Title page

Thesis title

Yaser Kaddoura

Note: When reviewer/examiner decides that the report is close to be finished, contact coordinator for a report number and instructions to produce a title and abstract page.

This page intentionally left blank.

Abstract

This page intentionally left blank.

Contents

List of Figures

List of Tables

List of Acronyms

(Might be needed to make introduction starts at odd page number)

This page intentionally left blank.

1 Introduction

Earlier in this century, floods around Lake Vänern and Arvika have costed Sweden an estimate of 11.1 billions Swedish Krona for damages and repairs [24]. Counties of Dalarna and Gävleborg suffered from flash floods in 2021 disturbing the daily life of their citizens and damaging public and private properties [4]. Flooding is a devastating natural disaster that threatens the lively hood of people and the infrastructure of communities around the world [10].

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 [3]. 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 source of data.

Researchers started harnessing this new wealth of information to aid the disaster management procedure. Twitter streaming Application Programming Interfaces (API) makes it possible to create a monitoring system for early event detection on a global [6] and local [2] scales. Another use for it would be identifying victims in real-time, locating their physical location, and communicating the information to rescue teams [27]. After the threat subsides, emergency managers can use relevant tweets to assess the impact and plan the recovery phase [2]. 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 Swedish Meteorological and Hydrological Institute (SMHI), a Swedish expert authority with a global perspective and a vital task in predicting changes in weather, water, and climate [28].

This thesis project uses 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 visualizations can be used to represent the results?
4. What insights can be extracted by using text analysis on the tweets?

This thesis project is a part of a research project, AI for Climate Adaptation [20] 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, physical locations are extracted from tweets mentioning flood events employing Named-entity recognition (NER) and gazetteers. Finally, the results are presented on a spatio-temporal visualization, and text analysis techniques are done on the tweets. For verification purposes, the pipeline is applied on a week worth of tweets after past flood events.

2 Literature Review

The massive and accessible volume of data that social media produces 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, visualizing the results, and applying text analysis.

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 [17], making it necessary to limit the number of tweets to fetch using the API to reduce computational power and downtime. Feng and Sester [9] only fetches geotagged tweets and then filters necessary them using 45 keywords in 7 languages; this approach filters out a big chunk of relevant tweets since 1% of tweets are geotagged [18]. A better approach is to fetch tweets using keywords related to the topic of interest in different languages. de Bruijn et al. [6] uses over 40 keywords associated with floods in 11 major languages in the query to fetch tweets.

In addition to using textual data, some researchers use other types of data to enhance their pipelines. Some tweets contain media attachments, such as images and videos that are potential visual information for the concerned research topic [1][25][21]; search engines are another resourceful source for images as well [9]. When it comes to flooding events, hydrological information can be a valuable source of information that can be extracted from a global precipitation dataset based on tweets' time stamps and location in the text [7]. Barker and Macleod [2] use Environment Agency flood-monitoring API[1] to get river gauge levels and flood warnings to identify at-risk flooding areas.

Processing text is a crucial part of any Natural Language Processing (NLP) pipeline to train an effective classifier. Research requires that the corpus is in multiple languages, so translating the text to one language (most likely English) is needed if the classifier can not handle multilingual data [27]. One of the most common text-processing tasks is removing unnecessary terms such as stopwords, Uniform Resource Locator URLs, numbers, and punctuation marks. User mentions in tweets don't provide useful information, so pipelines often remove or replace them with a generic term such as "@user"[7]. The location of the flood event is an important piece of information that is extracted from a term in the tweet, making it a potential target that includes biases in the dataset by overusing it; de Bruijn et al. [7] replaces these terms by the country name that the location is located in; on the other hand, Petersen and Styve [23] replace the terms by the word "place" if they get mentioned more than 0.5% of the size of the data set. Another way

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

to improve the performance of the classifier is to group the terms by converting them to lower-case and transforming them to their lexeme or word stem by lemmatisation or stemming respectively.

Some tweets are noisy or redundant, making them a target for filtering out. 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; other reasons could be for 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. [6] only considers one tweet from each user in the last 14 days mentioning a specific region; they also remove 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. [27] approaches this problem by only extracting tweets created from mobile phones and only considers 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 that 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 labeled data for training. A straightforward approach is to manually label a sample of the tweets [6][2]. Petersen and Styve [23] use Crisilext6 [22], a crowdsource labelled tweets, for training their classifiers that get evaluated on 88 million unlabelled tweets containing flood-related terms [5]. Feng and Sester [9] 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's a rainfall, the tweet is labelled positive, negative otherwise.

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 Term Frequency–Inverse Document Frequency (TF-IDF) [30] that reflect how important a word is to a document in a corpus. Word2vec [19] and its extension doc2Vec [15] are other word embedding techniques that 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 in the space.

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 [16]. Feng and Sester [9] use naive Bayes, random forest, logistic regression, Support Machine Vector (SVM) (RBF Kernel), and SVM (Linear Kernel) on

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

labelled data transformed using TF-IDF with accuracies of 0.7109, 0.7582, 0.7705, 0.7712, and 0.7739, respectively. Petersen and Styve [23] 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 and Petersen and Styve train a CNN model on word2vec embeddings with 0.7868 and 0.94611 accuracies, respectively.

Recently, transfer learning has been gaining popularity; it's the idea of transferring knowledge acquired by solving one problem to other related problems. In the case of text classification, Recurrent Neural Network (RNN) is a class of artificial neural networks that process sequences of data using Long Short-Term Memory (LSTM) making it a suitable algorithm for NLP tasks; Petersen and Styve [23] fine-tunes the pre-trained Universal Language Model Fine-tuning (ULMFit) [12] with an accuracy of 0.9499. Models using transforms architecture outperform RNN in many NLP tasks. One popular example is Bidirectional Encoder Representations from Transformers (BERT) [8], a deep learning-based NLP pre-training technique used by generalized models by training on a massive dataset; afterwards, they are fine-tuned on a smaller one for a specific task. Alam et al. [1] uses a pre-trained BERT model that works on one language with an accuracy of 0.853, and de Bruijn et al. [6] uses a multilingual model with 0.8 F1-score.

Visual data such as images are usually classified using CNN models by getting the results of the models after removing the output layer to get a feature vector to train classifiers. These models are pre-trained on massive datasets such as Imagenet [14] and places database [31]. Feng and Sester [9] use GoogLeNet (Inception-V3 model) [29] 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. [21] uses VGGNet [26], Inception V3, ResNet [11], and DenseNet201 [13] with 0.91 accuracy.

2.3 Location Extraction

- Location specific or global - global, local - methods - markov chains using data from user profile and historical tweets - twitter info (geotag, entities, etc.) - 1% of tweets are geotaged [cite:@middletonRealTimeCrisisMapping2014]

2.4 Visualization

2.5 Text analysis

Link to tentative notes for report

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. arXiv: 2011.14944 [cs]. URL: http://arxiv.org/abs/2011.14944 (visited on 10/18/2022).

[2] 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. URL: https://linkinghub.elsevier.com/retrieve/pii/S136481521830094X (visited on 09/07/2022).

[3] 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).

[4] 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).

[5] Jens de. *Flood Tweet IDs (Multilingual)*. Version V2. 2019. DOI: 10.7910/DVN/T3ZFMR. URL: https://doi.org/10.7910/DVN/T3ZFMR.

[6] 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. URL: https://www.nature.com/articles/s41597-019-0326-9 (visited on 10/04/2022).

[7] 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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0098300419308106 (visited on 11/28/2022).

[8] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 11/26/2022).

[9] 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. URL: https://www.mdpi.com/2220-9964/7/2/39 (visited on 09/07/2022).

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

[11] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385 [cs]. URL: http://arxiv.org/abs/1512.03385 (visited on 01/04/2023).

[12] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. May 23, 2018. DOI: 10.48550/arXiv.1801.06146. arXiv: 1801.06146 [cs, stat]. URL: http://arxiv.org/abs/1801.06146 (visited on 01/04/2023).

[13] Gao Huang et al. *Densely Connected Convolutional Networks*. Jan. 28, 2018. DOI: 10.48550/arXiv.1608.06993. arXiv: 1608.06993 [cs]. URL: http://arxiv.org/abs/1608.06993 (visited on 01/04/2023).

[14] 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. URL: https://dl.acm.org/doi/10.1145/3065386 (visited on 12/15/2022).

[15] Quoc V. Le and Tomas Mikolov. *Distributed Representations of Sentences and Documents*. May 22, 2014. DOI: 10.48550/arXiv.1405.4053. arXiv: 1405.4053 [cs]. URL: http://arxiv.org/abs/1405.4053 (visited on 12/30/2022).

[16] 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. URL: https://onlinelibrary.wiley.com/doi/10.1111/1468-5973.12194 (visited on 09/11/2022).

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

[18] 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. URL: http://ieeexplore.ieee.org/document/6692841/ (visited on 10/19/2022).

[19] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Sept. 6, 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781 [cs]. URL: http://arxiv.org/abs/1301.3781 (visited on 12/30/2022).

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

[21] 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. URL: https://www.mdpi.com/2220-9964/9/2/104 (visited on 09/11/2022).

[22] 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. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14538 (visited on 11/28/2022).

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

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

[25] Naina Said et al. *Floods Detection in Twitter Text and Images*. Nov. 30, 2020. DOI: 10.48550/arXiv.2011.14943. arXiv: 2011.14943 [cs]. URL: http://arxiv.org/abs/2011.14943 (visited on 11/26/2022).

[26] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556 [cs]. URL: http://arxiv.org/abs/1409.1556 (visited on 12/15/2022).

[27] 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. URL: https://doi.org/10.1007/s10479-017-2522-3 (visited on 09/07/2022).

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

[29] 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.

[30] Wikipedia contributors. *Tf–Idf — Wikipedia, the Free Encyclopedia*. 2022. URL: https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=1123031029.

[31] 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).

Appendices

A Dummy appendix

A.1 Dummy appendix