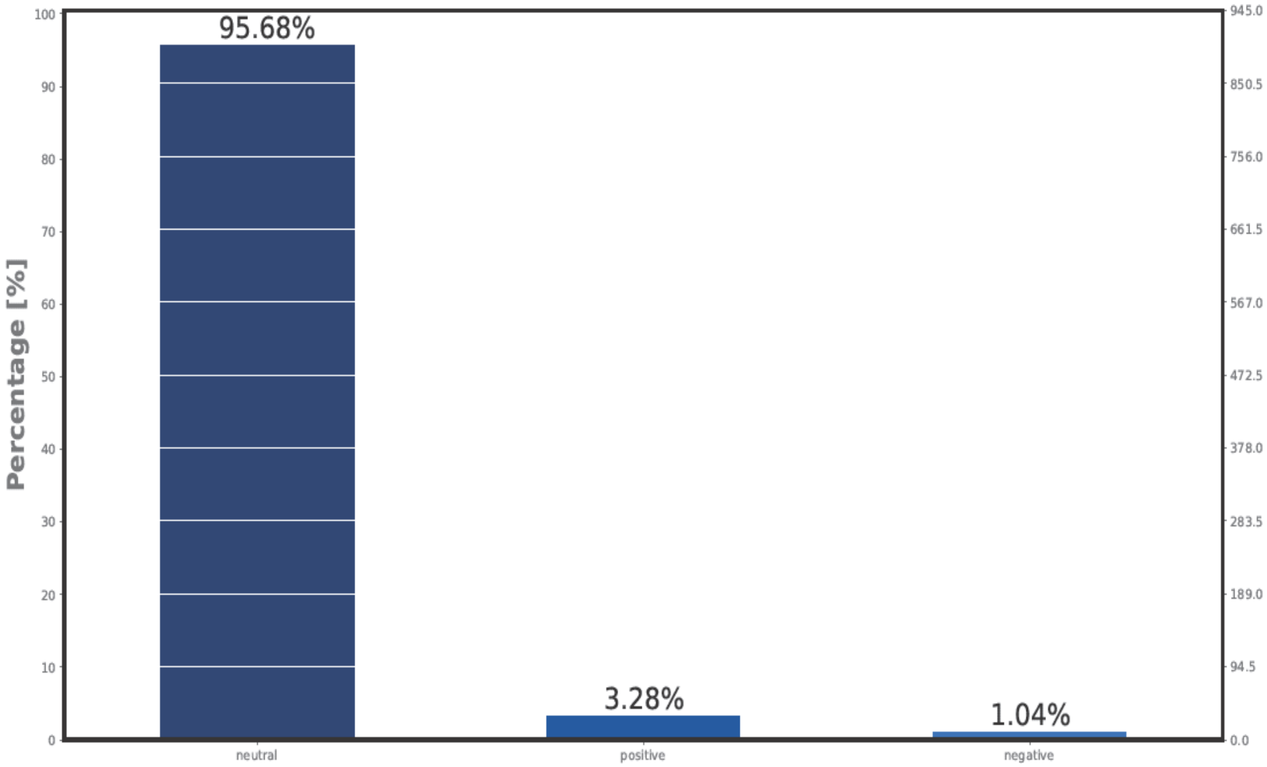


Figure 9: The sentiment analysis of the tweets

4.2.1 Neutral sentiment

After removing keywords such as “Ouagadougou” and “inundation,” which were the keywords for scraping, we can visualize the remaining text in a word cloud that showcases important word patterns and frequencies. Every word that appeared I n the tweets was calculated for its frequency of occurrence. The most frequent words in the neutral category are “*Pluie*”, “*grosse pluies*”, “*forte pluie*”, “*pluie diluvienne*”, “*nuit*”, and “*juillet*” (Figure 9). When they are put together, these words provide insight into the triggers of the flood events (e.g., heavy rainfall, extreme rain) and the timing (e.g., Nighttime and July). are repeated quite often. This gives real-time information about the situation. It requires intervention to participate in rescue and rehabilitation efforts.

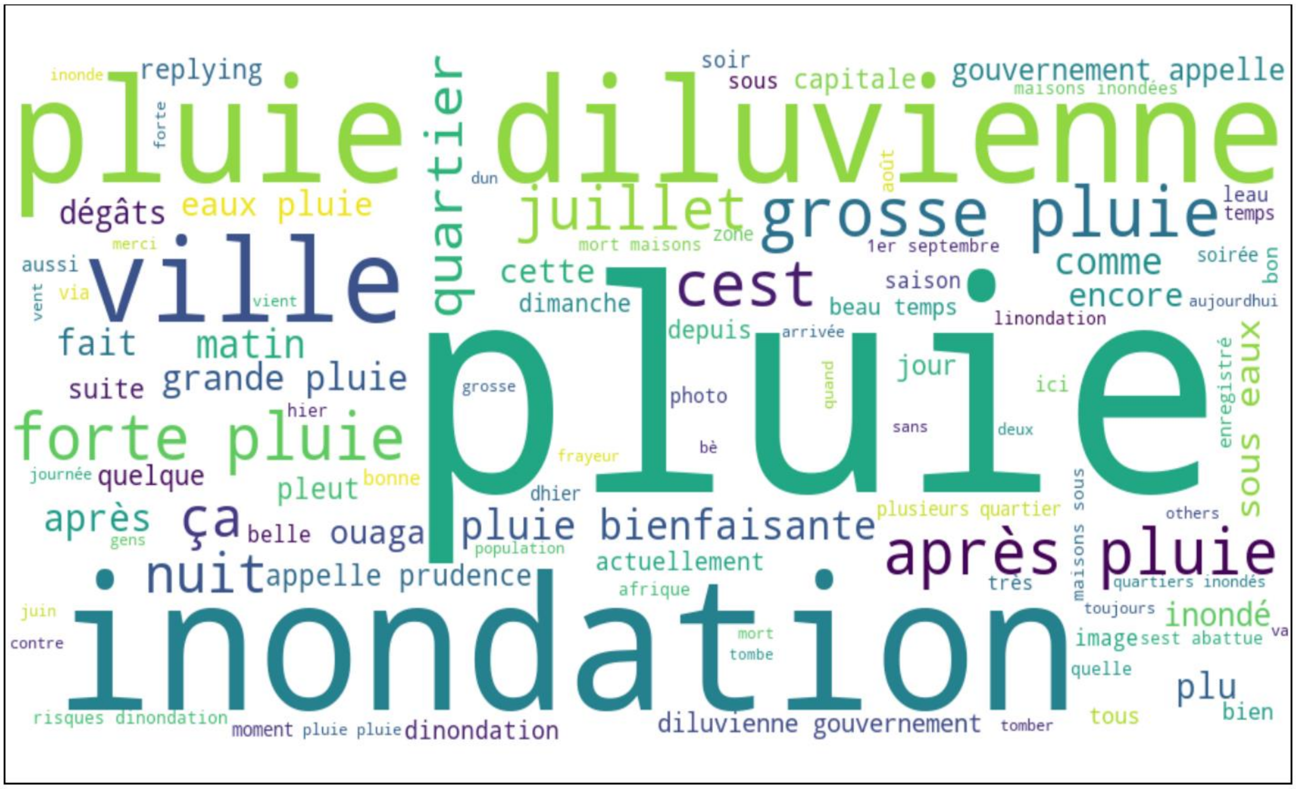


Figure 10:The most frequent words in the neutral category of tweets sent out between January 2015 and December 2020.

4.2.2 Negative sentiment

The category of negative sentiment tweets the most frequent are “*Pluie*”, “*insolite*”, “circulation”, “*crocodiles*”, “*deserte*”, “*effarement*”, “*eblouissante*”, “*matin*”, “*violent*” (Figure 10). The resulting echoes of these words indicate the effects or consequences of flood events in the City of Ouagadougou. Examples of what “eyewitness” reported from the CONASUR documents include the difficulty in traffic circulation due to road blockage, and in the natural parc of Bangr-Weogo, the” crocodiles” are generally seen outside the parc along the mains road about the parc when a flood occurs.

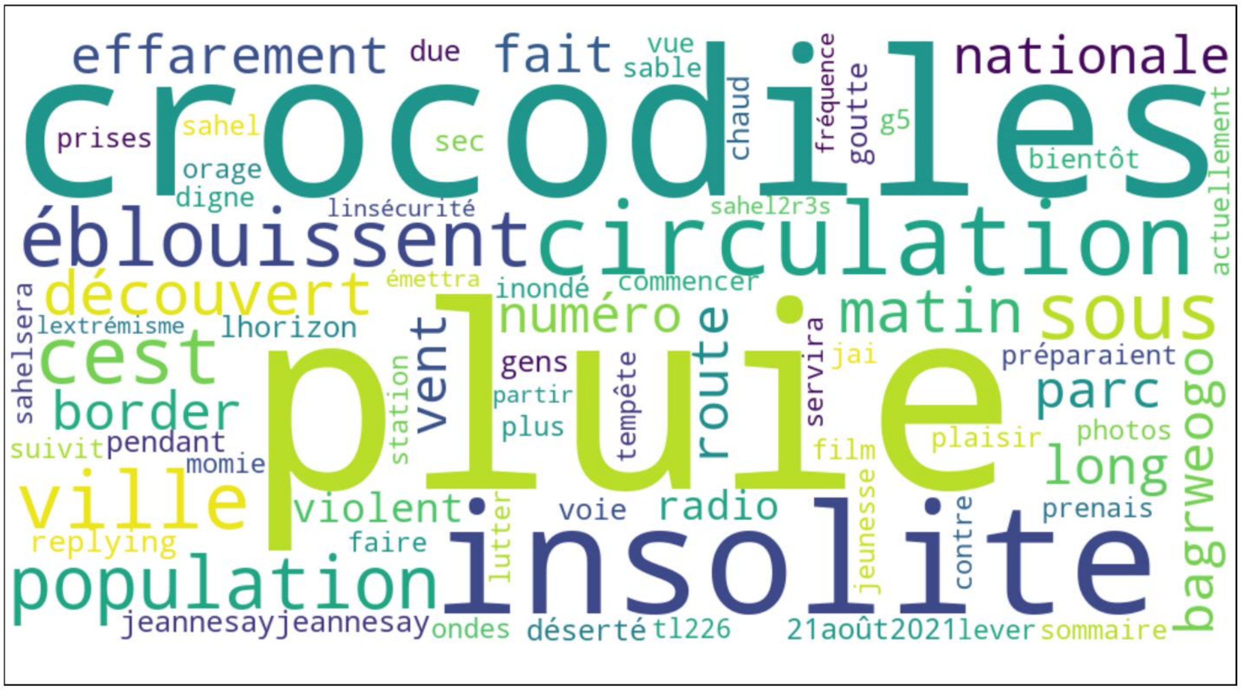


Figure 11: The most frequent word in the negative category of tweets sent out between January 2015 and December 2020.

4.2.3. Positive sentiment

Following Figure 11, the most frequent words in this category are *"Pluie*," "*bonne*," fine," "*super*," "*bienfaisante*," "urgent," and "matin." They tend to portray a more relaxed context after the flood events with some similar sentiments to the negative and neutral categories. Therefore, the positive feelings are mixed across neutral and negative emotions, although they sometimes contain a relaxed post-event context.

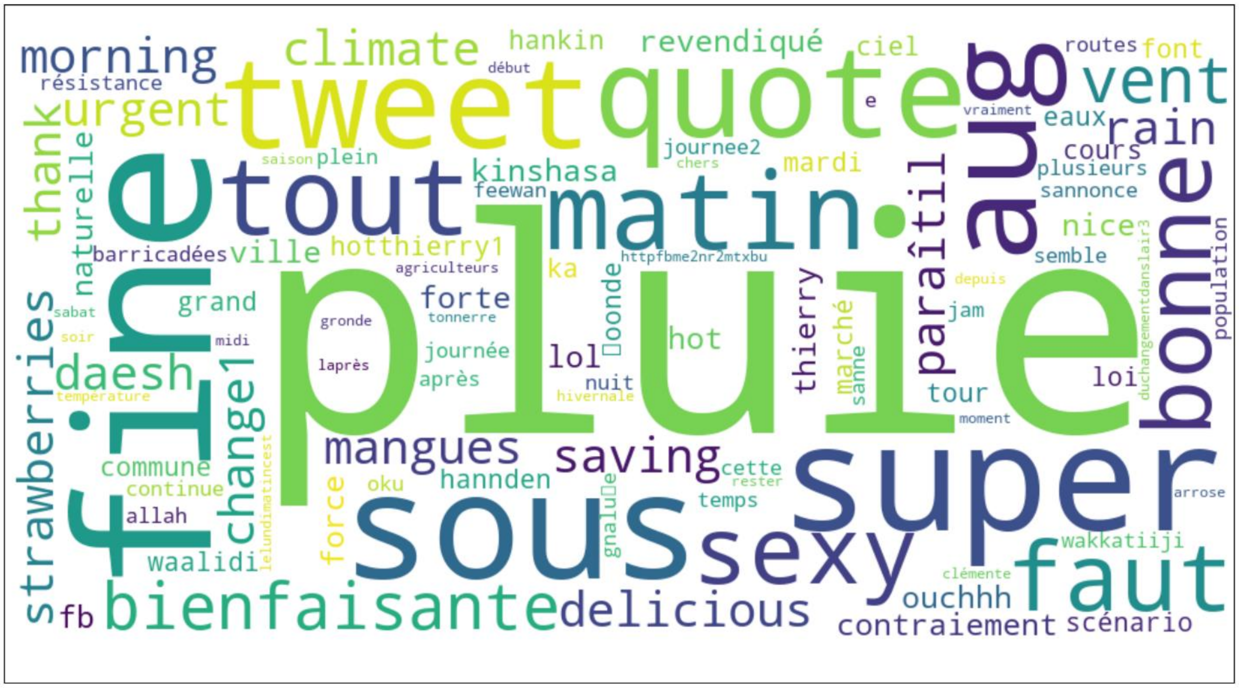


Figure 12: The most frequent word in the positive category of tweets sent out between January 2015 and December 2020.

From the results obtained we can conclude that Twitter data is a valuable asset for data collection, effect-based monitoring, and communication of urban floods which is visible in the word cloud generated. In a similar study, the first noticeable aspect of our qualitative analysis is that the most frequent terms identified in these word clouds reveal information related to weather warnings in the context of floods, as well as global climate warming and activism in the case of heat waves(Ponce-López et al., 2022.)

4.3 Performance Assessment of Twitter Data

4.3.1 Evaluation of the logistic regression

The effectiveness and performance of the model were evaluated by selected standard metrics such as accuracy, f1-score, precision, and recall. The classification accuracies obtained from the logistic regression were consistently high. The percentage accuracy of correctly classified using 945 tweets was at a 93.10% rate.

Indra et al. (2017) used the Logistic Regression algorithm for the text classification task in research that aims at developing a web-based application that can classify tweets of netizens into these four categories of topics. The set of features vector was applied. The trained classifier was evaluated using 1800 tweets with 450 for each case. The results showed that the accuracy of tweets classification in the selected topics is 92% which is considered very high (Indra et al., 2017).

In another study, feature selection methods Hindi-English language identification task was used to extract the most relevant for an efficient model. In this article, we apply different feature selection algorithms across various learning algorithms to analyze the algorithm's effect and the number of features on the performance of the task. Finally, an exhaustive comparative analysis is put forward concerning the overall experiments conducted for the job. The Logistic Regression classifier shows the best performance in the top-k selection and partially in the case of forwarding feature selection (Ansari et al., 2020)

4.3.2 Raingauge and Twitter data

The results generated from figure 12 below is a bar chart showing daily tweet counts plotted against the daily rainfall. Although the value of tweets is relatively low we deduced a positive correlation between the two datasets., we observed that the higher the number of tweets, the more elevated the rainfall amount. Thus, the tweets are posted more often when the rainfall is high, and this usually high amount of rain is what causes a flood.

Secondly, we observed that during the start and the end of the rainy season is when we have active posts of tweets on floods in Ouagadougou. From the plot, we can see that there are almost less than 5 five posts during the dry season. This result here can be used to forecast the onset and end of the rainy season.

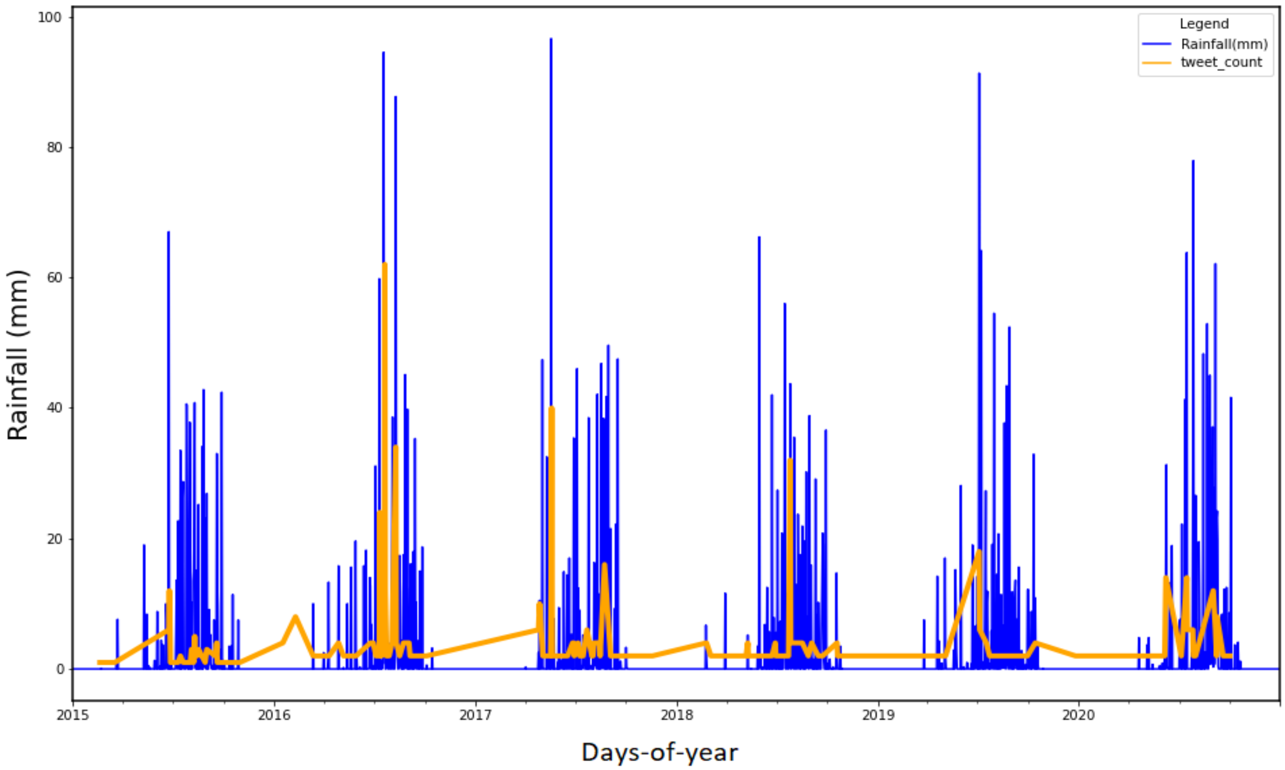


Figure 13:The plot above shows the tweets and the rainfall data.

4.3.3 Geo-location of Landfills in the City of Ouagadougou

The results obtained using the ODK to geolocate the landfills in Ouagadougou show how they can be a trigger of floods in Ouagadougou as most of these landfills are surrounding the reservoirs and drainages and some are located in the low-lying areas of the city(Figure 13). Municipal solid waste (MSW) landfills pose a long-lasting risk for humans and the environment. While landfill emissions under regular operating conditions are well investigated, landfill behavior and associated emissions in case of flooding are widely unknown, although damages have been documented (Laner et.al,2009). Waste disposal sites are mostly located in lowland areas close to residential areas inducing a long-term risk of potential environmental contamination due to flooding. the study outlines that in case of flooding or erosion of waste disposals, the hazardous waste released to the environment could lead to partly tremendous damages (Neuhold & Nachtnebel, 2011). Many of today’s cities expand at a rate much faster than flood management plans are developed and infrastructure is installed. Urban floods and solid waste management are only two of its numerous difficulties. Each sector frequently lacks adequate capacity, and the two interact as solid waste builds up and obstructs existing drainage routes, causing flooding, damage, and public health issues. By clogging drainage systems and accumulating debris, poor waste management can worsen the effects of urban flooding.

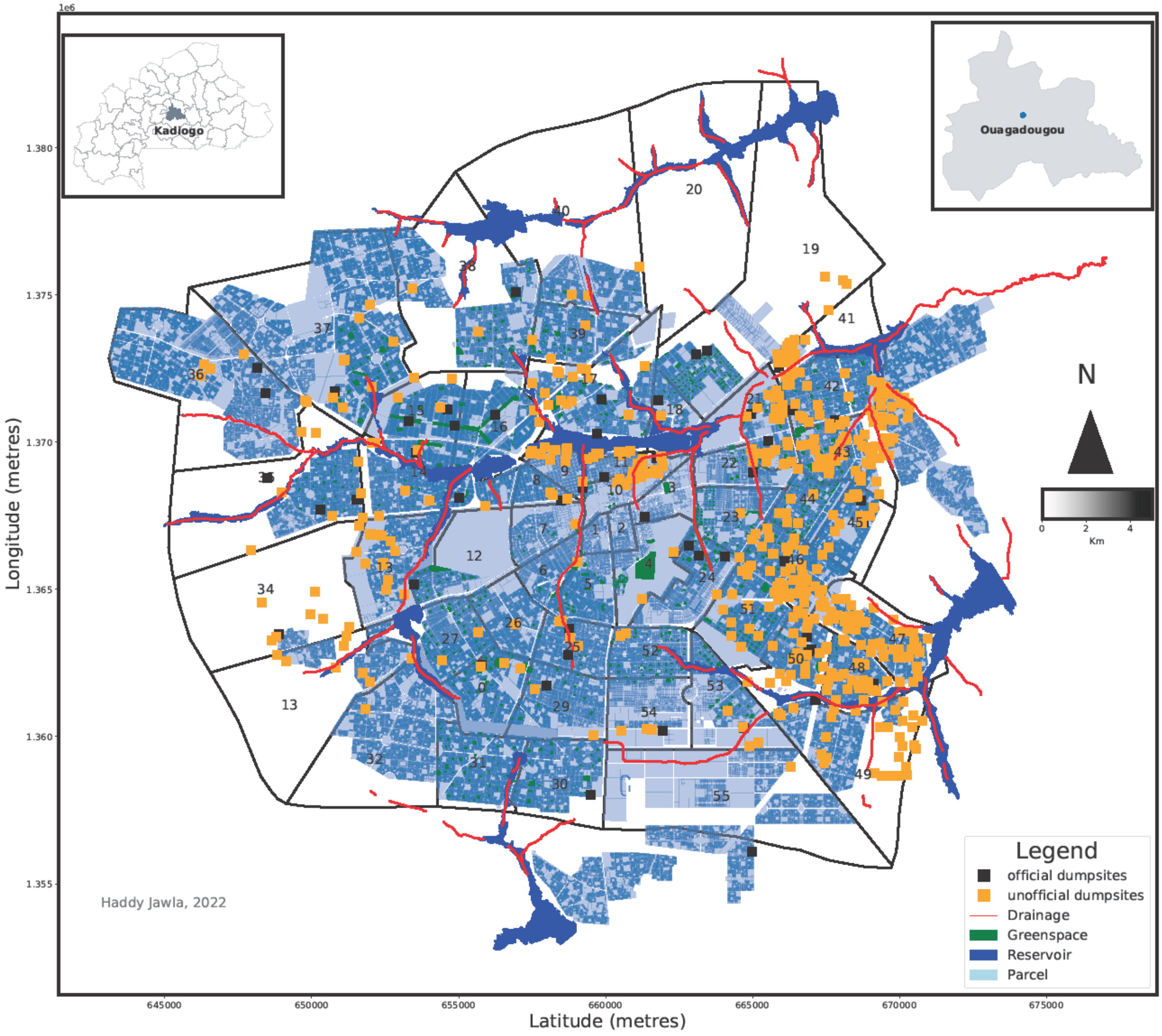


Figure 14: Map of the City of Ouagadougou with its districts (polygons), sectors (numbers), green spaces, drainage system, reservoirs, elements, and landfills (official & unofficial).

A study done by Samari in 2011,he identified these sectors (3, 4, 10, 11, 12, 23, 24, 13, 21, 22, 19, 18, 14, 15) as flood-prone areas in the city of Ouagadougou. These sectors are also identified in figure 14 as areas highly dense with dumpsites, thus we can conclude by saying that dumpsites are triggers of fthe lood.

[redacted] CONCLUSION AND PERSPECTIVES

A fundamental assumption when dealing with data obtained from social media is that it cannot be considered as reliable as reports from professional observers. Henceforth, it can be concluded that Twitter messages with some considerations are informative enough to be used to estimate flood situations. The text classification technique also proved that tweets contain valuable information and show how the use of Twitter data and instrumental data coupled with reports can give us helpful information not only about floods but shows how using Twitter data and instrumental data associated with words can provide us with valuable details only about floods but also about rainfall onset. These data can be trained using machine learning algorithms to make times series forecasting can conclude y saying that Twitter data is vital in flood risk management. Although there are several data analysis challenges encountered that need further research. We also faced challenges in the limitation of the Twitter dataset. From the results obtained Twitter dataset can be used by the disaster management agency for early intervention and researchers working in this area.

ODK has the potential to have a significant influence on data collection in the future, particularly in development applications where budgets are constrained and locations may be distant, but where mobile phone use is rising quickly due to the increase in service coverage. The ODK was useful in this research as most of our stakeholders had android phones to collect data, which we were receiving in real-time to develop our maps.

We are at a time of intense urbanization. Globally, the capacity and services of cities have been aggravated by the inflow of people into metropolitan regions. Urban floods and solid waste management are only two of its numerous difficulties. Each sector frequently lacks adequate capacity, and the two interact as solid waste builds up and obstructs existing drainage routes, causing flooding, damage, and public health issues that can worsen the effects of urban flooding. The relationship between the city of Ouagadougou's solid waste management and flooding, however, is not well understood. Lack of knowledge by residents on their role in municipal solid waste management was an issue that contributed to the failure in bringing about changed attitudes toward MSWM. Providing information to make people aware of their roles in MSWM as well as sensitization on proper disposal practices can reduce the environmental impact of flooding. In the sense of landfills as triggers of urban floods, future work can be done to derive indices from these landfills and use them in the flood risk equation to generate flood risk maps of the impact of landfills on urban flooding. The study can also use Twitter geolocation and compare if the tweets are coming from areas with more landfills to also help in monitoring and impact-based forecasting.

[redacted] REFERENCES

Anbalagan, B., & Valliyammai, C. (2017). #ChennaiFloods: Leveraging human and machine learning for crisis mapping during disasters using social media. *Proceedings - 23rd IEEE International Conference on High Performance Computing Workshops, HiPCW 2016, December 2016*, 50–59. https://doi.org/10.1109/HiPCW.2016.17

Anokwa, Y., Hartung, C., Brunette, W., Borriello, G., & Lerer, A. (2009). Open Source Data Collection in the Developing World. Computer, 42.

Ansari, M. Z., Ahmad, T., & Fatima, A. (2020). *Feature Selection on Noisy Twitter Short Text Messages for Language Identification*.

Dokken, D. (2012). special report of the intergovernmental panel on climate change managing the risks of extreme events and disasters to advance climate change adaptation.

Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y., & Barrios, S. (2020). Web Scraping versus Twitter API: A Comparison for a Credibility Analysis. *ACM International Conference Proceeding Series*, 263–273. https://doi.org/10.1145/3428757.3429104

Eilander, D., Trambauer, P., Wagemaker, J., & van Loenen, A. (2016). Harvesting Social Media for Generation of Near Real-time Flood Maps. *Procedia Engineering*, *154*, 176–183. https://doi.org/10.1016/j.proeng.2016.07.441

Frigerio, S., Schenato, L., Bossi, G., Mantovani, M., Marcato, G., & Pasuto, A. (2018). Hands-on experience of crowdsourcing for flood risks. An android mobile application tested in Frederikssund, Denmark. *International Journal of Environmental Research and Public Health*, *15*(9). https://doi.org/10.3390/ijerph15091926

Griscom, R. T. (2020). *Open Data Kit (ODK) for mobile linguistic metadata entry*. https://doi.org/10.5281/zenodo.3509475

Harrison, S. E., & Johnson, P. A. (2016). Crowdsourcing the Disaster Management Cycle. *International Journal of Information Systems for Crisis Response and Management*, *8*(4), 17–40. https://doi.org/10.4018/ijiscram.2016100102

Hazard Control, H. (2006). *Natural Disasters (*. http://www.ecri.org.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). *A Web Scraping Methodology for Bypassing Twitter API Restrictions*. http://arxiv.org/abs/1803.09875

Howe, J. (2006). The Rise of Crowdsourcing. Wired, 14. http://www.wired.com/wired/archive/14.06/crowds.html

Ijaz, S., Miandad, M., Mehdi, S., & Anwar, M. M. (2021). Spatio-temporal Analysis of Land Use Land Cover(LULC) in Abbottabad View project Tuberculosis View project. https://doi.org/10.17576/geo-2021-1701-01

Indra, S. T., Wikarsa, L., & Turang, R. (2017). Using logistic regression method to classify tweets into the selected topics. *2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, 385–390. https://doi.org/10.1109/ICACSIS.2016.7872727

Jeffrey-Coker, F., & Modi, V. (n.d.). Open Data Kit: Implications for the Use of Smartphone Software Technology for Questionnaire Studies in International Development.

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2019). Can volunteer crowdsourcing reduce disaster risk? A systematic review of the literature. In International Journal of Disaster Risk Reduction (Vol. 35). Elsevier Ltd. https://doi.org/10.1016/j.ijdrr.2019.101097

Karami, Amir & Shah, Vishal & Vaezi, Reza & Bansal, Amit. (2019). Twitter speaks: A Case of National Disaster Situational Awareness. Journal of Information Science. 016555151982862. 10.1177/0165551519828620.

Komolafe, A. A., Adegboyega, S. A. A., & Akinluyi, F. O. (2015). A review of flood risk analysis in Nigeria. In *American Journal of Environmental Sciences* (Vol. 11, Issue 3, pp. 157–166). Science Publications. https://doi.org/10.3844/ajessp.2015.157.166

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., Mach, K., Muir-Wood, R., Brakenridge, G. R., Kron, W., Benito, G., Honda, Y., Takahashi, K., & Sherstyukov, B. (2014). Le risque d'inondation et les perspectives de changement climatique mondial et régional. *Hydrological Sciences Journal*, *59*(1), 1–28. https://doi.org/10.1080/02626667.2013.857411

Laner, David & Fellner, Johann & Brunner, Paul. (2009). Flooding of municipal solid waste landfills - An environmental hazard?. The Science of the total environment. 407. 3674-80. 10.1016/j.scitotenv.2009.03.006.

Moreira, Ranieri & Degrossi, Livia & De Albuquerque, Joao. (2015). An experimental evaluation of a crowdsourcing-based approach for flood risk management. CIBSE 2015 - XVIII Ibero-American Conference on Software Engineering.

Myneni, Dr & Prasad, L.V. & Reddy, G.G.. (2017). Automatic assessment of floods impact using twitter data. International Journal of Civil Engineering and Technology. 8. 1228-1238.

Nath, B., Shri, K., Vaishno, M., Wankhade, M., Chandra, A., Rao, S., Dara, S., & Kaushik, B. (2017). *A Sentiment Analysis of Food Review using Logistic Regression*. *2*(7), 251–260. https://www.researchgate.net/publication/334654833

Neuhold, C., & Nachtnebel, H. P. (2011). Assessing flood risk associated with waste disposals: Methodology, application and uncertainties. *Natural Hazards*, *56*(1), 359–370. https://doi.org/10.1007/s11069-010-9575-9

OCHA (2021), West And Central Africa Flooding Situation (2021).

https://reliefweb.int/report/democratic-republic-congo/west-and-central-africa-flooding-situation-30-august-2021

Oliveira, Nuno & Cortez, Paulo & Areal, Nelson. (2016). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications. 73. 10.1016/j.eswa.2016.12.036.

Park, K., Choi, S. H., & Yu, I. (2021). Risk type analysis of building on urban flood damage. *Water (Switzerland)*, *13*(18). https://doi.org/10.3390/w13182505

Ponce-López, V., & Spataru, C. (n.d.). *Behaviour in social media for floods and heat waves in disaster response via Artificial Intelligence*.

Puttinaovarat, S., & Horkaew, P. (2020). Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques. *IEEE Access*, *8*, 5885–5905. https://doi.org/10.1109/ACCESS.2019.2963819

Rachunok, B., Fan, C., Lee, R., Nateghi, R., & Mostafavi, A. (2022). Is the data suitable? The comparison of keyword versus location filters in crisis informatics using Twitter data. *International Journal of*

Information Management Data Insights, *2*(1), 100063. https://doi.org/10.1016/J.JJIMEI.2022.100063

Rentschler, J., & Salhab, M. (2020). *People in Harm's Way Flood Exposure and Poverty in 189 Countries Poverty and Shared Prosperity 2020 Background Paper*. http://www.worldbank.org/prwp.

Ripberger, Joseph & Jenkins-Smith, Hank & Silva, Carol & Carlson, Deven & Henderson, Matthew. (2014). Social Media and Severe Weather: Do Tweets Provide a Valid Indicator of Public Attention to Severe Weather Risk Communication? Weather, Climate, and Society. 6. 520-530. 10.1175/WCAS-D-13-00028.1.

Salack S., Saley I. A., Lawson N. Z., Zabré I., & Daku E. K. (2018). Scales for rating heavy rainfall events in the West African Sahel. Weather and climate extremes, 21, 36-42.

Salack S., Sarr B., Sangare S. K., Ly M., Sanda I. S., & Kunstmann H. (2015). Crop-climate ensemble scenarios to improve risk assessment and resilience in the semi-arid regions of West Africa. Climate Research, 65, 107-121.

Salack S., Klein C., Giannini A., Sarr B., Worou O.N., Belko N., Bliefernicht J., Kunstmann H. (2016) Global warming induced hybrid rainy seasons in the Sahel. Environmental Research Letters, Volume 11, Pages 104008.

Salack S., Saley I. A., Lawson N. Z., Zabré I., & Daku E. K. (2018). Scales for rating heavy rainfall events in the West African Sahel. Weather and climate extremes, 21, 36-42.

Salack S., K. Hien, N. Z. Lawson, I. A. Saley, J.-E. Paturel, M. Waongo (2020). Prévisibilité des faux-départs de saison agricole au Sahel. In: Sultan B., Bossa A. Y., Salack S., Sanon M. (2020) Risques climatiques et agriculture en Afrique de l'Ouest. Edition de l'Institut de la Recherche pour le Development (1st Edition). p31-44.

Sanfo S., Neya O., Da S., Salack S., Amikuzuno J., Gandaa B. Z., Hackman K., P. (2022) Waste Recycling and Repurposing to Address SDG 11 in Burkina Faso: Do Multi-stakeholder Platforms Matter? *In*: Sustainable Development Goals Series, Sylvia Croese and Susan Parnell (Eds): Localizing the SDGs in African Cities, 978-3-030-95978-4, 496088_1 (Chapter 6).

Tavra, M., Racetin, I., & Peroš, J. (2021). The role of crowdsourcing and social media in crisis mapping: a case study of a wildfire reaching Croatian City of Split. *Geoenvironmental Disasters*, *8*(1). https://doi.org/10.1186/s40677-021-00181-3

Terpstra, Teun. (2012). Towards a realtime Twitter analysis during crises for operational crisis management.

UNDRR (2019).*"The human cost of disasters an overview of the last 20 years*. (2000- 2019)".

UNDP. Climate Change Adaptation: Burkina Faso, 2021. https://www.adaptation-undp.org/explore/western-africa/burkina-faso.

USAID (2022). Contry Profile: Burkina Faso

USAID (2012). Climate risks in food for peace and geographies: Burkina Faso.

Waongo M, Laux P, Kunstmann H. (2014) Adaptation to climate change: The impacts of optimized planting dates on attainable maize yields under rainfed conditions in Burkina Faso. Agricultural and Forest Meteorology 205 (2015) 23–39.

WMO/GWP. Urban flood risk management, a tool for integrated flood management, Associated Programme on Flood Management, World Meteorological Organization; 2008. https://library.wmo.int/doc_num.php?explnum_id=7342

WMO (2013), "Flood Forecasting and Early Warning", APFM Technical Document No. 19, integrated flood management tools series. Associated programme on flood management (WMO). (2013). https://www.floodmanagement.info/publications/tools/APFM_Tool_19.pdf

WMO(2017).The Associated Programme on Flood Management (APFM) is a joint initiative of the World Meteorological Organization (WMO) and the Global Water Partnership (GWP). (2017). www.apfm.infowww.gwp.org

World bank. Climate Change Knowledge Portal, 2021. https://climateknowledgeportal.worldbank.org/country/burkina-faso . Accessed: 15th May 2022.

World bank. *Flood-Resilient-Mass-Transit-Planning-in-Ouagadougou-2*. (2021).

https://openknowledge.worldbank.org/bitstream/handle/10986/35983/Flood-Resilient-Mass-Transit-Planning-in-Ouagadougou.pdf?sequence=1&isAllowed=y Accessed: 11 Febraury 2022.

Websites:

https://www.octoparse.com/blog/5-things-you-need-to-know-before-scraping-data-from-facebook#. Accessed: 12 February, 2022.

https://gs.statcounter.com/social-media-stats/all/burkina-faso. Accessed: 5th May, 2022.

https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ Accessed: 6th May 2022.

https://datareportal.com/reports/digital-2021-burkina-faso Accessed 23rd June, 2022

Thesis

Konstantinos Korovesis (2018). Sentiment Analysis for Tweets, Masters thesis, Athens University of Economics and Business, Athens, 42p.

[redacted]ANNEX 1: CODE FOR SCRAPING TWITTER

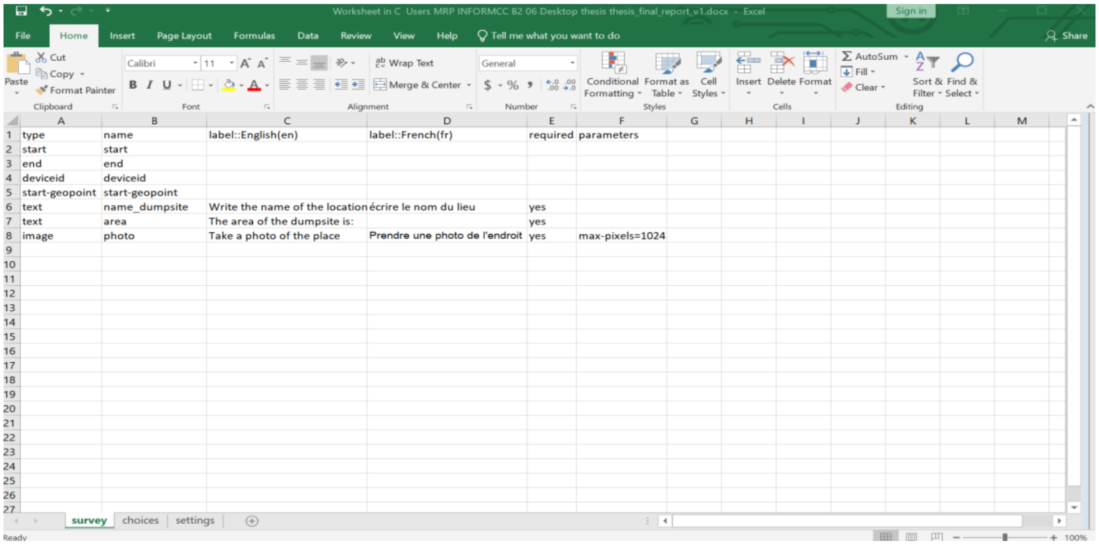
```
#!/usr/bin/env python

# coding: utf-8


# ### 1. Importing of Libraries

import re

import CSV

from getpass import getpass

from time import sleep

from selenium import webdriver

from selenium.webdriver.common.keys import Keys

from selenium.common.exceptions import NoSuchElementException

from selenium.webdriver import Firefox




# ### 2. Defining the data to be scraped

def get_tweet_data(card):

    username = card.find_element_by_xpath('.//span').text

    try:

        handle = card.find_element_by_xpath('.//span[contains(text(),"@")]').text

    except NoSuchElementException:

        return
    try:
        postdate = card.find_element_by_xpath('.//time').get_attribute("datetime")
    except NoSuchElementException:
        return

    comment = card.find_element_by_xpath('.//div[2]/div[2]/div[1]').text
    responding = card.find_element_by_xpath('.//div[2]/div[2]/div[2]').text
    text = comment + responding
    tweet = (username, handle, postdate, text,)
    return tweet

# ### 3. Instantiate the firefox browser., entering the user details and the search value.
# application variables
user = input('username: ')
my_password = getpass('Password: ')
search_term = input('search term: ')

# create instance of web driver
# firefox
driver = webdriver.Firefox(executable_path="./geckodriver")
driver.get("https://twitter.com/login")
```

```
# navigate to login screen
driver.get('https://www.twitter.com/login')
driver.maximize_window()

username = driver.find_element_by_xpath('//input[@name="session[username_or_email]"]')
username.send_keys(user)

password = driver.find_element_by_xpath('//input[@name="session[password]"]')
password.send_keys(my_password)
password.send_keys(Keys.RETURN)
sleep(10)

# find search input and search for term
search_input = driver.find_element_by_xpath('//input[@aria-label="Search query"]')
search_input.send_keys(search_term)
search_input.send_keys(Keys.RETURN)
sleep(10)

# navigate to historical 'latest' tab
driver.find_element_by_link_text('Latest').click()

# ### 4. Continuously scrolling the page.
# get all tweets on the page
data = []
tweet_ids = set()
last_position = driver.execute_script("return window.pageYOffset;")
scrolling = True
while scrolling:
    page_cards = driver.find_elements_by_xpath('//div[@data-testid="tweet"]')
    for card in page_cards[-15:]:
        tweet = get_tweet_data(card)
        if tweet:
            tweet_id = ''.join(tweet)
            if tweet_id not in tweet_ids:
                tweet_ids.add(tweet_id)
                data.append(tweet)

    scroll_attempt = 0
    while True:
        # check scroll position
        driver.execute_script('window.scrollTo(0, document.body.scrollHeight);')
        sleep(2)
        curr_position = driver.execute_script("return window.pageYOffset;")
        if last_position == curr_position:
            scroll_attempt += 1

            # end of scroll region

            if scroll_attempt >= 3:

                scrolling = False

                break

            else:

                sleep(2) # attempt another scroll

        else:

            last_position = curr_position

            break


# close the web driver

driver.close()


# ### 5. Saving the data in a csv file

with open("ouagadougou.csv", 'w', newline=", encoding='utf-8') as f:

    header = ['UserName', 'Handle', 'Timestamp', 'Text']

    writer = csv.writer(f)

    writer.writerow(header)

    writer.writerows(data)
```

[redacted]ANNEX 2: OPEN DATA KIT (ODK) QUESTIONNAIRE

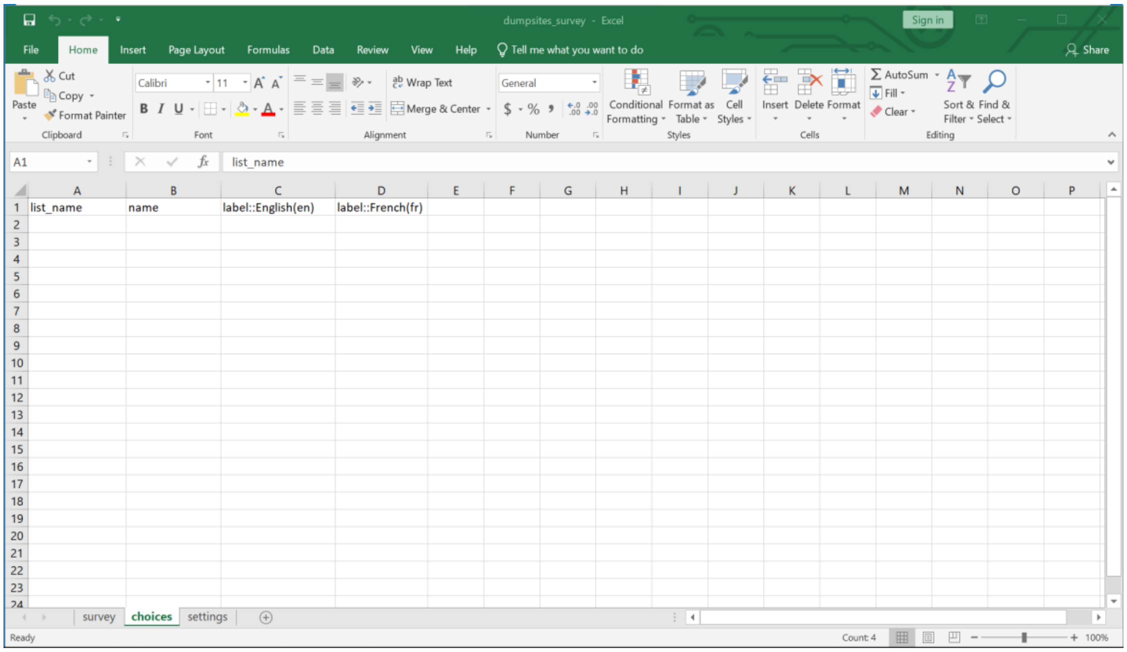
ODK Environment

- **Build:** create a data collection form or survey using the XLSForm.This form has three compulsory sheets.
    - **Survey sheet**


Worksheet in C Users MRP INFORMCC B2 06 Desktop thesis thesis_final_report_v1.docx - Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Tell me what you want to do
survey
choices
settings
Ready

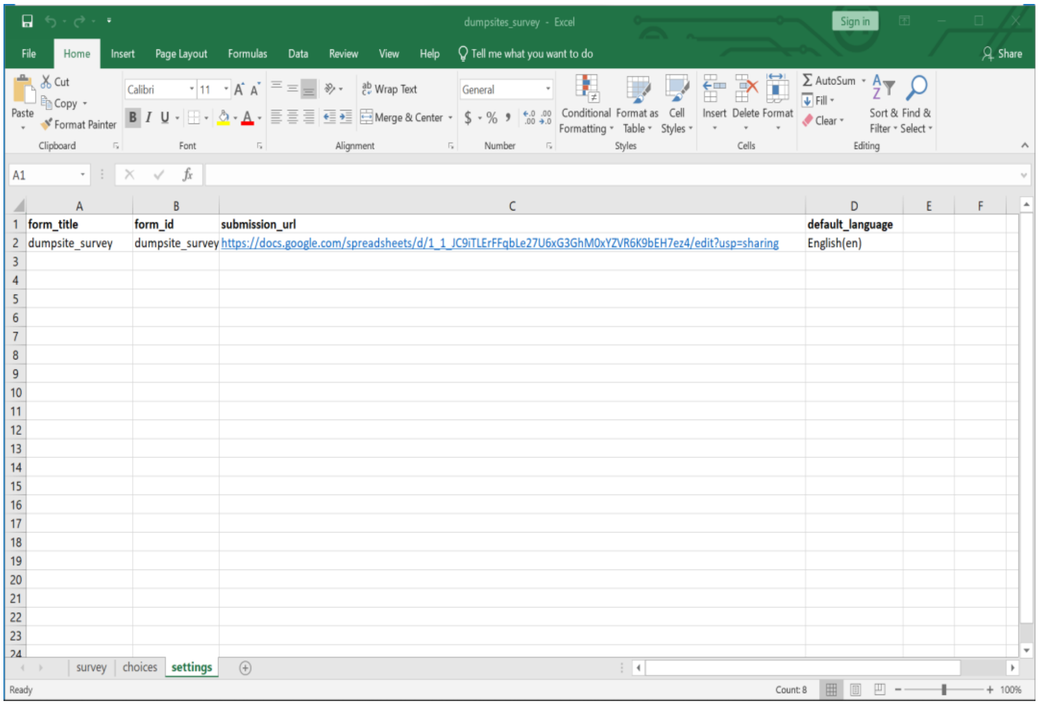
	A	B	C	D	E	F
1	type	name	label::English(en)	label::French(fr)	required	parameters
2	start	start				
3	end	end				
4	deviceid	deviceid				
5	start-geopoint	start-geopoint				
6	text	name_dumpsite	Write the name of the location	écrire le nom du lieu	yes	
7	text	area	The area of the dumpsite is:		yes	
8	image	photo	Take a photo of the place	Prendre une photo de l'endroit	yes	max-pixels=1024

- **Choices sheet**



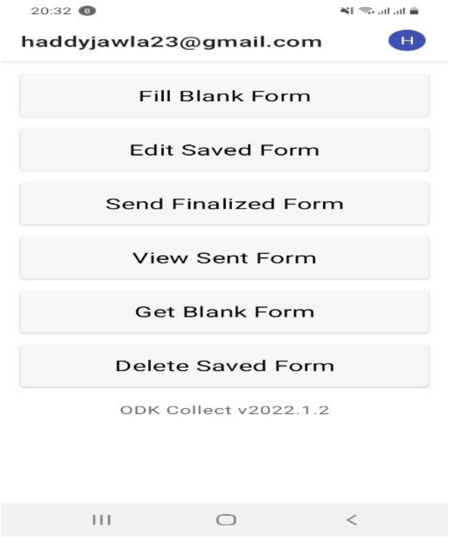
	A	B	C	D
1	list_name	name	label::English(en)	label::French(fr)

- **Settings sheet**

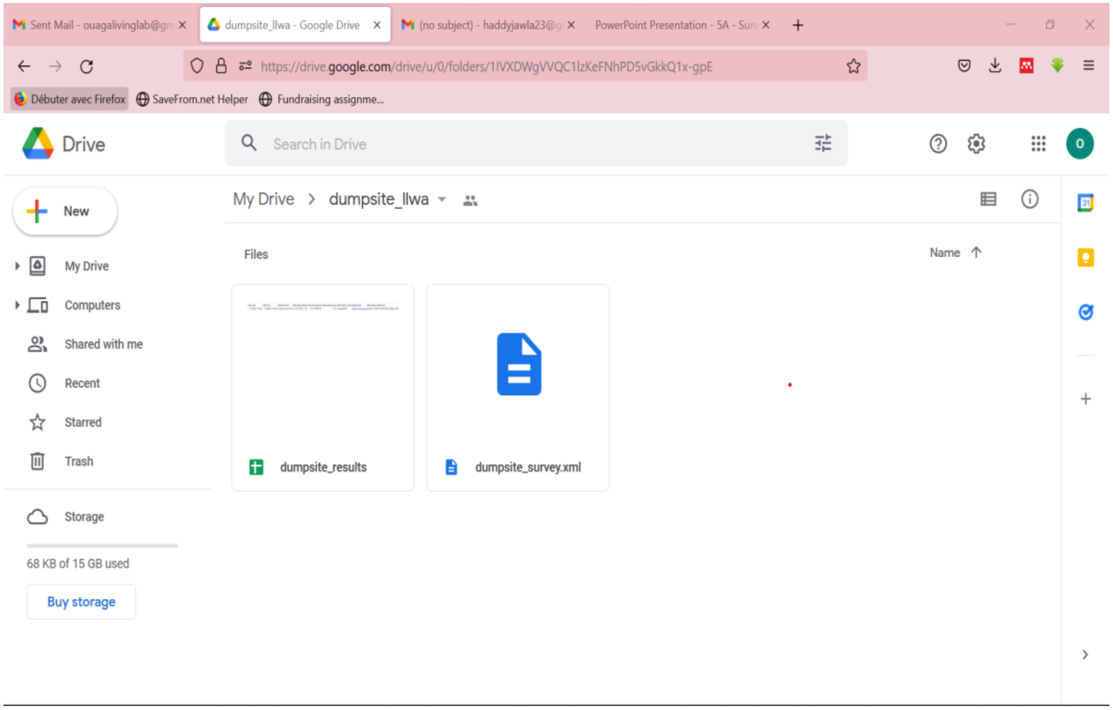


	A	B	C	D
1	form_title	form_id	submission_url	default_language
2	dumpsite_survey	dumpsite_survey	https://docs.google.com/spreadsheets/d/1_1_JC9iTLErFFqbLe27U6xG3GhM0xYZVR6K9bEH7ez4/edit?usp=sharing	English(en)

- **Collect:** compile data on a mobile device and send it to a server.
  - Download “ODK Collect” from Play Store



- **Aggregate:** put together collected data on a server (Google Drive) and extract it into useful formats.



https://drive.google.com/drive/u/0/folders/1IVXDWgVVQC1IzKeFNhPD5vGkkQ1x-gpE