**1. Project Name**

Crisis Informatics with Twitter Data

**2. Overview**

This project aims to understand the significance of contextual analysis, crisis prediction and recovery through textual data being created on the Twitter platform. Many new agencies and disaster relief organizations, including the UNDP, are interested in programmatically monitoring Twitter platform. However, there are many barriers to derive valuable insights using Twitter data as a reliable source of information. Utilizing Twitter data for crisis informatics application starts from determining the right and relevant tweet for a particular crisis. The key objective for this project is to utilize Twitter platform data for disaster preparedness, response, and recovery and build a proof-of-concept website that enables to analyze the information online. An example is a twitter alert system which can track rapid-onset disasters in a dashboard format.

**3. Background**

Twitter provides vast amounts of user-generated data, allowing a user to gain partial insights into the online public behavior. Twitter has become an important communication channel in times of emergency. Several factors that give Twitter considerable advantages over other social media platforms for analysis are, among others, the limited character size of tweets that provides a relatively homogeneous corpus. Moreover, the millions of tweets published every day allows access to large data samples, offering real-life information on many topics. Additionally, tweets are publicly available and easily accessible as well as retrievable via APIs with a developer account.

UNDP is an international organization operating and working on a broad spectrum of thematic areas. The organization can benefit greatly by analyzing real-life information coming from individuals who are the potential benefactor of the organization’s policy and programs. Multiple units within the UNDP have used social media data for tackling different issues related to development, such as food monitoring, public health, poverty reduction, urban dynamics and hate-speech detection.

Please, review the links below to have a better conceptual understanding of how Twitter data can be used in the field of development:

* [This study](https://www.unglobalpulse.org/project/analysing-social-media-conversations-to-understand-public-perceptions-of-sanitation-2014/) from UN Global Pulse analyses tweets to provide insights on the baseline of public engagement in the health sectorand explore ways to monitor a new sanitation campaign.
* [This study](https://www.unglobalpulse.org/document/using-twitter-data-to-analyse-public-sentiment-on-fuel-subsidy-policy-reform-in-el-salvador/) demonstrates that public opinion as expressed in social media can complement and potentially replace household survey data if such data is not available.
* [This project](https://marketplace.officialstatistics.org/now-casting-food-prices-in-indonesia-using-social-media-signals) investigates how people' self-reporting of commodity prices through Twitter can be used to provide real-time price indicators.
* [This project](https://www.unglobalpulse.org/project/nowcasting-food-prices-in-indonesia-using-social-media-signals-2014/) explored how Twitter data can be used to ‘nowcast’ or provide real-time food prices.
* [This study](https://www.unglobalpulse.org/project/feasibility-study-supporting-forest-and-peat-fire-management-using-social-media/) sought to explore early signals from Twitter relating to major forest fires or haze events to understand the relationship between communications trends and on-the-ground events.
* [This project](https://www.unglobalpulse.org/project/mining-citizen-feedback-data-for-enhanced-local-government-decision-making/) aims to explore the contribution of advanced data analytics to local government decision-making by generating insights from a combination of existing complaint systems and passive feedback from citizens on social media.
* [This study](https://www.unglobalpulse.org/project/using-twitter-to-understand-the-post-2015-global-conversation/) strives to provide real-time information on the development issues that most concern people around the world. By filtering Twitter every day for comments relevant to sixteen key development topics, the monitor shows which topics are most talked about in different countries over time.

**4. Key Objectives & Research Questions**

* To build a machine-learning model that predicts which tweets are about real disasters and which one’s aren’t using various NLP techniques and engage in a performance comparison.
* Develop a flagship website application (e.g. Dash of Plotly.js) to analyze Twitter data. The website application which may in the format of geographic dashboard.
* Review and consolidation of the resources for research on Crisis Informatics Topics from the [Crisis NLP Platform](https://crisisnlp.qcri.org/) (Qatar Computing Research Institute) and use the openly available data for the machine-learning model.
* Establish a solid Twitter analysis workflow that includes data gathering, data processing, analysis and evaluation if applicable.

Throughout the work, the research questions that need to be answered are the following:

RQ1: What components are required for twitter analysis workflow to derive contextual understanding for potential risks, crisis prediction, and recovery?

RQ2: How to quantify and identify misinformation in tweets?

RQ3: What are the best methods of practice to visualize time-series data such as twitter sentiments over time?

RQ4: How can user-generated contextual data such as tweet be used to provide insights on public opinion and further implemented into a decision-making process?

Please note that these suggestions are first and foremost guidelines. Consider the objectives as the focus, but may choose other methodologies and models that deems more suitable or accurate for the tasks.

**5. Workflows Ideas and Suggestions**

* Use a Twitter API to collect tweets related to a natural and human-made disaster, including potential risks such as climate change, violent conflict, and social unrest.
* After fetching the tweets, they need to be stored in a proper format such as JSON, CSV, TSV. This data may also be annotated for sentiments if we want to train our classifier.
* Performing exploratory data analysis that includes but not limited to finding an average number of words per tweet, N-gram frequency plots, word cloud and finding common words across datasets.
* Using an existing classifier to predict the sentiment for each tweet or train your classifier. Note that training will require to annotate the data for their sentiments. This data can be used as a training set to generate a machine-learning model (e.g. Logistic regression, LSTM, Naïve Bayes, SVM) and deep learning (e.g. BERT) techniques.
* Performing temporal data analysis that includes but not limited to finding the trends over the past 6 months to 1 year, and how the N-gram frequency plots and sentiments varied. Making predictions for future sentiments and risks.
* Creating visualizations (e.g. d3.js) that can present the results in a meaningful way and to be deployed on a web.

**6. Dataset and File Descriptions**

Dataset created using the official Twitter API would contain the following fields:

|  |  |
| --- | --- |
| Field | Description |
| created\_at | Date/time stamp of tweet creation. |
| id\_str | Unique tweet id. |
| screen\_name | Twitter username or handle. |
| display\_text\_range | Start and end word/character indices of the tweet. |
| text | Full text of the tweet. |
| location | Location information of the user at the time of creation. |
| hashtags | Contains the hashtags within the tweet. |
| indices | Start and end word/character indices of the hashtags |

| Index | Description[[1]](#footnote-2) |
| --- | --- |
| [Resource #1](https://crisisnlp.qcri.org/lrec2016/lrec2016.html) | This resource consists of Twitter data collected during 19 natural and human-induced disasters. Each dataset contains tweet-ids and human-labelled tweets of the event. Moreover, it contains a dictionary of out-of-vocabulary(OOV) words, a word2vec model, and a tweets downloader tool. |
| [Resource #2](https://crisisnlp.qcri.org/data/swdm2013_practical_dataset/SWDM2013_dataset.zip) | This resource consists of human-labelled tweets collected during the 2012 Hurricane Sandy and the 2011 Joplin tornado. |
| [Resource #3](https://crisisnlp.qcri.org/data/iscram2013_nuggets_datasset/ISCRAM2013_dataset.zip) | This resource consists of human-labelled tweets collected during the 2011 Joplin tornado and labelled into humanitarian categories. |
| [Resource #4](https://github.com/CrisisNLP/deep-learning-for-big-crisis-data) | This resource provides ready-to-use Python implementation of several neural networks and non-neural network-based classifiers for the classification of crisis-related Twitter data. |
| [Resource #5](https://crisisnlp.qcri.org/crisismmd) | This resource provides human-labelled multimodal datasets comprised of tweets and images collected during seven major natural disasters. |
| [Resource #6](https://crisisnlp.qcri.org/data/iscram2018_twitter_tale/ISCRAM18_datasets.zip) | This resource comprised of tweet-ids and a sample of raw tweets (50k) collected during three devastating hurricanes in 2017 namely Hurricane Harvey, Hurricane Irma, and Hurricane Maria. |
| [Resource #7](https://crisisnlp.qcri.org/data/acl_icwsm_2018/ACL_ICWSM_2018_datasets.zip) | This resource comprised of human-labelled tweets collected from the 2015 Nepal earthquake and the 2013 Queensland floods. |
| [Resource #8](https://crisisnlp.qcri.org/data/tools/TweetsRetrievalTool-v2.0.zip) | This resource is a Java-based tool to download full tweets content using tweet ids. This tool can make 180 API calls per 15 minutes, each API call downloads up to 100 tweets i.e. it can download up to 72,000 tweets per hour. |
| [Resource #9](https://crisisnlp.qcri.org/data/ASONAM17_damage_images/ASONAM17_Damage_Image_Dataset.tar.gz) | This corpus comprises images collected from Twitter during four natural disasters, namely Typhoon Ruby (2014), Nepal Earthquake (2015), Ecuador Earthquake (2016), and Hurricane Matthew (2016). In addition to Twitter images, it contains images collected from Google using queries such as "damage building", "damage bridge", and "damage road" to deal with labelled data scarcity problem. |
| [Resource #10](https://crisisnlp.qcri.org/data/acl_icwsm_2018/ACL_ICWSM_2018_datasets.zip) | This resource comprised of human-labelledd tweets collected from the 2015 Nepal earthquake and the 2013 Queensland floods. |
| [Resource #11](https://crisisnlp.qcri.org/data/california_wildfire_2018_tweet_ids_4m.csv) | This resource comprised of data related to the [2018 California wildfires (a.k.a Camp Fire)](https://en.wikipedia.org/wiki/Camp_Fire_(2018)). Specifically, it contains (1) the names of missing and found people, (2) web sources from which the names were taken, (3) hashtags related to missing, lost and found people. We will publish tweet-ids soon. |
| [Resource #12](https://crisisnlp.qcri.org/data/eyewitness_tweets_annotations_14k_public.zip) | This resource comprised of ~14,000labelledd tweets collected during several natural disasters including hurricanes, earthquakes, floods, and forest fires. The tweets were annotated following aeyewitnessss taxonomy defined in the below paper. |
| [Resource #13](https://crisisnlp.qcri.org/heritage) | This resource contains images and expert annotations for the detection of damaged heritage sites. The images were downloaded from Google and annotated using two annotation schemes. First, if an image shows a heritage site or not. Second, if a heritage image shows some damage content or not. |
| [Resource #14](https://crisisnlp.qcri.org/covid19) | This resource contains more than 50millionon tweets related to the COVID-19 pandemic. The geographic coverage of the dataset spans over 218 countries and 47K cities around the globe. Moreover, the dataset covers 62 international languages. |
| [Resource #15](https://crisisnlp.qcri.org/crisis-image-datasets-asonam20) | The crisis image benchmark dataset consis ofts data from several data sources such as Disasters on Social Media (DSM), CrisisMMD and data from AIDR. The purpose of this work wto as develop a consolidated dataset, and create non-overlapping train/dev/test set and pride a benchmark results for the community. |

**7. Deliverables**

* Advanced data visualizations of the Twitter derived insights generated in a web deployed application format. (e.g. Online dashboard with Flask and Pusher)
* Trained machine-learning models to predict which tweets are about real disasters and which ones are now. Model files with the optimized hyperparameters and compared accuracy between models.

1. Information from the Crisis NLP, Qatar Computing Research Institute: <https://crisisnlp.qcri.org/> [↑](#footnote-ref-2)