Title page

Extracting and Exploring Information about Flood Events from Twitter

Yaser Kaddoura

This page intentionally left blank.

Abstract

The more information about a disaster gets established, the more efficiently the disaster management is done by the concerned parties to handle the situation. People tend to share their experiences during disastrous events using social media, making them potential data sources. This thesis project implements a pipeline to extract knowledge from Twitter about flood events. It determines flood-relevant tweets using a classifier and identifies geographical locations mentioned in the tweets using a hybrid geoparsing approach. At the end of the pipeline, the spatial, temporal, and textual aspects of the results are presented using an interactive visual interface. The implemented pipeline is exemplified using historical tweets created during past flood events.

This page intentionally left blank.

Contents

List of Figures

List of Tables

List of Acronyms

Chapter 1

Introduction

Earlier this century, floods around Lake Vänern and Arvika have costed Sweden an estimated 11.1 billion Swedish Krona for damages and repairs [38]. Counties of Dalarna and Gävleborg suffered from flash floods in 2021, disturbing the daily life of their citizens and damaging public and private properties [5]. Flooding is a devastating natural disaster that threatens the livelihood of people and the infrastructure of communities around the world [13].

To facilitate the process of emergency management during these hazardous events, early warning systems analyse their risk, monitor and warn the public while ensuring their readiness [4]. Traditionally, meteorologists forecast the weather by relying on tools such as gauges, satellites, and radars for data extraction. The emergence of social media platforms such as Twitter provide individuals with a public space to share their experience, effectively creating another potential data source.

Researchers started harnessing this new wealth of information to aid the disaster management procedure. Twitter streaming Application Programming Interface (API) makes it possible to create a monitoring system for early event detection on a global [7] and local [3] scales. Another use for it would be identifying victims in real-time, locating their physical location, and communicating the information to rescue teams [43]. After the threat subsides, emergency managers can use relevant tweets to assess the impact and plan the recovery phase [3]. To prepare for future floods, authoritative entities can make informed actions by analysing historical data and determining the locations suffering from recurrent calamities. This newly acquired knowledge can augment weather warning systems' pipelines improving their accuracies such as SMHI, a Swedish expert authority with a global perspective and a vital task in predicting changes in weather, water, and climate [44].

Most research addresses the problem on a global or national scale, yet none has addressed Sweden specifically. This thesis project covers this gap by using Swedish tweets to extract relevant knowledge about flood events in Sweden with the focus on answering the following questions:

1. What methods can be used to classify Swedish tweets related to flood events?
2. How to extract the locations mentioned in the tweets?

3. What insights can be extracted by using text analysis on the tweets?

4. What visualizations can be used to represent the results?

The project focuses on providing a proof of concept for addressing the questions above, and it will not design a solution to make a production-ready product; i.e., the engineering challenges, such as automation, scalability, and ease of deployment are out of the focus of this thesis.

This thesis project is part of a research project, AI for Climate Adaptation [28] at Linköping University (LiU) in collaboration with the SMHI. It implements a pipeline that provides a visual representation of tweets related to flood events in Sweden. First, relevant tweets are pulled, processed, and classified from the Twitter API using data mining techniques. Second, geographical locations are identified from tweets mentioning flood events employing Named-entity recognition (NER) and gazetteers. Third, various text analysis techniques are applied to tweets. Finally, the results are presented using a visual interface. For verification purposes, the pipeline is applied on a week's worth of tweets after past flood events.

Chapter 2

Literature Review

The massive and accessible volume of data produced by social media has attracted the attention of many researchers as a valuable data source for their research topic; however, collecting and processing data of this nature pose many challenges to extracting useful information. This section mentions what other researchers focusing on disaster management topics did to address these challenges while using Twitter; it also discusses the different approaches used for identifying relevant tweets, extracting geographical location from them, making text analysis on the text, and visualizing the results.

2.1 Data Collection

Twitter's API enables developers to retrieve historical tweets using queries that are made of operators to match a variety of tweet attributes, such as a specific keyword, having a geotag provided by the user who created the tweet, and the language classified by Twitter. Users generate around 500-700 million tweets a day [23], making it necessary to limit the number of tweets to fetch using the API to reduce computational power and downtime. Feng and Sester [12] fetch geotagged tweets and filter them using 45 keywords in 7 languages; this approach disregards a massive portion of relevant tweets since 1% of tweets are geotagged [25]. A better way that doesn't only regard 1% of the tweets is to fetch tweets using keywords related to the topic of interest in different languages. de Bruijn et al. [7] use over 40 keywords associated with floods in 11 major languages in the query to fetch tweets.

In addition to textual data, some researchers use other data types to enhance their pipelines. Some tweets contain media attachments, such as images and videos that potentially provide additional visual information for the concerned research topic [1][39][29]; search engines are another resourceful source for images as well [12]. For flooding events, hydrological information can be a valuable source of information, which can be extracted from a global precipitation dataset based on tweets' time stamps and location in the text [8]. Barker and Macleod [3] use Environment Agency flood-monitoring API[1] to get river gauge levels and flood warnings to identify at-risk flooding areas.

[1] https://environment.data.gov.uk/flood-monitoring/doc/reference#flood-warnings

Text processing is crucial for any Natural Language Processing (NLP) pipeline to train performant classifiers. Some research deals with multilingual corpus, making it necessary to translate the text to one language (most likely English) if the used classifier can not handle multilingual data [43]. One of the most common text-processing tasks is removing unnecessary terms such as stopwords, Uniform Resource Locator (URL)s, numbers, and punctuation marks. "User mentions" (e.g. @user123) in tweets don't provide helpful information, so pipelines often remove or replace them with a generic term, such as "@user" [8]. The location of the flood event is a crucial piece of information that can be pinpointed from a term in the tweet, making it a potential target that includes bias in the dataset by overusing it; de Bruijn et al. [8] replace these terms by the country name that the location resides in; on the other hand, Petersen and Styve [35] replace the terms by the word "place" if they get mentioned more than 0.5% of the size of the data set. Another way to improve the performance of the classifier is to group the terms by converting them to lower-case and transforming them to their lexeme (e.g. better turns to good) or word stem (e.g. walking turns to walk) by lemmatisation [35] or stemming [12], respectively.

Some tweets are redundant or do not provide relevant information; these can be considered noise and filtered out. For example, retweets are identical to other tweets without additional context making them unneeded. Spam bots generate similar tweets for malicious reasons, such as spreading false content to manipulate the public, and utility reasons, such as creating a feed for users to check updates. These tweets introduce noise to the dataset that gets reduced by removing duplicate tweets. de Bruijn et al. [7] only consider one tweet from each user in the last 14 days mentioning a specific region; they also discard tweets containing more than five consecutive words that match with those in another tweet among the previous 100 talking about a location. Singh et al. [43] approach this problem by only extracting tweets created from mobile phones and only consider tweets from users who have followers/following < 1.

2.2 Text Classification

Identifying disaster events using social media requires a classifier to determine the relevant data. Textual data containing terms related to a disaster doesn't mean it discusses a disastrous event since words such as "flood" can be used figuratively in sentences (e.g., a flood of joy). A binary classifier labelling the data with "on-topic" and "off-topic" labels is needed to filter out irrelevant content.

Most classifiers use supervised Machine learning (ML) algorithms requiring labelled data for training. A straightforward approach is to manually label a sample of the tweets [7][3]. Petersen and Styve [35] use CrisisLex6 [30], a crowdsourced collection of labelled tweets, for training their classifiers that get evaluated on 88 million unlabelled tweets containing flood-related terms [6]. Feng and Sester [12] automatically label the tweets by checking if there is rainfall during the provided time and city location by using a weather API[2]; if there is rainfall, the tweet is labelled positive, otherwise negative.

[2]https://www.wunderground.com/weather/api/d/docs

A classifier needs a numerical representation of the textual data for training. Text is often represented in a real-valued vector by encoding words and their context. There are different word embedding techniques, such as TF-IDF, which reflect on how important a word is to a document in a corpus, Word2vec [27], and its extension doc2Vec [21], which capture the semantic and syntactic qualities of words via a vector space with several hundred dimensions, where each unique word in the corpus gets assigned to a vector.

There are three groups of approaches for NLP tasks: heuristics, ML, and deep learning. The heuristics approach is the oldest; it manually builds rules for a specific task by using dictionaries and thesauruses. ML techniques, including probabilistic modelling and likelihood maximization, are used on a numerical representation of the textual data to learn a model. Neural networks are a popular choice for handling complex and unstructured data, making them a suitable candidate for text.

There are different groups of ML algorithms to classify data for varying data types. Supervised algorithms are employed if the training data set is labelled; otherwise, a probabilistic approach can be used by training a naive Bayes classifier on labelled and unlabelled data [22]. Feng and Sester [12] use naive Bayes, random forest, logistic regression, Support Machine Vector (SVM) (RBF Kernel), and SVM (Linear Kernel) on labelled data transformed using TF-IDF with accuracies of 0.7109, 0.7582, 0.7705, 0.7712, and 0.7739, respectively. Petersen and Styve [35] results are more promising, where they train a logistic regression and random forest classifiers with 0.939 and 0.9253 accuracies, respectively. Deep learning approaches generally outperform classical algorithms; one example is Convolutional Neural Networks (CNN) trained on word embeddings for sentence classification. Feng and Sester [12] and Petersen and Styve [35] train a CNN model on word2vec embeddings with 0.7868 and 0.94611 accuracies, respectively.

RNN [17] is a common artificial neural network for NLP tasks, such as text classification, NER, and machine translation. Its memory enables it to take information from previous input to update the current input and output vector (called hidden state) as shown in Figure 2.1, taken from Tunstall et al.'s book [47], making it appropriate for sequential data, such as text.

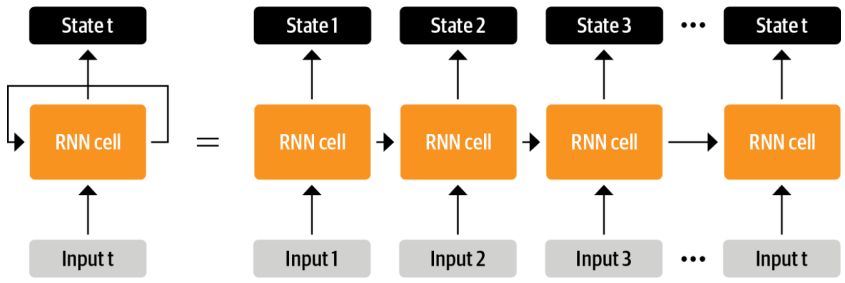


Figure 2.1: RNN example [47]

For tasks such as translation, an encoder-decoder architecture is needed; the encoder encodes the input sequence into a numerical representation (called the last hidden state)

that gets passed to the decoder for output sequence generation. Figure 2.2 shows an example of translating the English statement "Transformers are great!" to German.

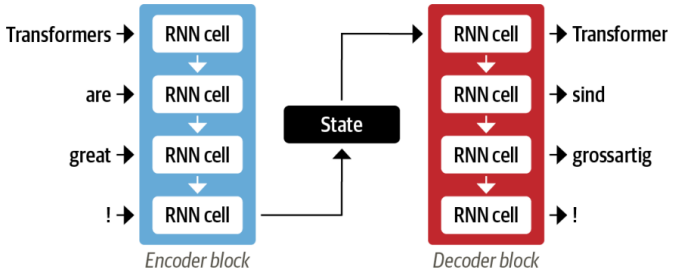


Figure 2.2: Two RNNs making an encoder-decoder architecture [47]

RNN has shortcomings when it tries to capture the context of long sequences of information since the encoder might lose the information at the start while forming the representation. RNN's weak memory can be addressed by using the attention mechanism that allows the decoder to access all the hidden states of the encoder. The main goal of attention is to enable the decoder to prioritize the states using weights it assigns at every decoding timestamp. Figure 2.3 shows an example for predicting the third token in the output sequence. Even though attention improves the accuracy of the translations, the computations are sequential and cannot be parallelized. In addition, most NLP tasks require training models using a large amount of labelled text data that might not be available. Transfer learning resolves this problem by transferring knowledge acquired from solving one problem to other related ones.

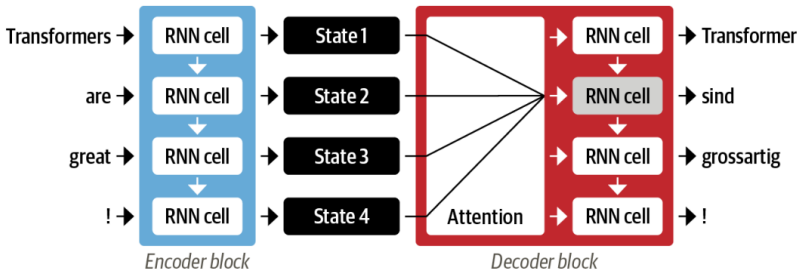


Figure 2.3: Two RNNs making an encoder-decoder architecture with attention mechanism [47]

Transfer learning has been used in computer vision before its introduction to NLP. The models are pre-trained on massive datasets, such as Imagenet [20] and places database [50] to learn the basic features of images, such as edges and colours; then, they are fine-tuned on downstream tasks with a smaller dataset. Feng and Sester [12] use GoogLeNet (Inception-V3 model) [45] pre-trained on ImageNet to train multilayer perceptron, random forest, gradient boosted trees, and XGBoost with accuracies of 0.8907, 0.9133,

0.9252, and 0.9295, respectively. Ning et al. [29] uses VGGNet [42], Inception V3, ResNet [15], and DenseNet201 [19] with 0.91 accuracy.

In 2017 and 2018, several research groups proposed new approaches to use transfer learning for NLP. Universal Language Model Fine-tuning (ULMFit) [18] introduced a general framework by pre-training Long Short-Term Memory (LSTM) models for various tasks. Petersen and Styve [35] fine-tune a pre-trained ULMFit model to classify flood-relevant tweets with an accuracy of 0.9499.

Transformers with transfer learning and their self-attention architecture, proposed by google researchers [49], made the training process much faster. The idea is to use attention on all states in the same neural network's layer. Figure 2.4 shows the self-attention mechanism on both the encoder and decoder with their outputs fed to feed-forward neural networks. Alam et al. [1] fine-tune a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [10] model that works on one language with an accuracy of 0.853, and de Bruijn et al. [7] use a multilingual model with 0.8 F1-score.

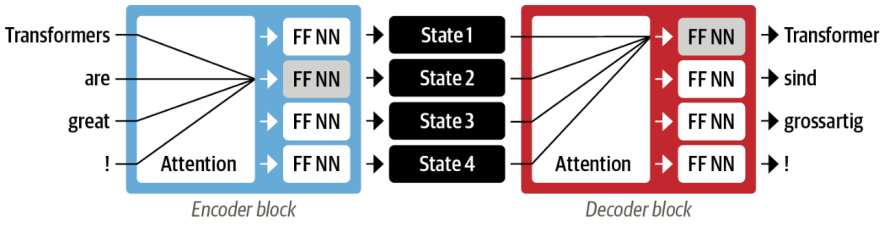


Figure 2.4: Transformer's encoder-decoder architecture [47]

2.3 Location Extraction

Identifying the locations of disasters is helpful for the disaster management cycle. Social media enables people to generate Volunteered Geographic Information (VGI), which is more advantageous over the more expensive accuracy testing done by official agencies because contributors have unique local knowledge. Detecting a disastrous event and its location as soon as possible can reduce its impact on society [7] by informing the citizens and the authority to prepare for it. During the event, the rescue teams' task becomes easier if they can locate the endangered people [43]. When the event wanes, assessing the most impacted spots can enable the authority to make informed decisions on a recovery plan.

Twitter users can assign an accessible property to their tweets, called "geotag", a geographical identification metadata. Adding "has:geo" to the query sent to the API will return geotagged tweets with metadata about the location, such as a display name, geo polygon, and a geo lat-log coordinate. The geotag is the most straightforward method to identify the locations [12], but unfortunately, only 1% of the tweets are geotagged [26].

Locations can be extracted using toponyms, a place's name, in tweets' text by using geoparsing, which is a process of converting free-text descriptions of places (such as

"twenty miles northeast of Jalalabad") into unambiguous geographic identifiers. A toponym can have more than one location candidate, such as "Boston", which is the name of several places, including "Boston, USA" and "Boston, UK"; this fact makes geoparsing tasks on a global scale harder than local ones. de Bruijn et al. [7] use TAGGS [9], a geoparsing algorithm, to extract countries, administrative areas, and settlements (i.e. cities, towns, and villages) mentioned within the tweets' text on a global scale. The process includes toponym recognition and toponym resolution. Toponym recognition extracts the toponyms that refer to one or more locations using a gazetteer, a geographical index, or a dictionary. Toponym resolution predicts the correct location for the toponyms in several steps. A score is assigned to each possible location using metadata related to the tweet, such as the user's timezone and hometown, the tweet's coordinates, and mentions of nearby locations. Then, the average score of grouped tweets mentioning the same toponym within a 24-hour is calculated. Finally, the groups of tweets with the location that has the highest score are assigned. Petersen and Styve [35] use geotag property, geoparsing using NER on text, and user's profile location to extract toponyms. If the text contains two toponyms, they pick one randomly if the locations are close with a distance threshold of 1500km. They use GeoPy[3] to assign geographical locations to toponyms, a Python package that is a client for several popular geocoding web services (e.g., GoogleV3 and GeoNames). Singh et al. [43] use the fact that people visit the same locations daily to generate a Markov chain model on historical tweets created by the same user to locate them.

2.4 Text Analysis

Besides text classification and location extraction, other text analysis techniques extract valuable information from text data. During hazards, disaster managers can use social media to get insights, such as how impactful an event is on society. They can visualize the results to understand the situation and act accordingly.

Gründer-Fahrer, Schlaf, and Wustmann [14] extract multiple relevant pieces of information from social media and present them to disaster managers via a searchable application. They extract the following: topics using HDP-CRF algorithm [46], locations using Openstreetmap[4] location markers, time using the social media metadata, and names of organizations using NER. They present the information using several interactive graphs such as pie charts, word clouds and line graphs.

Dimensionality reduction is a common preprocessing technique to reduce the complexity of textual data, preparing it for other tasks such as noise reduction, visualization, or clustering. Heusinger, Raab, and Schleif [16] use random projection to reduce the dimensions of tweets to predict their hashtags, making them easier to search on Twitter. Omuya, Okeyo, and Kimwele [31] extract features from social media using Principal Component Analysis to perform sentiment analysis. Sentiment analysis is a popular text analysis technique that shows people's sentiments during an event. Lu et al. [24] perform

[3]https://geopy.readthedocs.io/en/stable/
[4]https://www.openstreetmap.org/

sentiment analysis from Twitter about the Ebola virus using three different sentiment classifiers to measure the sentiment score of the tweet depending on the majority of the votes. Also, they calculate the inconsistency between the classifiers using an entropy measure [2]. The positive and negative sentiments are each presented in a density map using solid blue and red colours, respectively; if the inconsistency score is above a certain threshold, the colour is blurred. Periñán-Pascual [34] tries to extract the sentiment by calculating three scores for the tweets: (1) the reliability of how much the tweet discusses a problem during a hazard, (2) the impact of the tweet by using the user's activity and popularity as well as tweet's influence [33], (3) and the impact of the problem using the previous scores. They present the mean of the scores on a time frame basis on a line graph.

2.5 Visualization

Visualization of the results of an NLP pipeline is common practice for several reasons. The massive and complex data can communicate the needed knowledge for different audiences to understand the underlying situation and take action accordingly. The developers can use the visualization to validate that the pipeline is working as intended. The authority can check the spatio-temporal data to identify places that have recurring floods and reinforce their infrastructure to prepare for future flooding. Also, the plots make event detection and monitoring much faster and more straightforward.

de Bruijn et al. [7] use historical and real-time data to show flooding events on different levels (countries, administrative areas, and settlements). It is powered using a JavaScript library, leaflet[5]. The application, seen in figure 2.5, contains a map showing the flooding events with an adjustable timeline and a list of tweets for the selected location.

[5]https://leafletjs.com/

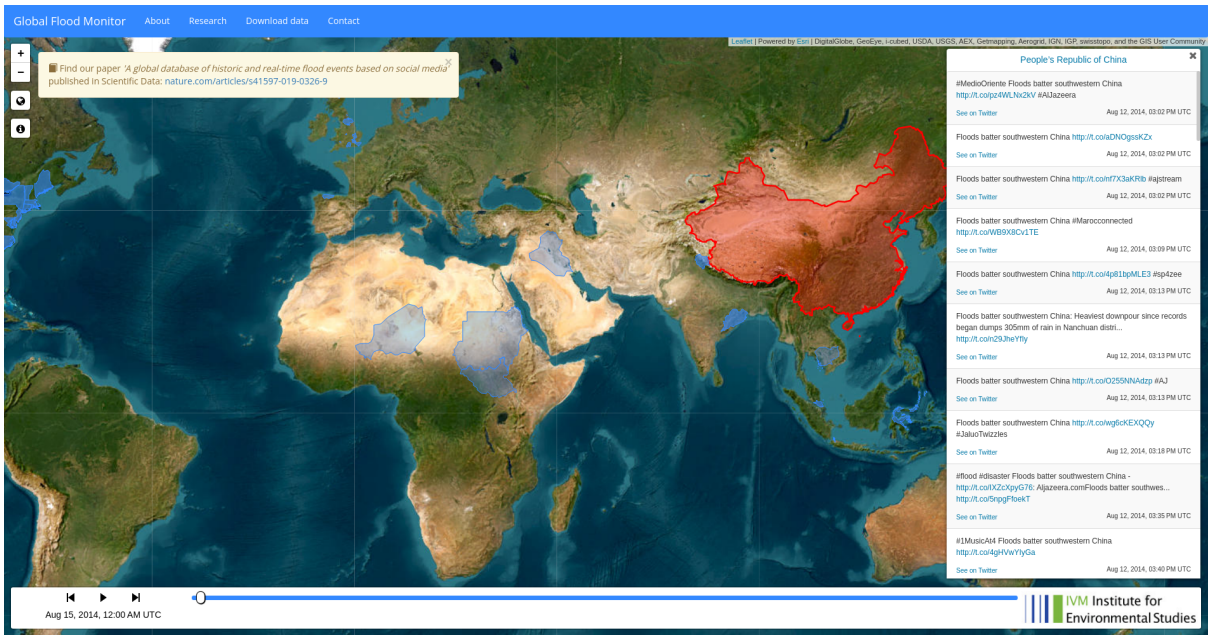


Figure 2.5: Global Flood Monitor application showing flood events

Petersen and Styve [35] provide multiple plots with sophisticated methods to configure the interface and filter the tweets. Their visualization is powered using the python libraries, Plotly[6] and Dash[7]. The app, shown in Figure 2.6, provides an interface to showcase the different aspects of the data: spatial via a map, temporal via a histogram, and textual via a list of tweets and word cloud. They use a scatter map to show the locations extracted from the tweets, where the colour of each point represents the method used to identify the location. To resolve the problem of tweets overlapping each other due to the discussion of the same location, the identical points are spread by adding Gaussian noise to their coordinates points. As for representing the timestamps, they use a histogram aggregated by each day with a time slider. Researchers can pinpoint repetitive or interesting topics by navigating the word cloud to see the most frequent keywords or manually navigating the list of tweets. The plots are interactive, where actions in one of them would influence others. The data can be filtered in different ways: keywords in the text, the method used to extract the location, tweet type (a retweet or not), a map, and a histogram. In addition, there is a drop-down to change the map graph type and the algorithm used to classify the tweets.

[6]https://plotly.com/python/
[7]https://dash.plotly.com/

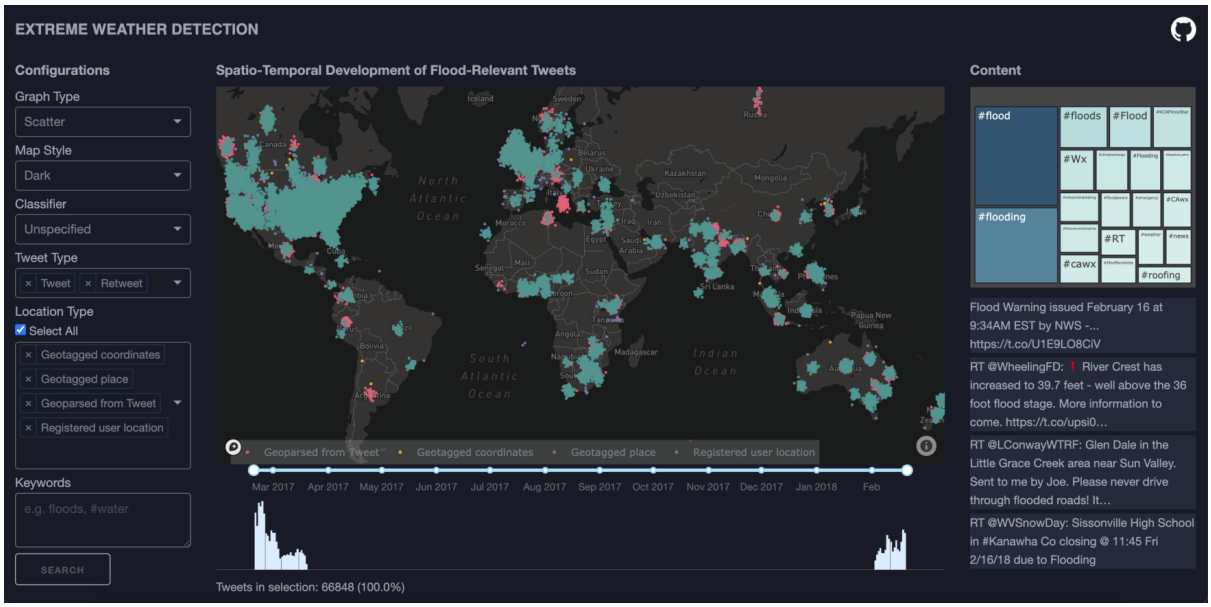


Figure 2.6: Petersen and Styve [35] application

Feng and Sester [12] use leaflet to plot a map showing flooding events as observed in Figure 2.7. They use Getis-Ord Gi* [32] to detect statistical hot spots and present them as a choropleth map. The light blue circles represent the spatio-temporal clusters of events, and the circles with numbers at the centre indicate clusters of tweets in that area with their total. The markers indicate individual tweets with a pop-up containing information about it.

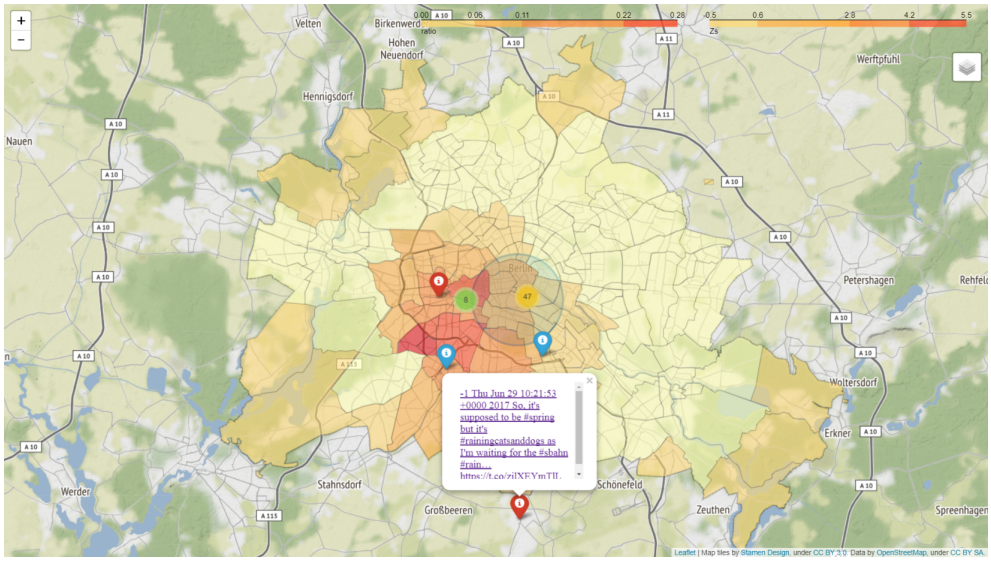


Figure 2.7: Web map application with pluvial flood in Berlin by Feng and Sester [12]

Barker and Macleod [3] visualize the tweets using different map plots created by leaflet. The map plot in Figure 2.8 consists of clickable pointers for pop-up boxes of the tweets.

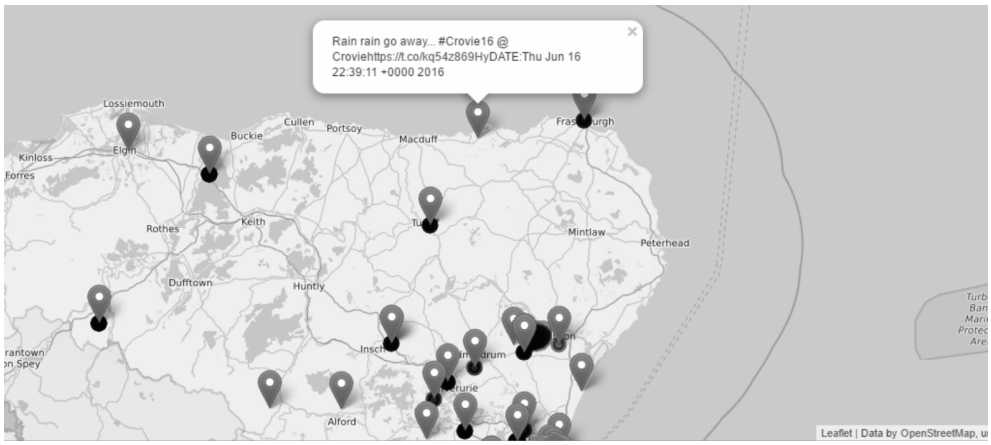


Figure 2.8: Map with tweet markers in by Barker and Macleod [3]

The bubble map in Figure 2.9 displays the tweets with the size of the circles representing the area of the place and colour indicating the number of tweets talking about the location.

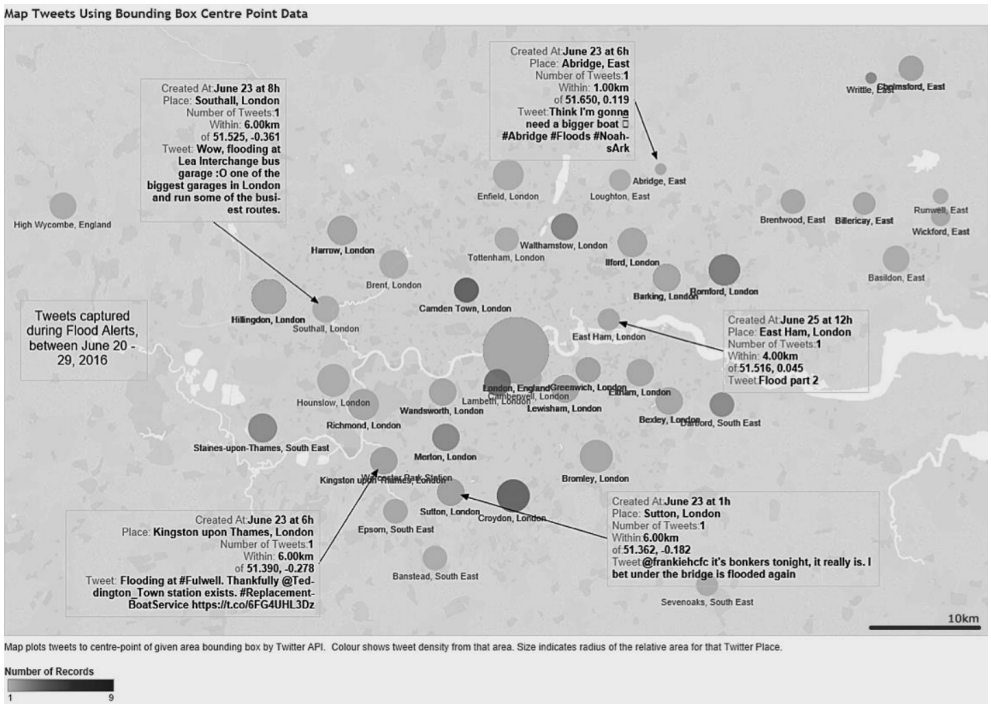


Figure 2.9: Bubble map of tweets by Barker and Macleod [3]

Chapter 3

Methods

This section discusses and motivates the methods used in the project. Figure 3.1 shows a flow chart of the steps for the pipeline (an enlarged and more detailed copy is available in Figure A.1 of Appendix A). The pipeline consists of the following steps: Data collection, text classification, location extraction, text analysis, and visualization. Python is the primary programming language used for the project because of the rich ecosystem surrounding it, especially when it comes to data science-related tasks. The code base is available on a GitHub repository[1] accompanied with a `README.md` containing instructions to set up the environment and run the project.

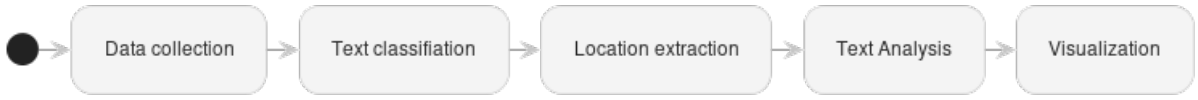


Figure 3.1: Flow chart for the pipeline

3.1 Data Collection

Finding a good quality data source is the first step to having a lean start for most research questions. The pipeline trains an ML classifier using three manually labelled datasets. Two of them are crowdsourced datasets provided by CrisisLex6; the tweets are from the 2013 flood events in Alberta[2] and Queensland[3], and there are around 10,000 records for each one with the tweet's ID, tweet's text, and a label about the relevance of the tweet regarding the event. The third dataset was collected and annotated in the scope of the AI4ClimateAdaptation project and includes flood events in Sweden, spanning between 2015 and 2021; it contains 4899 tweets, most of them are in the Swedish language, with attributes presented in Table 3.1. The text and metadata of the tweets are extracted from Twitter's API using the IDs. The trained model performance is verified using

[1] https://github.com/YasserKa/Classification-and-visualization-of-natural-disasters-using-Twitter
[2] https://en.wikipedia.org/wiki/2013_Alberta_floods
[3] https://en.wikipedia.org/wiki/Cyclone_Oswald

Table 3.1: Dataset attributes

Field	Type	Description
ID	Int	ID of the tweet
On Topic	Bool	Text discusses an event
Informative sarcastic	Bool	Text contains relevant information about the event
Contains IMPACT info	Bool	Text discusses the impact of the event
Explicit location	Bool	Text mentions the location of the event

extracted tweets from the API using Tweepy[4], a python library for accessing Twitter API. Figure 3.2 shows both the source and usage of the data in the pipeline.

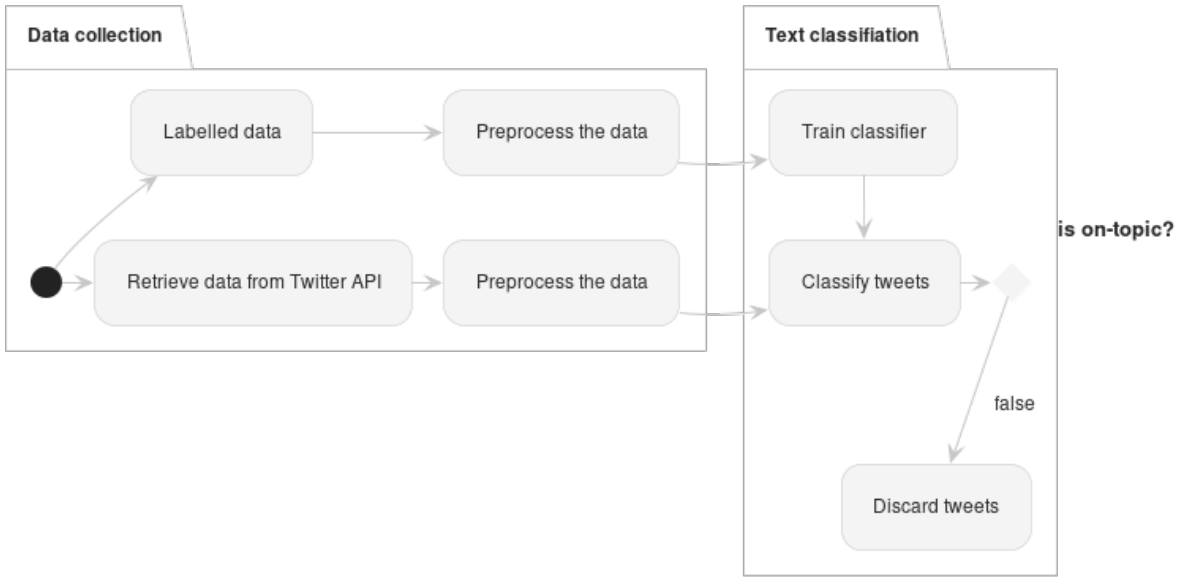


Figure 3.2: Data collection and text classification steps of the pipeline

After retrieving the data, they are pre-processed to prepare them for the upcoming tasks, such as training an ML algorithm, text analysis, and visualization. Parts of the text that do not contribute to the context are discarded: URLs, emojis, mentions, hashtag signs, numbers, new lines, punctuation, and stopwords (provided by spaCy[5], an NLP python library). Afterwards, duplicate tweets, tweets with no text, and retweets are discarded from the dataset. The trained model requires the text to be in English, and since Sweden is the focus of the research, most of the text is in Swedish; thus, the Swedish text is translated to English using google translate[6] by a python library wrapper deep-translate[7].

[4]https://docs.tweepy.org/en/latest/index.html
[5]https://spacy.io/
[6]https://translate.google.com/
[7]https://deep-translator.readthedocs.io/en/latest/

Table 3.2: Tweet attributes used

Attribute	Type	Description
id	Int	The unique identifier of the requested Tweet
text	Str	The actual UTF-8 text of the Tweet
created at	Date	Creation time of the Tweet
author id	Str	The unique identifier of the tweet creator

Data needs to be stored and managed to accommodate policies and regulations. Twitter's developer policy[8] has a content redistribution section stating that only tweets' IDs can be shared online. Thus, the tweets can not be available publicly on the internet, such as GitHub (the service that hosts the publicly available code base). To this end, the data is stored after each step of the pipeline using google drive using Data Version Control (DVC)'s[9] data management capabilities.

Twitter's API provides an extensive list of information about the tweets[10]. It shares the engagement metrics of the tweet, including like count, reply count, and retweet count; as well as, an NLP analysis of its own, such as the language used, and entities parsed from the text. Table 3.2 shows the tweet's attributes used in this project for the following reasons: the id to generate the URL of the tweet, the text for NLP tasks, the created date for temporal analysis, and author's id to reduce spam.

3.2 Text Classification

This project uses the DistilBERT transformer[40], a variant of BERT, for text classification. The main advantage of this model is that it achieves comparable performance to BERT while being significantly smaller than BERT and more efficient. The used DistilBERT pre-trained model is provided by Hugging Face[11], a framework that provides a unified API for over more than 50 architectures, making it easier for users to integrate NLP models into their applications. The learning rate for the neural network is $5 \times e^{-5}$ with 100 warmup steps over four epochs using 90% of the labelled tweets as training data, 5% as test data, and 5% for validation. Text classification's purpose is to identify the tweets that discuss flood events, so model training is done with the "On Topic" attribute as a label.

Training the model in the local environment takes a long time with the available resources, so the training is done using Amazon SageMaker[12], a service that covers tools to build, train, and deploy ML models, to accelerate the process. The data is uploaded to Amazon Simple Storage Service (Amazon S3) to make it accessible for the Hugging Face training script, which gets executed in an instance available in the cloud. After

[8] https://developer.twitter.com/en/developer-terms/policy
[9] https://dvc.org/doc
[10] https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet
[11] https://huggingface.co/
[12] https://aws.amazon.com/sagemaker/

Table 3.3: Confusion matrix

		Predicted: Positive	Predicted: Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

completing the training, the fine-tuned model and the evaluation metrics are downloaded. The evaluation metrics consist of the following:

- Confusion matrix: a matrix showing the classifier's predictions for a labelled dataset corresponding to its actual values (Table 3.3).
- Accuracy: a fraction of the number of correctly classified instances (i.e., true positives and true negatives) among all instances (i.e., whole dataset) (equation 3.1).

$$\text{Accuracy} = \frac{TN + TP}{TN + FN + TP + FP} \tag{3.1}$$

- Precision: a fraction of the number of correctly classified relevant instances (i.e., true positives) among the total number of instances classified as relevant (i.e., true positives and false positives) (equation 3.2).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

- Recall: a fraction of the correctly classified relevant instances (i.e., true positives) among all relevant instances (i.e. true positives and false negatives) (equation 3.3).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

- F_1 score: a harmonic mean of precision and recall (equation 3.4).

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

3.3 Location Extraction

The project uses a hybrid approach for geoparsing to extract locations. For toponym recognition, the tokens representing locations are identified using the KBLab/bert-base-swedish-cased-ner model[13], which is based on BERT and fine-tuned for NER using The

[13]https://huggingface.co/KBLab/bert-base-swedish-cased-ner

Stockholm-Umeå Corpus, a collection of Swedish texts from the 1990s that consists of one million words. As for toponym resolution, the location tokens are disambiguated using Nominatim and GeoNames geocoders through Geopy[14], a Python client for several popular geocoding web services. Nominatim retrieves different fields about the location from OpenStreetMap. An output example is available in Appendix B. The descriptions for the attributes are available in the documentation[15]. The project uses the lat, lon, and display_name to represent the location on a map. In some cases, the text might contain two locations, the one with the smaller bounding box (area of corner coordinates) is used, which is, in most cases, a more specific place located in the bigger one (e.g. a street within a municipality). The geocoder services provide the ability to limit the search of locations within a specific country. Since the project is limited to Sweden, the output is limited using this option, reducing the false positives that happen when different countries have places with the same name. Tweets not containing location terms identifying a geographical location are discarded, as shown in Figure 3.3.

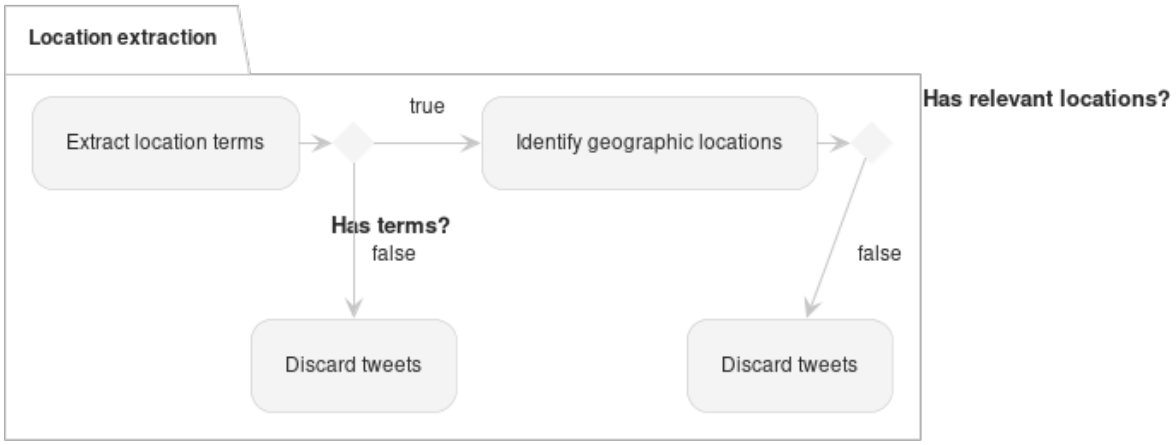


Figure 3.3: Flow chart for the location extraction step of the pipeline

3.4 Text analysis

Further pre-processing is done on the dataset to prepare for text analysis tasks. Lemmatisation is done on the text, using Natural Language Toolkit (NLTK)[16], to reduce words to their lemmas; then, the corpus is tokenized. Terms occurring in less than 20 documents or 5% of the documents are removed, as well as the terms mentioned in more than 75% of the documents. Bigrams that occur more than 20 times in the corpus are included, such as traffic jam, and climate change. To reduce the impact of spambots, tweets created by the same user who tweeted about the same location the past week are discarded.

[14]https://geopy.readthedocs.io/en/latest
[15]https://nominatim.org/release-docs/develop/api/Output/
[16]https://www.nltk.org/

The text analysis used in the project are LDA [36] [11], TF-IDF, and t-SNE [48]. LDA is a topic modelling method that generates topics (a set of terms) in a corpus and assigns the relevance of each topic in each document. Categorizing tweets can show important insights about a flood event, such as incidents caused by floods and their impact on society. The LDA model is initialized and trained using Gensim [37], where the number of discovered topics is adjustable in the visualization. The second text analysis technique used is TF-IDF, using scikit-learn[17], to extract interesting terms by checking their average weight and frequency in the corpus. Lastly, t-SNE is a visualization method for high-dimensional data by reducing their dimensions to two or three-dimensional maps. In this project, t-SNE reduces the dimensions of a TF-IDF matrix generated from the corpus to 2-dimensional space while using the euclidean distance between tweets; afterwards, the tweets are presented on a scatter plot with clusters generated by Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a density-based clustering non-parametric algorithm. Applying dimensionality reduction on data reduces their information, so the clustering is done before applying t-SNE on the tweets. Presenting the tweets using a clustered 2-dimensional space makes it easier to explore tweets for insights about the discussed flood event, such as accidents and infrastructure damages.

3.5 Visualization

Visualization is the final and most significant step of the pipeline since it provides a framework for domain experts to interpret the data and gain actionable insights. Given the nature of the problem the pipeline addresses, which is extracting information about disastrous events, it is crucial to have several plots representing the different aspects of the events. Most hazards impact certain regions during a time interval, so describing the spatio-temporal information of the events is needed to analyze them. Also, the tweets being mainly text allows them to be portrayed visually after applying text analysis techniques. Another direct and beneficial use for visualizing the data is to validate that the pipeline is functioning as intended by navigating through the plots and checking for suspicious results. The web application proposed in this thesis is made by Dash[18] and Plotly[19] python packages. Dash Bootstrap Components[20] is used as well for an easier way to use Bootstrap components for Plotly Dash, such as buttons, input, and tables.

Figure 3.4 shows the visual interface containing all the graphs enabling spatial, temporal, and textual exploration of the tweets: (A) a table showing tweets' properties, (B) a map containing clusters of tweets, (C) a scatter plot for 2-dimensional representation of tweets, (D) tables populated with terms mentioned in the tweets, and (E) a histogram for tweets' creation dates. Users can add filtering rules for the tweets in all the plots using their creation dates, location, and textual properties.

[17] https://scikit-learn.org/stable/
[18] https://dash.plotly.com/
[19] https://plotly.com/python/
[20] https://dash-bootstrap-components.opensource.faculty.ai/

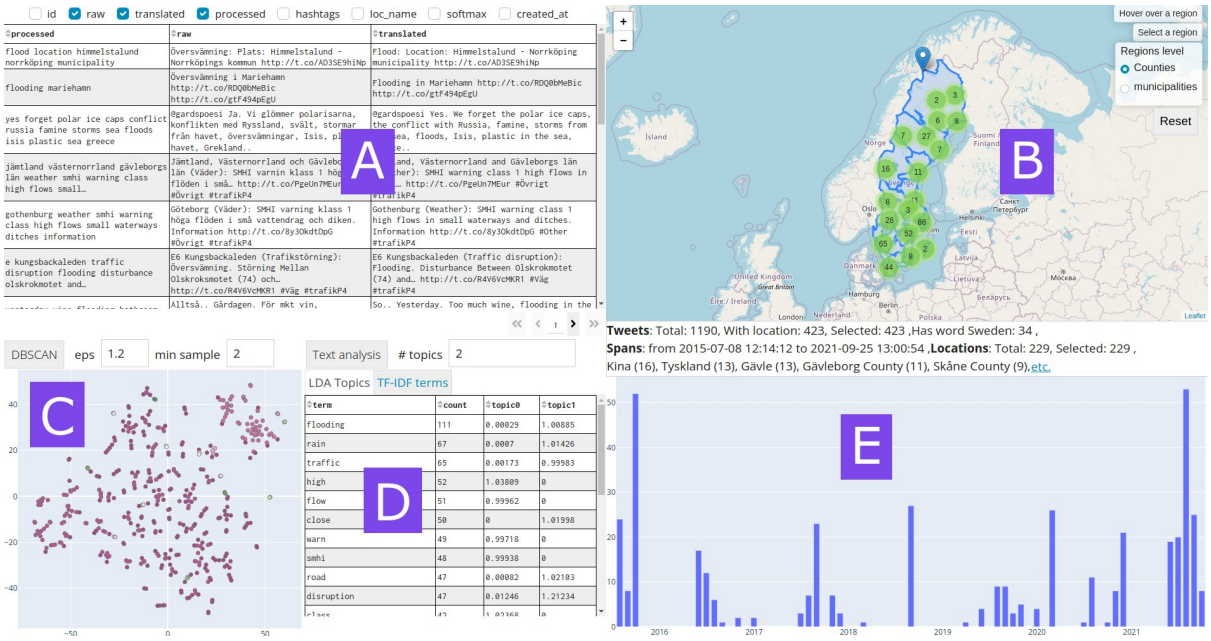


Figure 3.4: The proposed visual interface with 5 plots: (A) tweets table, (B) map for extracted locations, (C) t-SNE scatter plot, (D) tables for LDA and TF-IDF terms, and (E) histogram for tweets' creation dates

The visual interface is designed to fulfil the basic principles for the Visual Information Seeking Mantra: "Overview first, zoom and filter, then details-on-demand"[41]; most of the plots adhere to five of the seven tasks mentioned in the same paper:

- Overview: Having an overview of the whole dataset.
- Zoom: Zooming in on an entry of interest.
- Filter: Disregarding uninteresting entries.
- Details-on-demand: Getting details on a selected entry or group when needed.
- Relate: Viewing the connections among entries.

The metadata in Figure 3.5, which are displayed in the interface between the map (B) and histogram (E) in Fig. 3.4, handles the overview task for the entire dataset. It contains the total number of tweets, the number of selected tweets, the oldest and newest tweet creation dates, the total number of locations, the number of selected locations, and a list of locations' names with the number of their occurrence with an "etc." button to show a pop-over with the rest of the locations.

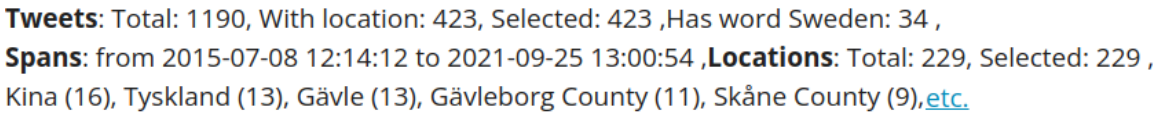


Figure 3.5: Metadata about the visual interface

Zooming helps the users to focus on the portion of tweets they are interested in, and Plotly supports it using the cursor for the map, histogram, and scatter plots. All plots retain the filtering done on each one of them, so the information of the selected tweets is presented after each filtering step, making the selected tweets easier to compare and explore; thus, satisfying the "Details-on-demand" and "Relate" tasks.

Figure 3.6 shows the spatial distribution of tweets using an interactive map containing clickable clustered pointers for the tweets. It makes finding locations mentioned in the tweets more intuitive than using their toponymy. Zooming in or out disperses or congregates clusters, respectively, providing enough clusters at any given moment. Hovering over a map pointer shows a pop-up with the location name extracted by the pipeline. Clicking on a cluster or a region selects the tweets they contain; this will zoom in to cover them while filtering out the unselected pointers from the map. The top right section of the map has several elements: text elements showing the name of the hovered and selected regions, a reset button to remove the current filter, and a radio element to change the spatial resolution of the available regions between counties and municipalities, enabling different granularities to filter tweets.

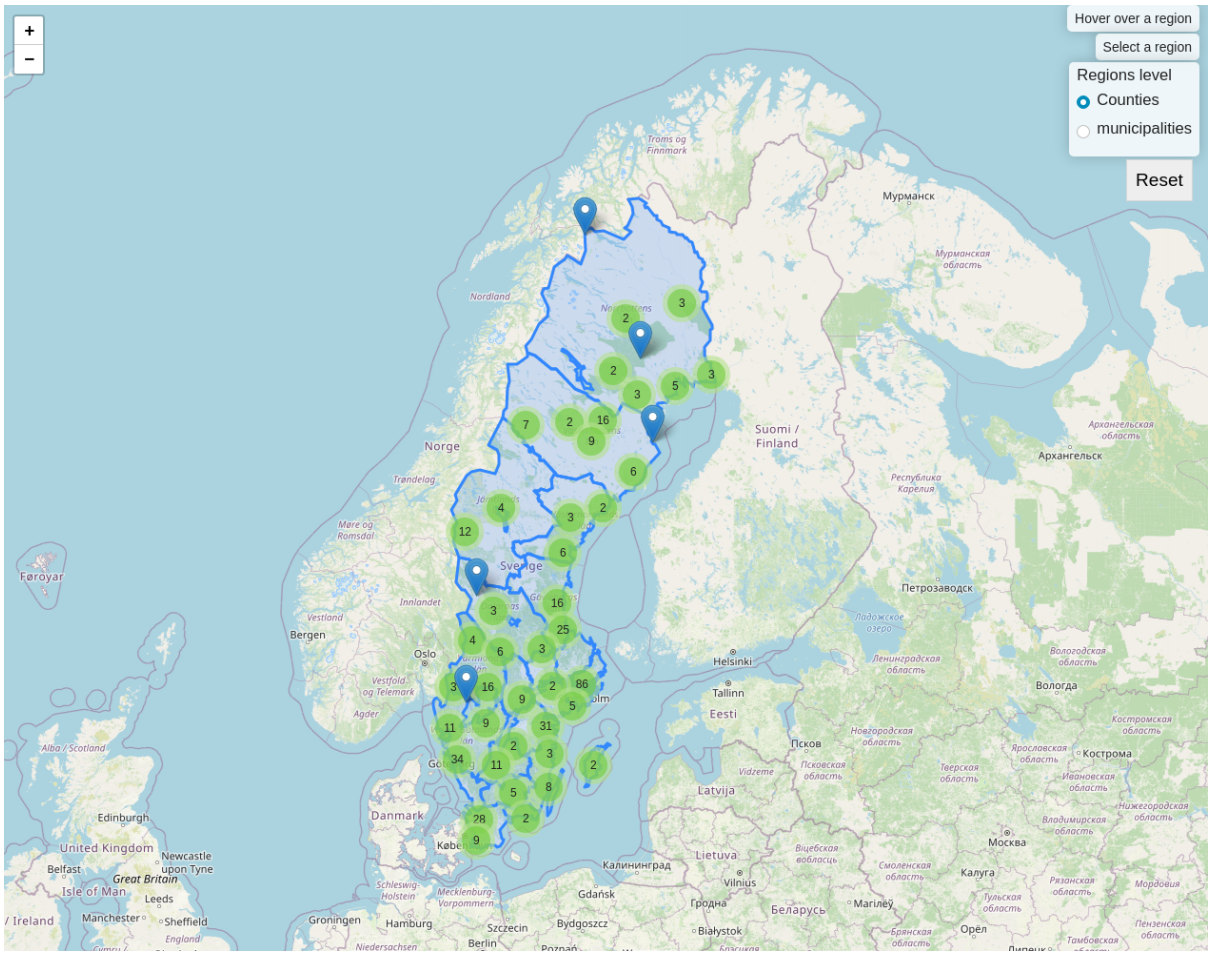


Figure 3.6: Map showing clusters of tweets

Another way to explore the tweets is by showing the temporal distribution of their creation dates using a histogram as seen in Figure 3.7. Event detection is a good use for histograms since they attract the attention of individuals, leading to sudden peaks that dissipate afterwards. The dates are aggregated by day if they span a month or less; otherwise, by month. Tweets selection is done using a select box between two dates, where the blue and red parts of the bars represent the selected and unselected tweets, respectively. Hovering over the bars shows the date and the number of tweets.

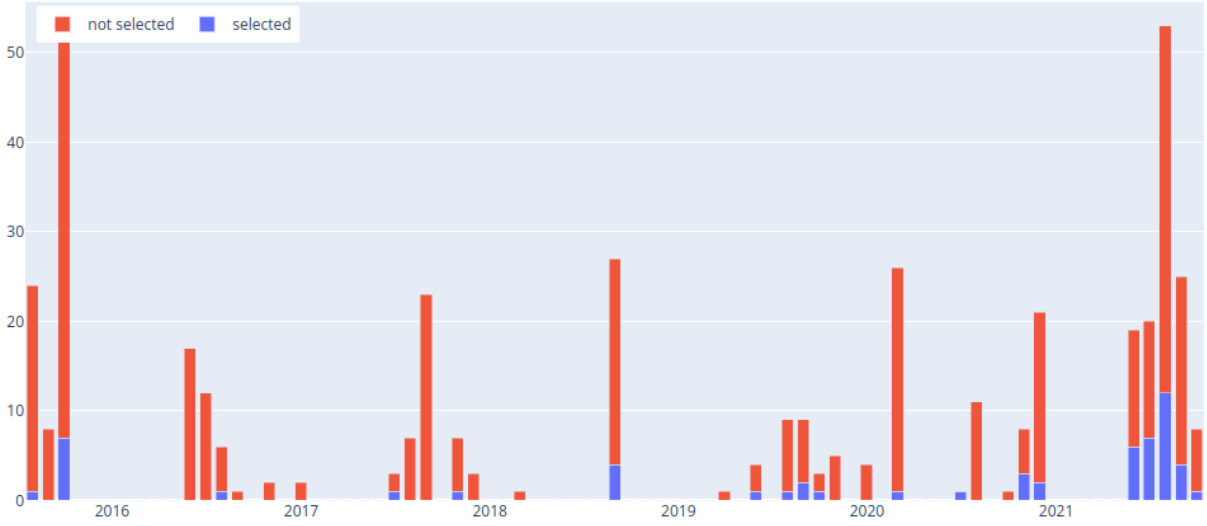


Figure 3.7: Histogram for tweets' creation dates

Tables show all the properties for data in a straightforward way to find patterns while searching and comparing the entries among each other. The table in Figure 3.8 shows the selected tweets with several of their attributes: the ID, the raw (original) text, the translated text, the processed text, the hashtags used, the location extracted, the softmax value for the prediction, and the creation date. The table is sortable and paginated, yet it lacks the features to satisfy the "Filtering" task.

id raw translated processed hashtags loc_name softmax created_at

created_at	hashtags	loc_name	processed	raw	softmax	translated
2015-07-08T12:14:12+00:00		Himmelstalund	flood location himmelstalund norrköping municipality	Översvämning: Plats: Himmelstalund - Norrköpings kommun http://t.co/AD3SE9hiNp	0.945	Flood: Location: Himmelstalund - Norrköping municipality http://t.co/AD3SE9hiNp
2015-07-08T17:31:06+00:00		Mariehamn	flooding mariehamn	Översvämning i Mariehamn http://t.co/RDQ0bMeBic http://t.co/gtF494pEgU	0.95	Flooding in Mariehamn http://t.co/RDQ0bMeBic http://t.co/gtF494pEgU
2015-07-09T19:44:45+00:00		Isis	yes forget polar ice caps conflict russia famine storms sea floods isis plastic sea greece	@gardspoesi Ja. Vi glömmer polarisarna, konflikten med Ryssland, svält, stormar från havet, översvämningar, Isis, plast i havet, Grekland..	0.935	@gardspoesi Yes. We forget the polar ice caps, the conflict with Russia, famine, storms from the sea, floods, Isis, plastic in the sea, Greece..
2015-07-25T22:25:33+00:00	#Övrigt, #trafikP4	Gävleborg County	jämtland västernorrland gävleborgs län weather smhi warning class high flows small…	Jämtland, Västernorrland och Gävleborgs län (Väder): SMHI varnin klass 1 höga flöden i små… http://t.co/PgeUn7MEur #Övrigt #trafikP4	0.942	Jämtland, Västernorrland and Gävleborgs län (Weather): SMHI warning class 1 high flows in small… http://t.co/PgeUn7MEur #Övrigt #trafikP4
2015-07-25T22:28:34+00:00	#Övrigt, #trafikP4	Gothenburg	gothenburg weather smhi warning class high flows small waterways ditches information	Göteborg (Väder): SMHI varning klass 1 höga flöden i små vattendrag och diken. Information http://t.co/8y3OkdtDpG #Övrigt #trafikP4	0.938	Gothenburg (Weather): SMHI warning class 1 high flows in small waterways and ditches. Information http://t.co/8y3OkdtDpG #Other #trafikP4
2015-07-26T06:36:14+00:00	#Väg, #trafikP4	Olskroksmotet	e kungsbackaleden traffic disruption flooding disturbance olskrokmotet and…	E6 Kungsbackaleden (Trafikstörning): Översvämning. Störning Mellan Olskroksmotet (74) och… http://t.co/R4V6VcMKR1 #Väg #trafikP4	0.947	E6 Kungsbackaleden (Traffic disruption): Flooding. Disturbance Between Olskrokmotet (74) and… http://t.co/R4V6VcMKR1 #Väg #trafikP4
2015-07-26T09:12:04+00:00		Turkiet	yesterday wine flooding bathroom work mom came home turkey today happy me	Alltså.. Gårdagen. För mkt vin, översvämning i badrummet och jobb på de. Och mamma kom hem från Turkiet idag. Om hon va glad att se mig?	0.934	So.. Yesterday. Too much wine, flooding in the bathroom and work on them. And mom came home from Turkey today. If she was happy to see me?
2015-07-26T10:25:40+00:00		Dunsjöfjället	smhi warns västra götaland county sjuhäradsbygden göta river warning class high flows small watercourses gothenburg	SMHI varnar - Västra Götalands län, Sjuhäradsbygden och Göta älv: Varning klass 1, höga flöden, i små vattendrag i Göteborg.	0.945	SMHI warns - Västra Götaland county, Sjuhäradsbygden and Göta river: Warning class 1, high flows, in small watercourses in Gothenburg.
2015-07-26T13:08:48+00:00		Gävleborg County	class warning high flows downgraded class applies counties jämtland västernorrland gävleborg	Klass 2-varning för mycket höga flöden har gått ner till klass1. Gäller Jämtlands, Västernorrlands och Gävleborgs län http://t.co/Oqvfd3WmEp	0.946	Class 2 warning for very high flows has been downgraded to Class 1. Applies to the counties of Jämtland, Västernorrland and Gävleborg http://t.co/Oqvfd3WmEp
2015-07-26T18:24:15+00:00	#nyheter, #sverige	Molkom	floods lightning molkom	Översvämningar och blixtnedslag i Molkom - http://t.co/Vs8iLT5K7g #nyheter #sverige	0.941	Floods and lightning in Molkom - http://t.co/Vs8iLT5K7g #nyheter #sverige
2015-07-26T20:10:05+00:00	#Övrigt, #trafikP4	Gävleborg County	jämtlan västernorrland county amp gävleborg county weather smhi warning class high flows small…	Jämtlan, Västernorrlands län & Gävleborgs län (Väder): SMHI-varning klass 1: höga flöden i små… http://t.co/gQSPHURAso #Övrigt #trafikP4	0.944	Jämtlan, Västernorrland County & Gävleborg county (Weather): SMHI warning class 1: high flows in small… http://t.co/gQSPHURAso #Övrigt #trafikP4
2015-07-30T22:41:48+00:00		Lycksele kommun	smhi västerbotten county inland warning class high flows umeälven upstream lycksele	SMHI: Västerbottens län inland: Varning klass 1, höga flöden, Umeälven, uppströms Lycksele. http://t.co/SFNwuUyfmt	0.947	SMHI: Västerbotten county inland: Warning class 1, high flows, Umeälven, upstream Lycksele. http://t.co/SFNwuUyfmt
2015-07-30T22:41:48+00:00		Jämtland County	smhi jämtland county mountains warning class high flows small medium…	SMHI: Jämtlands län utom fjällen: Varning klass 1, höga flöden, små och medelstora… http://t.co/HWFvq2jSyf	0.947	SMHI: Jämtland county except the mountains: Warning class 1, high flows, small and medium… http://t.co/HWFvq2jSyf
2015-07-30T22:41:48+00:00		Gävleborg County	smhi gävleborg county inland warning class high flows small mediumsized watercourses in…	SMHI: Gävleborgs län inland: Varning klass 1, höga flöden, små och medelstora vattendrag i… http://t.co/GWrwebBKm	0.945	SMHI: Gävleborg county inland: Warning class 1, high flows, small and medium-sized watercourses in… http://t.co/GWrwebBKm
2015-07-30T22:41:48+00:00		Östergötland County	smhi östergötland county warning class high flows small mediumsized watercourses…	SMHI: Östergötlands län: Varning klass 1, höga flöden, små och medelstora vattendrag… http://t.co/tRjCJnEuax	0.945	SMHI: Östergötland county: Warning class 1, high flows, small and medium-sized watercourses… http://t.co/tRjCJnEuax
2015-07-31T02:49:18+00:00		Kalixälven	smhi norrbotten county inland warning class high flows kalixälven upstream…	SMHI: Norrbottens län inland: Varning klass 1, höga flöden, Kalixälven, uppströms… http://t.co/6fMMqUZW75	0.946	SMHI: Norrbotten county inland: Warning class 1, high flows, Kalixälven, upstream… http://t.co/6fMMqUZW75
2015-07-31T02:49:18+00:00		Piteälven	smhi norrbotten county inland warning class high flows piteälven upper	SMHI: Norrbottens län inland: Varning klass 1, höga flöden, Piteälven, övre delen. http://t.co/HulTOVHxlC	0.948	SMHI: Norrbotten county inland: Warning class 1, high flows, Piteälven, upper part. http://t.co/HulTOVHxlC
2015-07-31T02:49:18+00:00		Umeälven	smhi västerbotten county southern lapland mountains warning class high flows umeälven…	SMHI: Västerbottens län, södra Lapplandsfjällen: Varning klass 1, höga flöden, Umeälven… http://t.co/6apTulnDOc	0.945	SMHI: Västerbotten County, Southern Lapland Mountains: Warning class 1, high flows, Umeälven… http://t.co/6apTulnDOc
2015-07-31T09:21:19+00:00		Lillå	smhi norrbotten county inland warning class high flows lilla luleälven	SMHI: Norrbottens län inland: Varning klass 1, höga flöden, Lilla Luleälven. http://t.co/vfg6H8rVMk	0.949	SMHI: Norrbotten county inland: Warning class 1, high flows, Lilla Luleälven. http://t.co/vfg6H8rVMk
2015-07-31T09:21:20+00:00		Jämtland County	smhi jämtland county mountains warning class high flows medium lakes…	SMHI: Jämtlands län utom fjällen: Varning klass 1, höga flöden, medelstora sjörika… http://t.co/2Eh30WbeGM	0.944	SMHI: Jämtland County except the mountains: Warning class 1, high flows, medium lakes… http://t.co/2Eh30WbeGM

« ‹ 1 › »

Figure 3.8: Table showing the tweets

Textual data are hard to visualize, requiring a processing step to reduce their complexity; dimensionality reduction transforms the data from a high-dimensional space to a lower one to represent it visually using plots. Figure 3.9 shows a scatter plot for the t-SNE's 2-dimensional space representation of tweets using the euclidean space with DBSCAN clustering, as discussed in the previous section. The clusters can be recalculated after adjusting the properties using the text inputs above the scatter plot: "eps", the maximum distance between two samples for one to be considered as in the neighbourhood of the other; and "minimum samples", The number of samples (or total weight) in a neighbourhood for a point to be considered as a core point including the point itself. Hovering over the points show a pop-up of the text for the tweets, and the points can be selected using a box or lasso selection, meeting the "Filtering" task.

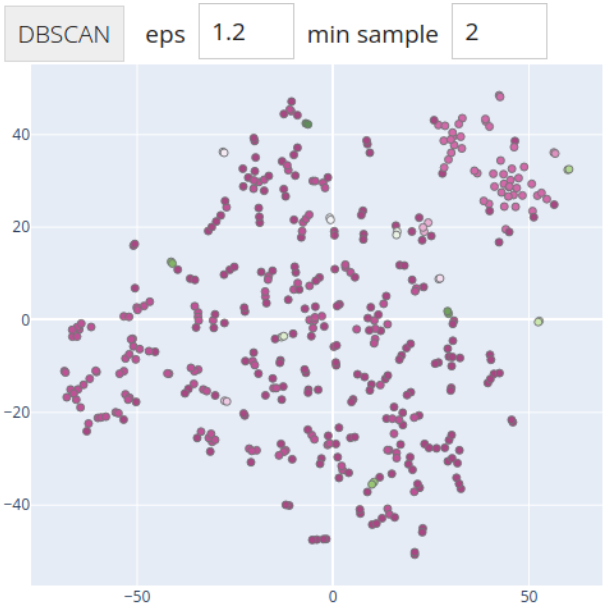
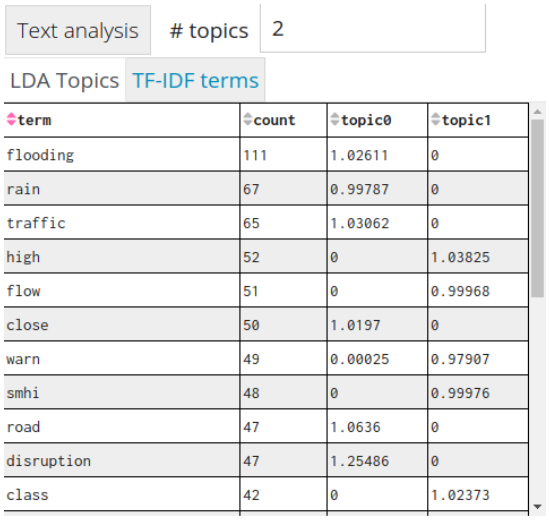


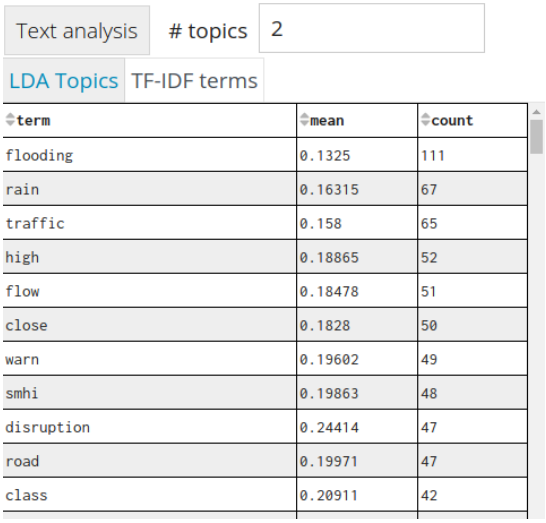
Figure 3.9: Scatter plot for t-SNE's space

The results of LDA and TF-IDF are displayed in two tables (shown in Figure 3.10a and Figure 3.10b, respectively) showing the frequency of the terms and their mean weights. The tables provide a suitable presentation to explore terms by checking their frequency and the topics they belong to. Users can alter the number of topics generated by LDA using a text input and regenerate the tables on the selected tweets by clicking the button.



term	count	topic0	topic1
flooding	111	1.02611	0
rain	67	0.99787	0
traffic	65	1.03062	0
high	52	0	1.03825
flow	51	0	0.99968
close	50	1.0197	0
warn	49	0.00025	0.97907
smhi	48	0	0.99976
road	47	1.0636	0
disruption	47	1.25486	0
class	42	0	1.02373

(a) LDA topic weights



term	mean	count
flooding	0.1325	111
rain	0.16315	67
traffic	0.158	65
high	0.18865	52
flow	0.18478	51
close	0.1828	50
warn	0.19602	49
smhi	0.19863	48
disruption	0.24414	47
road	0.19971	47
class	0.20911	42

(b) TF-IDF weights

Figure 3.10: Tables showing terms with respect their frequency and their weights

Chapter 4

Results

This section presents the results of the methods used by evaluating each step of the pipeline using three datasets from flooding events in Sweden : (1) classifying flood-relevant tweets, (2) extracting geographical locations from tweets, (3) finding useful insights using textual analysis techniques, and (4) interactively visualizing the results.

4.1 Text Classification

Table 4.1 shows the evaluation metrics mentioned in Section 3.2 for the trained DistilBERT model on the dataset and a balanced version by doing undersampling using imbalanced-learn's RandomUnderSampler method[1]. The metrics show that the classifier is performing well with the trained data. Table 4.2 shows falsely classified tweets from the Swedish dataset translated to English.

Table 4.1: Evaluation metrics

	Accuracy	Precision	Recall	F_1 Score	Confusion Matrix
Original	0.9231	0.8944	0.9181	0.9061	$\begin{bmatrix} 381 & 34 \\ 568 & 45 \end{bmatrix}$
Undersampled	0.9137	0.9091	0.9138	0.9115	$\begin{bmatrix} 350 & 33 \\ 370 & 35 \end{bmatrix}$

4.2 Experiments

This section presents the results by applying the pipeline to three unlabelled collections and showing the most noteworthy results from the visualizations. One week's worth of

[1]https://imbalanced-learn.org/stable/

Table 4.2: Miss-classified tweets

Translated tweet	Processed tweet	Predicted	Actual
The road has rained away outside our driveway!! Damn storm https://t.co/wU6uuZo7El	road rained away outside driveway damn storm	0	1
Right now! Stormy weather in southern Norway. Functionality affected - all resources prioritized to save lives, correct in Vestfold.	right stormy weather southern norway functionality affected resources prioritized save lives correct vestfold	0	1
AFTER LIGHT! Basement full of water? Do you live in #Stockholm and are affected by this weekend's #flooding? Call reporter Nadya bums	light basement water live affected weekends reporter nadya bums	0	1
Impressed by efforts and people's patience. Here is the latest municipal information. #Hallsberg #Flooding #orepol #svpol http://t.co/C0sCxEDtLT	impressed efforts peoples patience latest municipal information	0	1
world Floods, war, famine, terror. Goodnight world.	floods war famine terror goodnight	1	0
Flooding in the bathtub?	flooding bathtub	1	0
A basement was flooded when a water main leaked in #Vårberga in #Borgå #borgåvatten https://t.co/zX08QDqJv9	basement flooded water main leaked	1	0
storm flood assumption years Storm flood assumption off by about 2,500 years https://t.co/v14XtEcbTC	storm flood assumption years	1	0

tweets are extracted from Twitter's API starting from the date of the beginning of the events using a query created by experts at a workshop in SMHI containing flood-relevant terms in Swedish:

```
"atmosfärisk flod" OR "hög vatten" OR åskskur
OR regnskur OR dagvattensystem OR dränering OR "höga vågor"
OR "höga flöden" OR dämmor
OR snösmältning OR blött OR oväder OR stormflod OR vattenstånd
OR vattennivå OR åskväder OR regnstorm
OR "mycket regn" OR "kraftig regn" OR översvämningsskador
OR översvämningar OR översvämning
```

Its English translation is the following:

```
"atmospheric river" or "high water" or thunderstorms
 Or "rain shower" or "day water system" or drainage or "high waves"
 Or "high flows" or dams
 Or "snow melt" or wet or storm or "storm river" or "water level"
 Or "water level" or thunderstorms or rainstorm
 Or "very rain" or "heavy rain" or "flood damage"
 Or floods or flood
```

4.2.1 Gävleborg and Dalarna on 18 August 2021

Gävleborg and Dalarna counties had flood events on 18 August 2021[2], damaging their infrastructure, such as houses and roads. After extracting 1589 tweets from Twitter's API and processing them, 910 were left, of which 700 were classified as flood-relevant, and the classifier did a reasonable job at labelling the tweets with few misclassifications. Table 4.3 shows some of the false negatives.

Table 4.3: Miss-classified tweets for floods in Gävleborg and Dalarna

Translated tweet	Processed tweet
Ovädret och det kraftiga regnandet i Gävle har tvingat Brynäs att stänga sin hemmaarena på grund av översvämning. #twittpuck #Brynäs https://t.co/hrZA9icAy7	Ovädret and the heavy rains in Gävle have forced Brynäs to close its home on the ground of overturning. #twittpuck #Brynäs https://t.co/hrZA9icAy7
Blött i Gävle sa Bull.. https://t.co/fV1ChW7ZTR	Wet in Gävle said Bull.. https://t.co/fV1ChW7ZTR
Att tänka på mycket regn bakåt i tiden o tänka på bl.a. ån i Halland som steg o ställde till det !	Thinking about a lot of rain back in time and thinking about e.g. the river in Halland that rose and made it happen!
Nån som vet om det är lite blött i Gävle?	Anyone know if it's a bit wet in Gävle?

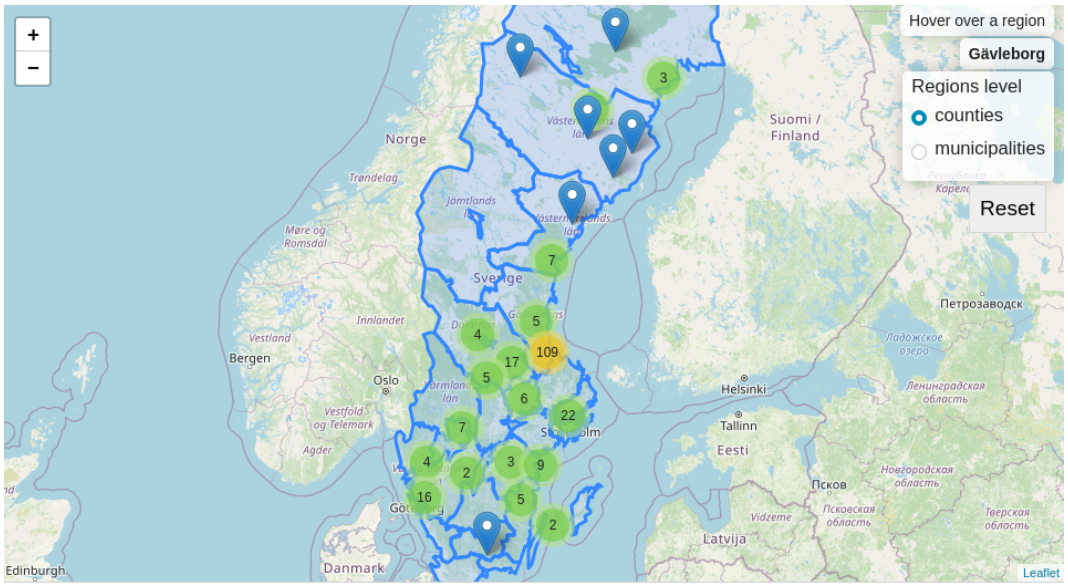
[2]https://floodlist.com/europe/central-sweden-floods-august-2021

The map and metadata in Figure 4.1 show that the tweets mention 104 locations in Sweden within 247 tweets. The location of the incident (Gävle) seems to be identified since it is the highest identified location, with 114 tweets mentioning spots in Gävleborg county tweets, such as "Gävle" (96). Note that the clusters with more than 100 tweets are coloured orange since it is a property of the clustering python package. According to the histogram in Figure 4.2, there's an influx of tweets created on the day of the event, the 18th of August, where 81 out of the 146 tweets have locations in Gävleborg county and 22 out of 51 on the 19th, and 6 out of 16on the 20th. Different types of locations are distinguished correctly, such as counties, municipalities, lakes, and streets; yet, some terms identify locations in other countries or spots, such as:

- **Original tweet**: Dödssiffran stiger i turkiska översvämningar #Turkiet #svpol https://t.co/K6kLRmxQdw
  Translated tweet: Death toll rises in Turkish floods #Turkey #svpol https://t.co/K6kLRmxQdw
  Identified location: Turkiet, a hamlet[3] in Uppsala county.
  Actual location: Turkey, the country.

- **Original tweet**: Information. Det kraftiga regnovädret över Gävle har orsakat översvämningar i arenan. Detta innebär att all verksamhet i Monitor ERP Arena, vilket inkluderar bland annat aktivitet på isen samt restaurangverksamheten, tills vidare är pausad. Vi återkommer med mer information. https://t.co/gHDfirq9VS
  Translated tweet: Information. The heavy rain over Gävle has caused flooding in the arena. This means that all activities in the Monitor ERP Arena, which includes activities on the ice as well as restaurant operations, are paused until further notice. We will return with more information. https://t.co/gHDfirq9VS
  Identified location: Årena, an isolated dwelling[4] in Kalmar county.
  Actual location: Gävle.

[3] isolated settlement
[4] consist of not more than 2 households



Tweets: Total: 700, With location: 247, Selected: 247 ,Has word Sweden: 31 ,
Spans: from 2021-08-17 08:15:09 to 2021-08-22 20:44:07 ,**Locations**: Total: 104, Selected: 104 ,
Gävle (96), Tyskland (7), Brynäs (5), Sundsvall (5), Gävleborg County (5),etc.

Figure 4.1: Map showing tweets about flood event in Gävleborg

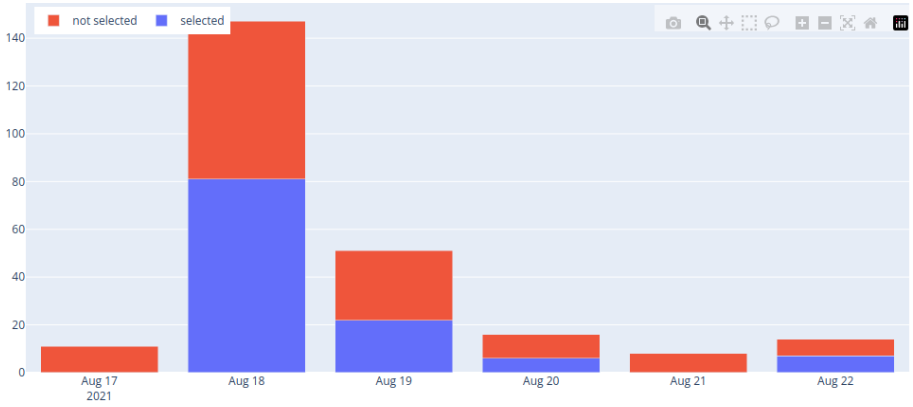
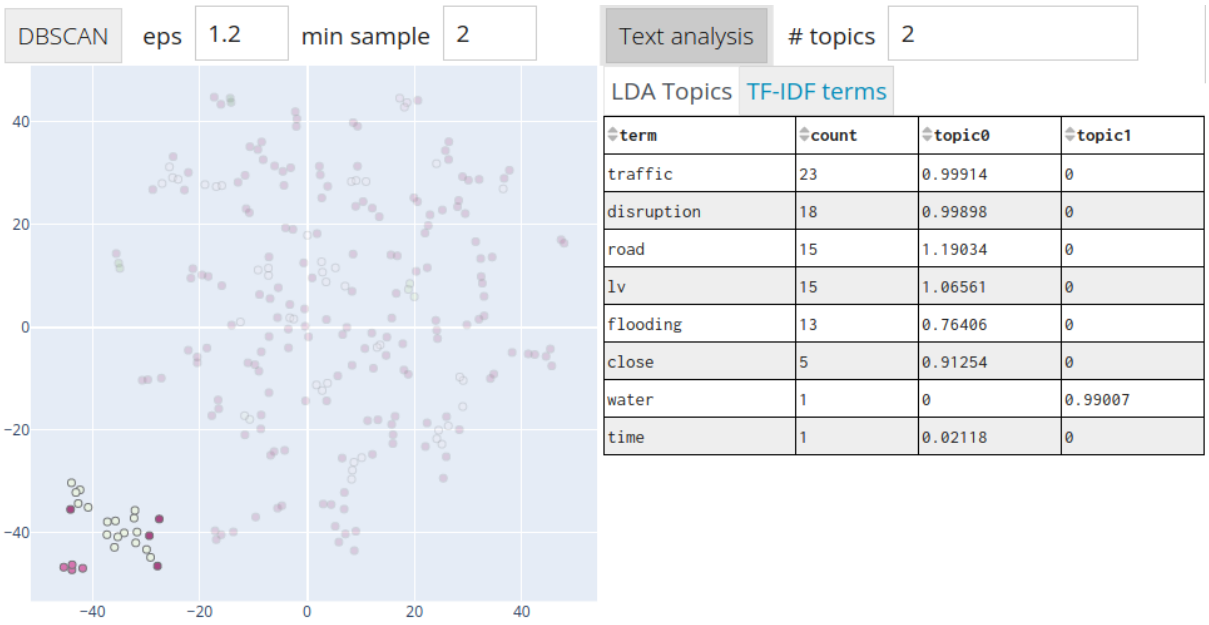


Figure 4.2: Histogram showing tweets about flood event in Gävleborg

After testing with several values for clustering properties, the ones shown in Figure 4.3 identified a traffic disruption through the bottom left cluster. The text in the tweets table (shown in Figure 4.4) and the terms found by LDA show that the tweets discuss a traffic disruption caused by flooding, where LDA is done with two topics only because the number of selected tweets is too small.



term	count	topic0	topic1
traffic	23	0.99914	0
disruption	18	0.99898	0
road	15	1.19034	0
lv	15	1.06561	0
flooding	13	0.76406	0
close	5	0.91254	0
water	1	0	0.99007
time	1	0.02118	0

Figure 4.3: Scatter plot, and LDA table showing a cluster of tweets about flood event in Gävleborg

id raw translated processed hashtags loc_name softmax created_at

processed	raw	translated
e västerås traffic disruption flooding tpl skälbymotet -tpl bäckbymotet direction enköping	E18 Västerås (Trafikstörning) Översvämning. Tpl Skälbymotet (129)-tpl Bäckbymotet (130) i riktning mot Enköping https://t.co/AUwDv7KJyt	E18 Västerås (Traffic disruption) Flooding. Tpl Skälbymotet (129)-tpl Bäckbymotet (130) in the direction of Enköping https://t.co/AUwDv7KJyt
e västerås traffic disruption flooding tpl rocklundamotet -tpl vallbymotet direction örebro	E18 Västerås (Trafikstörning) Översvämning. Tpl Rocklundamotet (132)-tpl Vallbymotet (131) i riktning mot Örebro https://t.co/xUAkMbyRCf	E18 Västerås (Traffic disruption) Flooding. Tpl Rocklundamotet (132)-tpl Vallbymotet (131) in the direction of Örebro https://t.co/xUAkMbyRCf
lv sundsvall traffic disturbance bad road surface storm diversion road outside sawmill	Lv 615 Sundsvall (Trafikstörning) Dålig vägbana efter oväder, på omledningsvägen. Utanför sågverket https://t.co/MAWCkQg308	Lv 615 Sundsvall (Traffic disturbance) Bad road surface after storm, on the diversion road. Outside the sawmill https://t.co/MAWCkQg308
rv borlänge-falun traffic disruption flooding height skommartjärn	Rv 50 Borlänge-Falun (Trafikstörning) Översvämning. I höjd Skommartjärn https://t.co/wMRRQkwbLJ	Rv 50 Borlänge-Falun (Traffic disruption) Flooding. In height Skommartjärn https://t.co/wMRRQkwbLJ
rv delsbo traffic disruption flooding height staffansgården	Rv 84 Delsbo (Trafikstörning) Översvämning I höjd med Staffansgården https://t.co/RHDhhsCG8P	Rv 84 Delsbo (Traffic disruption) Flooding At height of Staffansgården https://t.co/RHDhhsCG8P
lv säter-skenshyttan traffic disruption road closed flood	Lv 655 Säter-Skenshyttan (Trafikstörning) Vägen avstängd. Översvämning https://t.co/KB8voETPtX	Lv 655 Säter-Skenshyttan (Traffic disruption) Road closed. Flood https://t.co/KB8voETPtX
e sandviken-gävle traffic disruption blocking road floods damaged road exit tpl	E16 Sandviken-Gävle (Trafikstörning) Sättning i vägen. Översvämningar har skadat vägen Strax innan avfarten mot tpl Nybo i riktning	E16 Sandviken-Gävle (Traffic disruption) Blocking the road. Floods have damaged the road Just before the exit towards tpl Nybo in the

« ‹ 1 › »

Figure 4.4: Tweet table showing the selected cluster of tweets

Checking the map in Figure 4.5 shows that the same tweets discuss traffic disruptions

in the southern parts of Gävleborg and Dalarna counties. Besides this, the text analysis techniques didn't find anything interesting because the tweets are of small size and composed of other elements besides text to capture their context.

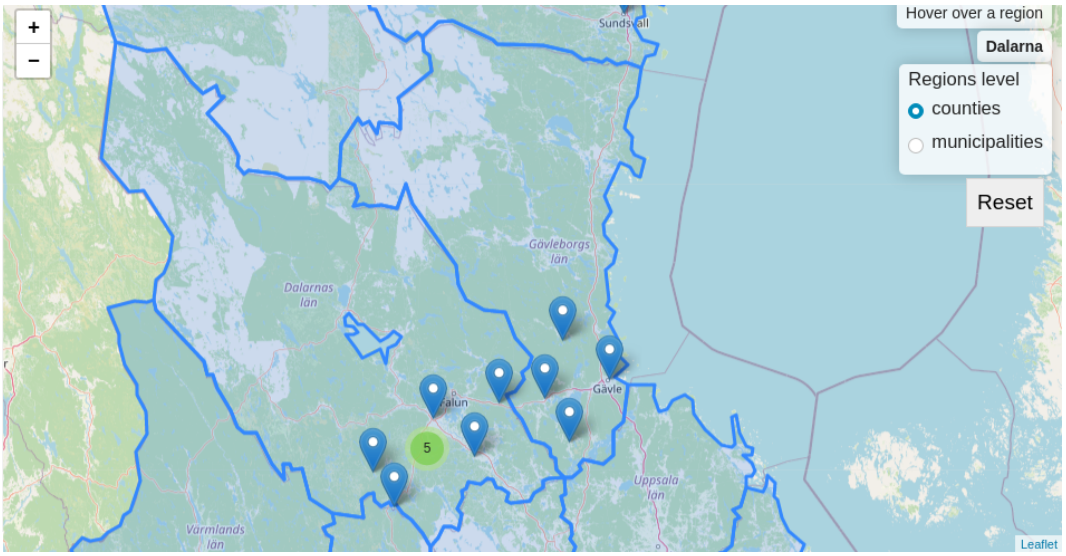


Figure 4.5: Map showing tweets discussing traffic disruption

4.2.2 Gothenburg on 11 September 2019

Heavy rain caused flooding in Gothenburg on 11 September 2019[5]. After extracting 315 tweets from Twitter's API and processing them, 243 were left, of which 117 were classified as flood-relevant, of which 53 contained toponyms. The spatial and temporal distributions of the plots in Figure 4.6 show the true location and starting date of the event, which is evident from the 18 flood-relevant tweets in Gothenburg county, of which 12 were created on the 11th and three on the 12th. With that said, the location extraction step made an error by identifying "Spanien", which was found in 16 tweets, as an isolated dwelling in Stockholm; in fact, the tweets are discussing floods in the country Spain[6]. There are false negatives for classifying flood-relevant tweets, such as "It was a little wet. https://t.co/PcroA3s1A2", where the tweet contains a URL for an article mentioning the flood event.

[5]https://floodlist.com/europe/sweden-flash-floods-gothenburg-september-2019
[6]https://www.svt.se/nyheter/utrikes/stora-oversvamningar-har-drabbat-sodra-spanien

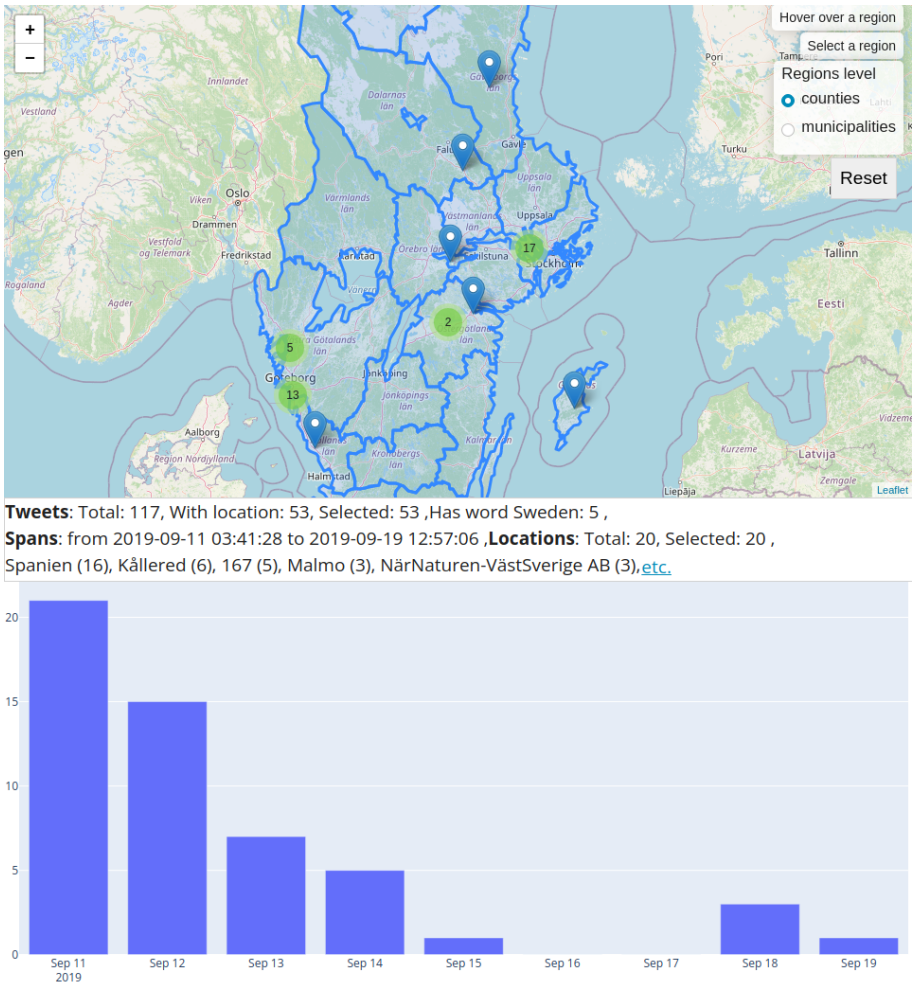


Figure 4.6: Tweet table showing tweets about flood event in Gothenburg

4.2.3 Halland, Värmland and Västra Götaland on 18 and 19 August 2014

On the 18 and 19 August 2014, Halland, Värmland and Västra Götaland counties had floods lasting four days caused by heavy rain[7]. After extracting 1508 tweets from Twitter's API and processing them, 995 were left, of which 503 were classified as flood-relevant, of which 226 contained locations in Sweden. The map in Figure 4.7 shows that Halland, Värmland, and Västra Götaland counties have 74, 41, and 27 tweets, respectively. The histogram shows 15 tweets created on the 18th, 67 on the 19th, and 47 on the 20th.

[7]https://floodlist.com/europe/four-days-floods-sweden

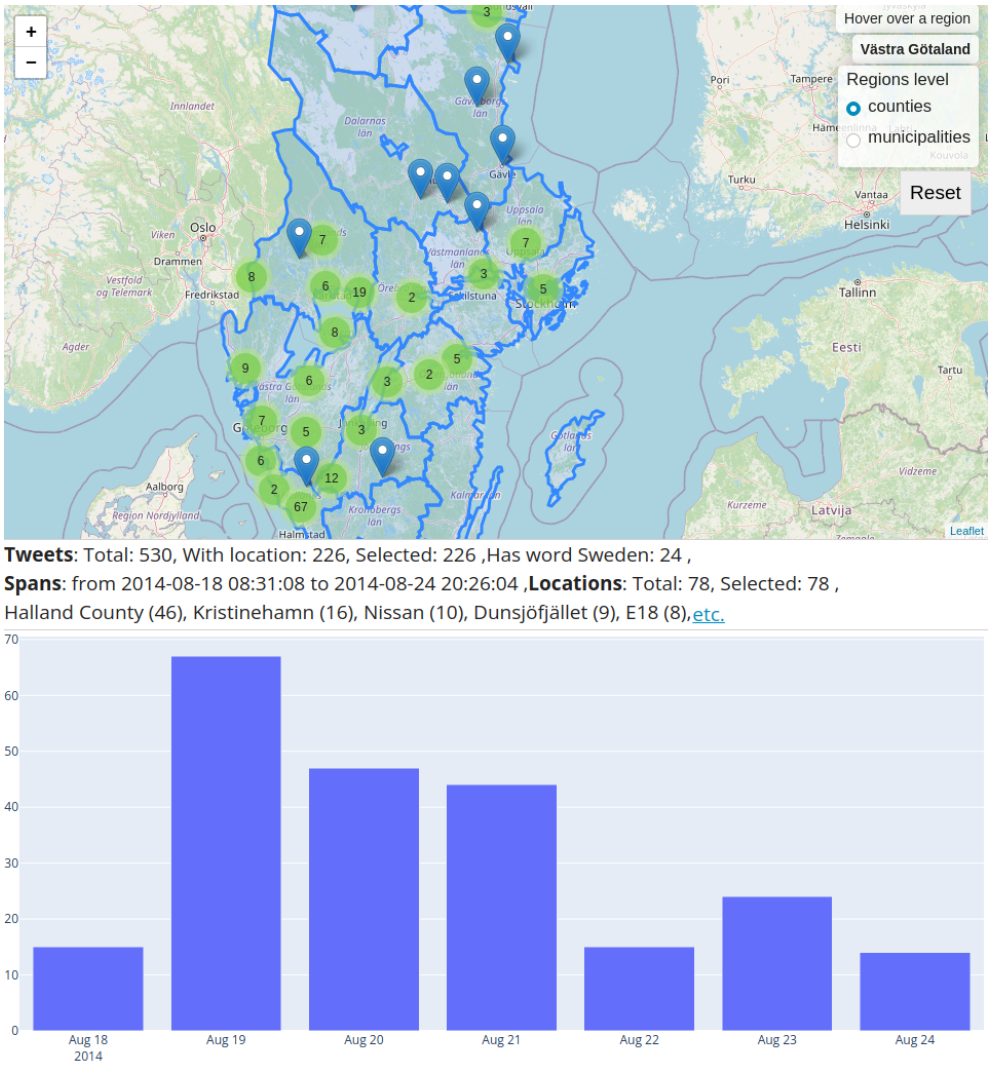


Figure 4.7: Map and histogram showing tweets about flood event in Swedish counties

The selected cluster in the scatter plot shown in Figure 4.8 contains tweets discussing SMHI warnings which is evident from the text in the tweets table and topic 1 in the LDA table.

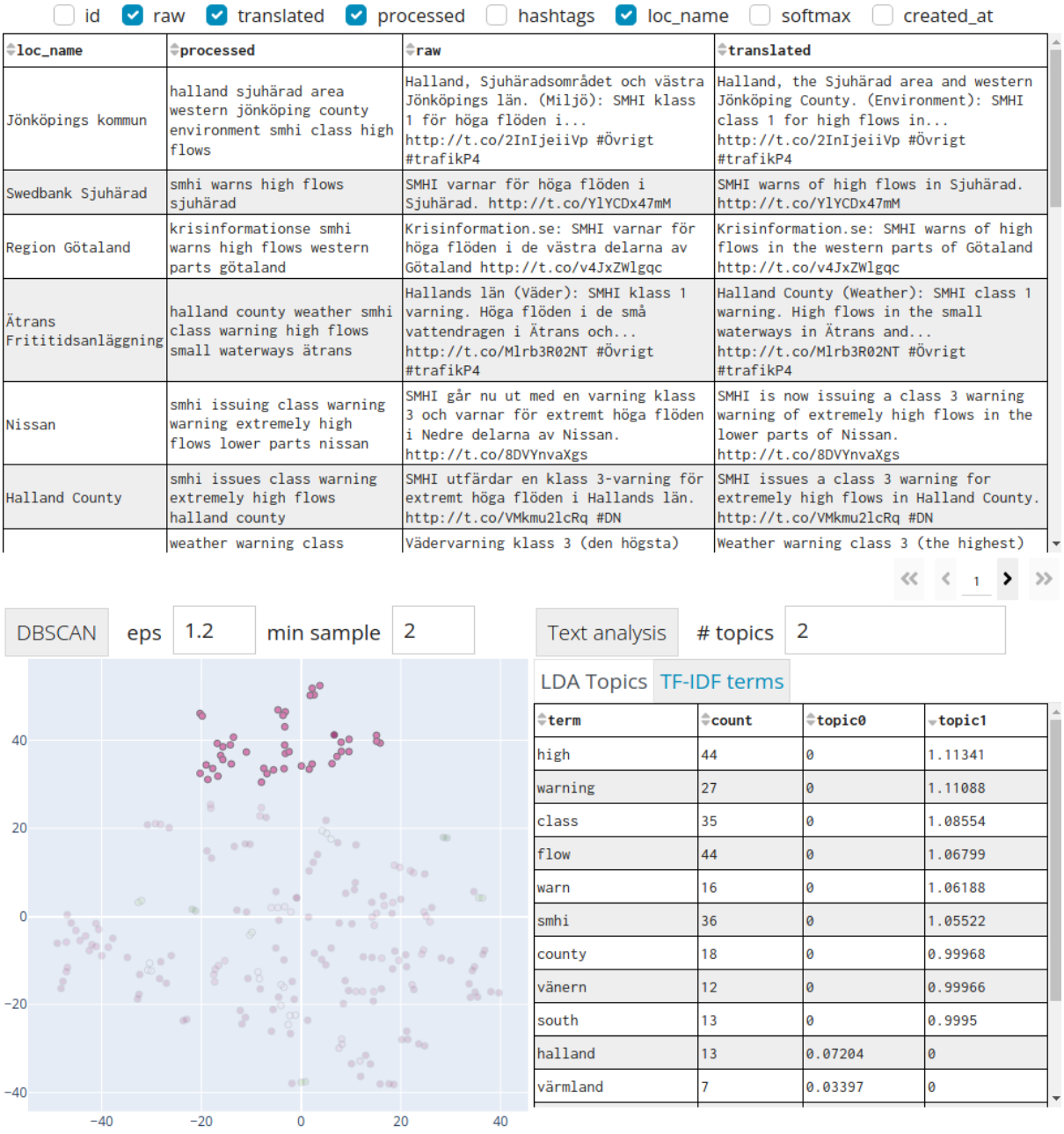


Figure 4.8: Tweet table, scatter plot, and LDA table showing a selected cluster of tweets talking about SMHI warnings

Another cluster of tweets discusses traffic disruptions, as shown in Figure 4.9, and the map in Figure 4.10 shows the locations discussed in these tweets.

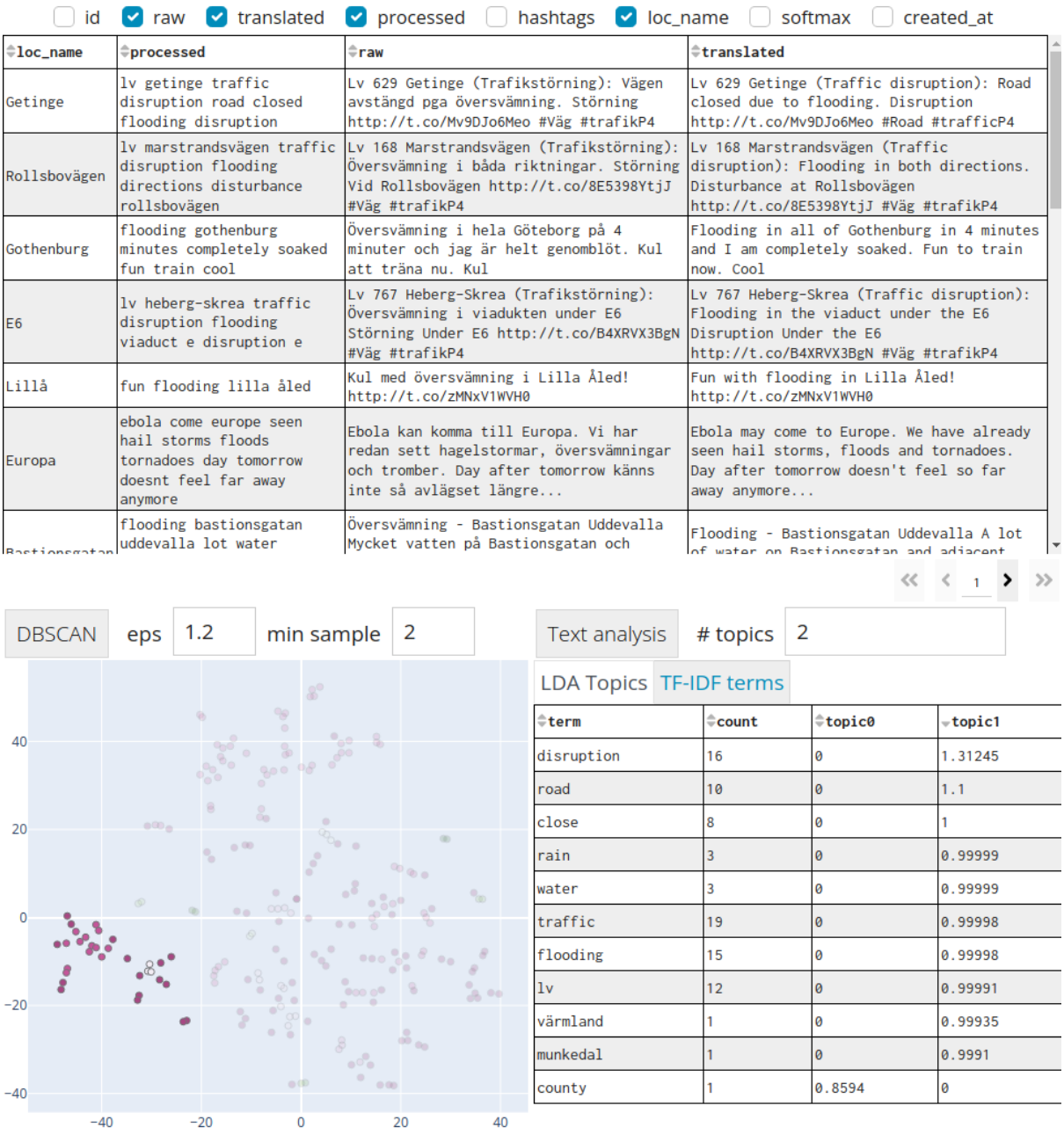


Figure 4.9: Tweet table, scatter plot, and LDA table showing a selected cluster of tweets talking about traffic disruptions

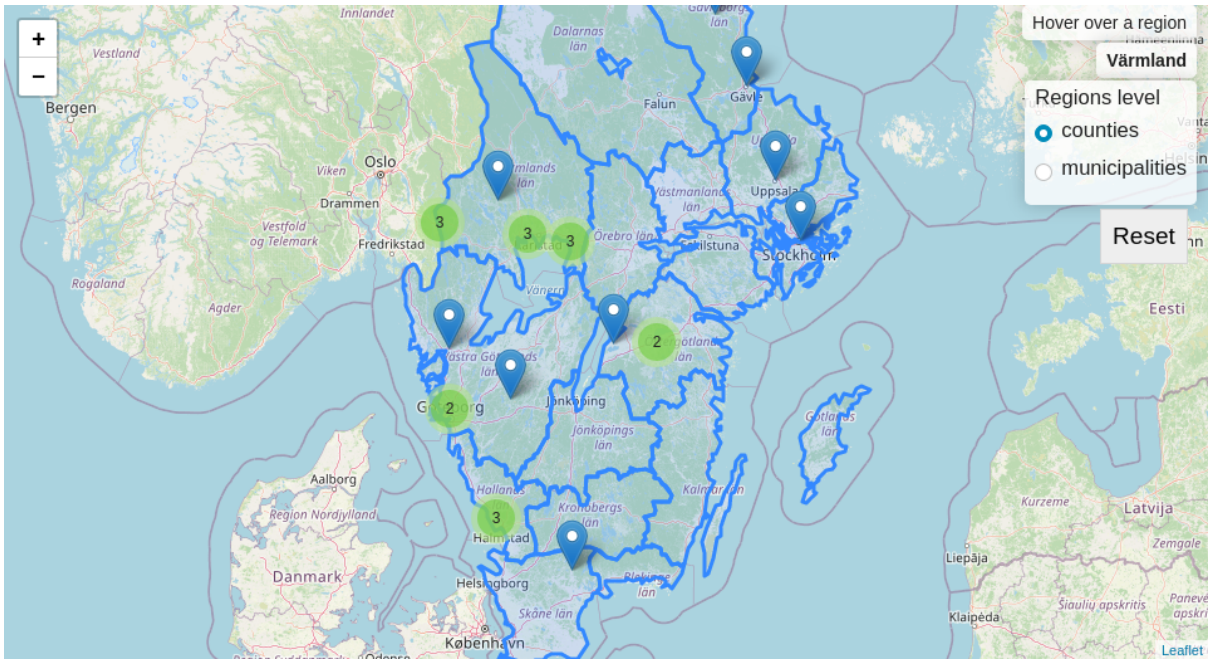


Figure 4.10: Map showing tweets mentioning traffic disruptions

Chapter 5

Discussion and Further Work

This section discusses the results of the methods used and their reliability in addressing the research questions mentioned in the introduction section while suggesting improvements to enhance the results.

The more tweets obtained for an event, the more information can be extracted about it, and the results of the experiments show that some flood events have more tweets than others; one possible reason is that these events are more impactful on society, causing them to attract the attention of the affected citizens. One limitation of this approach is that some events happen without getting reported because they are either out of sight from the public eye, or didn't attract sufficient attention. Social media won't provide sufficient data, if any, about these events to process; integrating more data sources into the pipeline, such as meteorological data, reports from governmental agencies, Global Database of Events, Language and Tone (GDELT)[1], and other social media might supplement this deficiency. The pre-processing used filters out tweets with the same text verbatim, leaving out near-identical tweets generated by bots for malicious or utility (e.g. acting as a feed generator). Some users post tweets that are flood-relevant yet fake to provide the public with misleading information for malicious reasons, polluting the data in the process. One potential improvement to the pipeline is to include methods to detect and filter out fake news to preserve the integrity of the data. Another approach is to blacklist user accounts with suspicious activities in their past tweets and network.

The evaluation metrics for the classifier based on the DistilBERT transformer seem promising on the training datasets, with an accuracy of 92.31% and F_1 score of 91.81%; yet, the experiments show that the model has high precision and low recall. The nature of the text in tweets makes it more difficult for them to get classified. The tweets are microblogs (i.e. small text documents), where there is not enough context in the tweets to classify them correctly (e.g. "It's wet in location X", and "It rained a lot yesterday"); also, since the context of the tweets includes other elements than text, such as emojis, images, hashtags, and URLs, the classifier will not be able to indicate the actual intent of the user, failing in categorizing the tweet correctly (e.g. It's very wet here https://t.co/PcroA3s1A2). Introducing these elements into the classification pro-

[1] https://www.gdeltproject.org/

cess would increase the algorithm's accuracy, which is possible with images and URLs, by using an image classifier and a web scraper. The language of the text is another aspect of the data that impacts the model's accuracy, since the classifier works only on English text, forcing the translation step on the data, reducing their quality, and ultimately impacting the performance of the trained model. One way to mitigate this problem is to remove the translation step and use a classifier that accepts Swedish text.

The pipeline identifies the locations of the events used in the experiments by showing a high frequency of tweets mentioning the location using the plots. Yet, it failed to identify the correct geographical locations for some keywords, such as Spain and Turkey; this is because some terms can refer to multiple locations existing in the world, and one method to identify the intended location within the tweet is by generating a confidence score using several factors, such as the "importance"[2] attribute obtained from the Nominatim package. Another improvement to the location extraction step would be using a better way to handle the existence of several locations in one tweet instead of picking the one with the smallest parameter only.

The plots in the visual interface present the tweets' textual, spatial, and temporal aspects interactively. The map shows the distribution of the geographical locations mentioned in the tweet and enables filtering using regions; yet, the map does not allow box or lasso selections, and the pop-up of the pointers shows the name of the location only, where it could be more informative by including more information related to the tweet (e.g. text and date created). The histogram shows the temporal distribution for the creation date of the tweets, where the number of tweets is the highest at the start of the event then it reduces gradually afterwards. This information can be a factor in calculating the impact of the flood on society since the event attracts more attention the more it influences the citizens. The tweets table provides a way to check some features of the selected tweets, and it could be improved by adding a text filter to focus on tweets that contain specific terms; this feature would support the "Filter" task for the textual data. The test cases show a potential use case for textual filtering by limiting the tweets to the terms related to trends found by exploring the clusters in the t-SNE scatter plot, such as traffic disruption and SMHI warnings. The clustered 2-dimensional space provides a method to find similar tweets using their spatial proximity, potentially referring to similar events, assessed using the spatio-temporal distribution in the map and histogram plots. Besides these trends, the results of the text analysis techniques shown in the t-SNE's scatter plot, LDA table, and TF-IDF table didn't provide any insights due to the nature of the text. Using other clustering techniques in the scatter plot, such as K-means, might bring better results. Obtaining more data and changing the pre-processing approach might improve the results of these techniques. More text analysis techniques can be used, such as sentiment analysis, which gives the ability to quantify the impact of the event.

Further work can include the following:

- Applying the pipeline to other type of events, such as earthquakes, by changing the query and the training dataset for the classifier.

[2]https://nominatim.org/release-docs/develop/customize/Importance/

- Applying the pipeline to other countries by changing the map used in the visualization.
- Using streaming for live event detection to identify flood events by using some criterion, such as sudden bursts of tweets talking about flooding.
- Augmenting warning systems pipeline by including this project's pipeline to detect and visualize flood events.

Chapter 6

Conclusion

This project shows that Twitter can be a great data source candidate to facilitate disaster management tasks. The pipeline extracts information about flood events using the following steps: (1) extracting flood-relevant tweets, (2) identifying geographical locations, (3) finding insights through text analysis, and (4) presenting the results using a visual interface. Even though the methods have room for improvement, they can extract relevant information about past flood events, showcasing the potential of knowledge extraction from social media for disastrous events.

Natural disasters impact human lives severely and will not disappear; they will only worsen due to climate change. With that said, people can try to reduce its impact by preparing for it and repairing the damages it made after dissipating, which is possible by using social media as a data source to predict and analyze events. One problem with this approach is the lack of people's participation, making the amount of data limited, thus leading to inaccurate results or the absence of information to reach them. If this framework gets established globally and people become aware of it, they will be more inclined to share their knowledge on social media to enhance its results. It's a solution for the people and by the people.

Appendix A

Diagrams

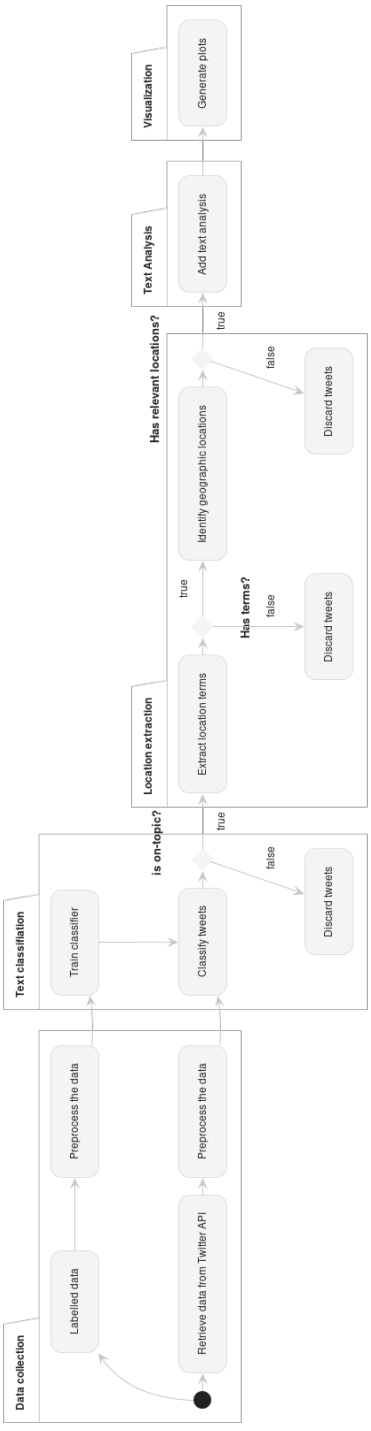


Figure A.1: Flow chart for the pipeline

Appendix B

Examples

B.1 Nominatim output example


```
{
  "place_id": "100149",
  "licence": "Data © OpenStreetMap contributors,
             ODbL 1.0. https://osm.org/copyright",
  "osm_type": "node",
  "osm_id": "107775",
  "boundingbox": ["51.3473219", "51.6673219",
                  "-0.2876474", "0.0323526"],
  "lat": "51.5073219",
  "lon": "-0.1276474",
  "display_name": "London, Greater London, England,
                   SW1A 2DU, United Kingdom",
  "class": "place",
  "type": "city",
  "importance": 0.9654895765402,
  "icon": "https://nominatim.openstreetmap.org/
          images/mapicons/poi_place_city.p.20.png",
  "address": {
    "city": "London",
    "state_district": "Greater London",
    "state": "England",
    "ISO3166-2-lvl4": "GB-ENG",
    "postcode": "SW1A 2DU",
    "country": "United Kingdom",
    "country_code": "gb"
  },
  "extratags": {
    "capital": "yes",
    "website": "http://www.london.gov.uk",
```

```
    "wikidata": "Q84",
    "wikipedia": "en:London",
    "population": "8416535"
  }
}
```

References

[1] Firoj Alam et al. *Flood Detection via Twitter Streams Using Textual and Visual Features*. Version 1. Nov. 30, 2020. DOI: 10.48550/arXiv.2011.14944. (Visited on 10/18/2022).

[2] S. Argamon-Engelson and I. Dagan. "Committee-Based Sample Selection for Probabilistic Classifiers". In: *Journal of Artificial Intelligence Research* 11 (Nov. 15, 1999), pp. 335–360. ISSN: 1076-9757. DOI: 10.1613/jair.612. (Visited on 01/19/2023).

[3] J.L.P. Barker and C.J.A. Macleod. "Development of a National-Scale Real-Time Twitter Data Mining Pipeline for Social Geodata on the Potential Impacts of Flooding on Communities". In: *Environmental Modelling & Software* 115 (May 2019), pp. 213–227. ISSN: 13648152. DOI: 10.1016/j.envsoft.2018.11.013. (Visited on 09/07/2022).

[4] Wikipedia contributors. *Early Warning System*. In: *Wikipedia*. 1119015319th ed. Wikipedia, The Free Encyclopedia, 10/30/2022, 06:41:00 AM. URL: https://en.wikipedia.org/w/index.php?title=Early_warning_system&oldid=1119015319 (visited on 11/17/2022).

[5] Richard Davies. *Sweden – Flash Floods in Dalarna and Gävleborg After Record Rainfall*. FloodList. Aug. 19, 2021. URL: https://floodlist.com/europe/central-sweden-floods-august-2021 (visited on 11/17/2022).

[6] Jens de. *Flood Tweet IDs (Multilingual)*. Version V2. 2019. DOI: 10.7910/DVN/T3ZFMR.

[7] Jens A. de Bruijn et al. "A Global Database of Historic and Real-Time Flood Events Based on Social Media". In: *Scientific Data* 6.1 (1 Dec. 9, 2019), p. 311. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0326-9. (Visited on 10/04/2022).

[8] Jens A. de Bruijn et al. "Improving the Classification of Flood Tweets with Contextual Hydrological Information in a Multimodal Neural Network". In: *Computers & Geosciences* 140 (July 2020), p. 104485. ISSN: 00983004. DOI: 10.1016/j.cageo.2020.104485. (Visited on 11/28/2022).

[9] Jens A. de Bruijn et al. "TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response". In: *Journal of Geovisualization and Spatial Analysis* 2.1 (Dec. 26, 2017), p. 2. ISSN: 2509-8829. DOI: 10.1007/s41651-017-0010-6. (Visited on 10/04/2022).

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. DOI: 10.48550/arXiv.1810.04805. (Visited on 11/26/2022).

[11] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies". In: *Genetics* 164.4 (Aug. 1, 2003), pp. 1567–1587. ISSN: 1943-2631. DOI: 10.1093/genetics/164.4.1567. (Visited on 01/26/2023).

[12] Yu Feng and Monika Sester. "Extraction of Pluvial Flood Relevant Volunteered Geographic Information (VGI) by Deep Learning from User Generated Texts and Photos". In: *ISPRS International Journal of Geo-Information* 7.2 (2 Feb. 2018), p. 39. ISSN: 2220-9964. DOI: 10.3390/ijgi7020039. (Visited on 09/07/2022).

[13] *Floodlist*. FloodList. Aug. 19, 2021. URL: https://floodlist.com/europe/central-sweden-floods-august-2021 (visited on 11/17/2022).

[14] Sabine Gründer-Fahrer, Antje Schlaf, and Sebastian Wustmann. "How Social Media Text Analysis Can Inform Disaster Management". In: *Language Technologies for the Challenges of the Digital Age*. Ed. by Georg Rehm and Thierry Declerck. Vol. 10713. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 199–207. ISBN: 978-3-319-73705-8 978-3-319-73706-5. DOI: 10.1007/978-3-319-73706-5_17. (Visited on 01/17/2023).

[15] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. (Visited on 01/04/2023).

[16] Moritz Heusinger, Christoph Raab, and Frank-Michael Schleif. "Dimensionality Reduction in the Context of Dynamic Social Media Data Streams". In: *Evolving Systems* 13.3 (June 1, 2022), pp. 387–401. ISSN: 1868-6486. DOI: 10.1007/s12530-021-09396-z. (Visited on 02/14/2023).

[17] J J Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." In: *Proceedings of the National Academy of Sciences* 79.8 (Apr. 1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. (Visited on 01/24/2023).

[18] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. May 23, 2018. DOI: 10.48550/arXiv.1801.06146. (Visited on 01/04/2023).

[19] Gao Huang et al. *Densely Connected Convolutional Networks*. Jan. 28, 2018. DOI: 10.48550/arXiv.1608.06993. (Visited on 01/04/2023).

[20] Alex Krizhevsky, Ilya Sutskever, and zz Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Communications of the ACM* 60.6 (May 24, 2017), pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386. (Visited on 12/15/2022).

[21] Quoc V. Le and Tomas Mikolov. *Distributed Representations of Sentences and Documents*. May 22, 2014. DOI: 10.48550/arXiv.1405.4053. (Visited on 12/30/2022).

[22] Hongmin Li et al. “Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach”. In: *Journal of Contingencies and Crisis Management* 26.1 (Mar. 2018), pp. 16–27. ISSN: 0966-0879, 1468-5973. DOI: 10.1111/1468-5973.12194. (Visited on 09/11/2022).

[23] Quanzhi Li et al. “How Much Data Do You Need? Twitter Decahose Data Analysis”. In: July 2016.

[24] Yafeng Lu et al. “Visualizing Social Media Sentiment in Disaster Scenarios”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15: 24th International World Wide Web Conference. Florence Italy: ACM, May 18, 2015, pp. 1211–1215. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2741720. (Visited on 12/16/2022).

[25] Stuart E. Middleton, Lee Middleton, and Stefano Modafferi. “Real-Time Crisis Mapping of Natural Disasters Using Social Media”. In: *IEEE Intelligent Systems* 29.2 (Mar. 2014), pp. 9–17. ISSN: 1541-1672. DOI: 10.1109/MIS.2013.126. (Visited on 10/19/2022).

[26] Stuart E. Middleton et al. “Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging”. In: *ACM Transactions on Information Systems* 36.4 (June 13, 2018), 40:1–40:27. ISSN: 1046-8188. DOI: 10.1145/3202662. (Visited on 10/19/2022).

[27] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Sept. 6, 2013. DOI: 10.48550/arXiv.1301.3781. (Visited on 12/30/2022).

[28] Tina Neset. *AI4ClimateAdaptation*. Linköping University. URL: https://liu.se/en/research/ai4climateadaptation (visited on 11/18/2022).

[29] Huan Ning et al. “Prototyping a Social Media Flooding Photo Screening System Based on Deep Learning”. In: *ISPRS International Journal of Geo-Information* 9.2 (2 Feb. 2020), p. 104. ISSN: 2220-9964. DOI: 10.3390/ijgi9020104. (Visited on 09/11/2022).

[30] Alexandra Olteanu et al. “CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 16, 2014), pp. 376–385. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v8i1.14538. (Visited on 11/28/2022).

[31] Erick Odhiambo Omuya, George Okeyo, and Michael Kimwele. “Sentiment Analysis on Social Media Tweets Using Dimensionality Reduction and Natural Language Processing”. In: *Engineering Reports* (Oct. 11, 2022). ISSN: 2577-8196, 2577-8196. DOI: 10.1002/eng2.12579. (Visited on 02/14/2023).

[32] J. K. Ord and Arthur Getis. “Local Spatial Autocorrelation Statistics: Distributional Issues and an Application”. In: *Geographical Analysis* 27.4 (Sept. 3, 2010), pp. 286–306. ISSN: 00167363. DOI: 10.1111/j.1538-4632.1995.tb00912.x. (Visited on 01/08/2023).

[33] Aditya Pal and Scott Counts. "Identifying Topical Authorities in Microblogs". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM'11: Fourth ACM International Conference on Web Search and Data Mining. Hong Kong China: ACM, Feb. 9, 2011, pp. 45–54. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935843. (Visited on 01/19/2023).

[34] Carlos Periñán-Pascual. "Assessing the Impact of Tweets in Flood Events". In: *1st International Workshop on Social Media Analysis for Intelligent Environment* (Jan. 1, 2020). URL: https://www.academia.edu/44757497/Assessing_the_Impact_of_Tweets_in_Flood_Events (visited on 12/16/2022).

[35] Julie Maria Petersen and Lise Styve. "Identification and Exploration of Extreme Weather Events From Twitter Data". Linköping University, 2021.

[36] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of Population Structure Using Multilocus Genotype Data". In: *Genetics* 155.2 (June 1, 2000), pp. 945–959. ISSN: 1943-2631. DOI: 10.1093/genetics/155.2.945. (Visited on 01/26/2023).

[37] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 22, 2010, pp. 45–50.

[38] *River Floods Sweden*. ClimateChangePost. Nov. 6, 2022. URL: https://www.climatechangepost.com/sweden/river-floods/ (visited on 11/17/2022).

[39] Naina Said et al. *Floods Detection in Twitter Text and Images*. Nov. 30, 2020. DOI: 10.48550/arXiv.2011.14943. (Visited on 11/26/2022).

[40] Victor Sanh et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter". In: *ArXiv* abs/1910.01108 (2019).

[41] B. Shneiderman. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". In: *Proceedings 1996 IEEE Symposium on Visual Languages*. 1996 IEEE Symposium on Visual Languages. Boulder, CO, USA: IEEE Comput. Soc. Press, 1996, pp. 336–343. ISBN: 978-0-8186-7508-9. DOI: 10.1109/VL.1996.545307. (Visited on 02/14/2023).

[42] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015. DOI: 10.48550/arXiv.1409.1556. (Visited on 12/15/2022).

[43] Jyoti Prakash Singh et al. "Event Classification and Location Prediction from Tweets during Disasters". In: *Annals of Operations Research* 283.1 (Dec. 1, 2019), pp. 737–757. ISSN: 1572-9338. DOI: 10.1007/s10479-017-2522-3. (Visited on 09/07/2022).

[44] *SMHI*. SMHI - Who we are. Apr. 30, 2021.

[45] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[46] Yee Whye Teh and Michael I. Jordan. "Hierarchical Bayesian Nonparametric Models with Applications". In: *Bayesian Nonparametrics*. Ed. by Chris Holmes et al. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2010, pp. 158–207. ISBN: 978-0-521-51346-3. DOI: 10.1017/CBO9780511802478.006. (Visited on 01/19/2023).

[47] Lewis Tunstall et al. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Revised edition. Sebastopol: O'Reilly, 2022. 383 pp. ISBN: 978-1-09-813679-6.

[48] L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE". In: *Journal of Machine Learning Research* 9 (nov 2008), pp. 2579–2605. ISSN: 1532-4435.

[49] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. DOI: 10.48550/arXiv.1706.03762. (Visited on 01/24/2023).

[50] Bolei Zhou et al. "Learning Deep Features for Scene Recognition Using Places Database". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://papers.nips.cc/paper/2014/hash/3fe94a002317b5f9259f82690aeea4cd-Abstract.html (visited on 12/15/2022).