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

In early twenties, floods around Lake Vänern and Arvika have costed Sweden an estimate of 11.1 billions Swedish Krona for damages and repairs [7]. Counties of Dalarna and Gävleborg has suffered from flash floods in 2021 disturbing the daily life of their citizens and damaging public and private properties [3]. Flooding is a devastating natural disaster that threatens the lively hood of people and the infrastructure of communities around the world [5].

To facilitate the process of emergency management during these hazardous events, early warning systems analyze their risk, monitor and warn the public while ensuring their readiness [2]. 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 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's stream API makes it possible to create a monitoring system for early event detection on a global [4] and local [1] scales. Another use for it would be identifying victims in real time, locate their physical location, and communicate the information to rescue teams [8]. After the threat subsides, emergency managers can use relevant tweets to assess the impact and plan the recovery phase [1]. To prepare for future floods, authoritative entities can make informed actions by analyzing historical data and determine the locations suffering from recurrent calamity. This new acquired knowledge is able to augment weather warning systems' pipelines improving their accuracy [6].

This thesis project 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 gazetteer. Finally, the identified locations with relevant information from tweets are presented on a spatio-temporal visualization. For verification purposes, the pipeline is applied on a week worth of tweets after past flood events.

References

[1] 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).

[2] 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).

[3] 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).

[4] 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).

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

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

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

[8] 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).

2 Draft

```
*** Abstract
Write abstract last
- What did you do?
  - Extract context from tweets regarding floods
- Why did you do it? What question were you trying to answer?
  - Floods has negative impact
- How did you do it? State methods.
  - Identify tweets that talks about floods
  - Identify locations
  - Visualize the output
- What did you learn? State major results.
- Why does it matter? Point out at least one significant implication
  - Helps for disaster management
  - flood detection specific for swedish lang/sweden
*** Introduction
A monitoring system Using streaming API provided by twitter, a
- How usage of social media can solve the problem
  - Integration with SMHI
  - Forcasting
  - Mention voulenteered geographic information VGI
  - Early detection
  - During disaster management (helping in need)
  - Post disaster analysis (assessing the damage)
  - [[https://www.physio-pedia.com/Disaster_Management][Disaster Management - Physiop
- What to expect in thesis (Problems tackled)
  - focus on floods in Sweden
  - Extracting historical tweets that are flood related
  - extracting locations
  - Visualize the output
  - The pipeline is verified on some past flood events
- [[https://en.wikipedia.org/wiki/Early_warning_system][Early warning system - Wikipe
- [[https://en.wikipedia.org/wiki/Disaster_response][Disaster response - Wikipedia]]
- [[https://en.wikipedia.org/wiki/Emergency_management#Preparedness][Emergency manage
- [[https://www.climatechangepost.com/sweden/river-floods/][River floods - Sweden - C
*** Literature review
- Introduction
- Data collection
  - labeled, unlabaled, autolabeling (page 8 of [cite:@petersenIdentificationExplorat
  - Using twitter api, streaming, 3rd party, images
  - Keywords used (multinlingual)
  - Data processing
    - spam, filtering
```

```
- 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]
- Flood detection
  - transformers
  - CNN
  - word embeding
  - Mention the progression from using basic classifiers to transformers
- Data visualization
- What problems they address
  - Early detection of events via monitiring a stream
  - Disaster management [[https://12ft.io/proxy?q=https://www.physio-pedia.com/Disast
- Maybe evaluate other works
- The location of my solution in the context of the existing literature
*** Methods
- git repo and major libraries used and for what purpose
- reason for picking each method
- packages used in each step
- how to replicate (Refer to README in git repo)
**** Data collection
- Preprocessing
  - Text
  - Tweets
  - Maybe, mention reprocessing in different sections, since each one's
  preprocessing can be different
  - Swedish only, translated to english for classification
- Data sources
  - Twitter API
    - query used
    - major parameters extracted
    - Used for verification an use cases
  - manually annotated Swedish
    - features
    - Available columns
    - amount
    - Training dataset doesn't explicit mention the location,
    - percentage of labeled relevant (is it imbalanced?)
    - language
- Spams
  - why they are bad
  - Reasons for spam
```

  - How to handle them
**** Flood classification
- Transformers - Self-attention - took the world of NLP by a storm -
Can be used for other tasks Their need for big dataset -
Pretrained models that can be fine tuned with smaller datasets to
improve their performance for a more specific task.
Summary of architecture
overthrowing RNN from its throne by a long shot making itself even feasible
for other tasks
More performant (parallize)
RNN attention and its capabilities

HuggingFace is 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.
Supports TensorFlow and PyTorch
It's model hub Contains tens of thousands of pretrained models
Ease of access to the state-of-the-art pretrained model
Has a big amount of datasets.

DistilBERT.1 The main advantage of this model is that it achieves comparable
performance to BERT,
while being significantly smaller than BERT while being significantly smaller and
more efficient.
This enables us to train a classifier in a few minutes.
Use swedish directly or translate it then use english transformer. Given that
the best transformers use english and that
the existence of English labeled datasets, it's more advantageous to translate it.
**** Location extraction
- Talk about Toponym recognition & Toponym resolution
- Usage of NER & gazetteer (Mention packages)
- Extract Swedish locations and one location if more than one two is given in
one tweet (smallest area)
- Mention that geo-referencing isn't viable since users barely use it
(Show reference)
**** Visualization
- Requirements
- Purposes
- What's the difference between this section and the one in Results?
  - Should I discuss the plots here and mention how they fulfill their roles in
  results section?
- Discuss every plot
- Discuss their interactivety
- Mention any significant logic that's being done under the hood

*** Results
- Introductory section
- Flood classification training ::
  eval_accuracy eval_f1 eval_precision eval_recall
  0.95885 0.93506 0.93506 0.93506
- Location extraction
  - What types of locations are being extracted
- Visualization
- Verification (Use cases)
  - Talk about the pipeline (Tweets extracted during a time interval of one week after the flood event)
  - Mention the query and package used
*** Discussion
- Mention briefly that the pipeline successfully identified the concerned flood event by using the
  results as evidence
- How reliable the methods are
  - Classification
- Discuss results
  - Locations
    - How many locations are extracted for that concerned location
    - Comparison between extracted locations and the manually annotated dataset
    - Same keyword can refer to more than one location
- Verification
  - Does the results seem reliable?
  - Maybe make a comparison between two flooding events
  - How accurate each part of the pipeline
    - Are the classified tweets reasonable?
    - Are the extracted locations mention the concerned flood event
    - Show screenshots of the plots while discussing points
    - Does the accuracy differ among the use cases
*** Further works
- Streaming for live event detection
- Identifying flood events explicitly
- Include other types of data images to the pipeline
- Include other sources of data such as GDELT, other social media for training and analysis Handling noise (spammers)
- Increase accuracy for different parts of the pipeline
- Use pipeline for other locations and topics by updating a part of the pipeline (e.g. twitter API query, text pre-processing for different languages)
- Using other languages
- Visualization improvements
- Augment warning systems pipeline by including this approach
- Text analysis for assessing damages (If topic modeling isn't used in main project)

*** Conclusion
- How the results show that this medium has potential to assist in the problems
- List some practical uses the solution gives for disaster management
- Making a grand scale system for anything based on different sources social media and other resources.
  - Challenges that could be faced
- Emphasis that the problem (disasters) won't disappear and they will only become worse in the future and that making use of solutions augmenting the coping for them.

Appendices

A Dummy appendix

A.1 Dummy appendix