Using Transformers to Extract Relevant Information From Twitter During Flood Events

201374125

September 15, 2021

Abstract

Social media presents a source of real-time information provided by individual users in emergency situations. However, due to its unstructured nature and high volume, key information is challenging to extract from these continuous data streams. My dissertation considers the ability to identify relevant flood related Tweets from a Twitter corpus, extracted during dates relating to a historic archive of past severe flooding events in the United Kingdom. A deep neural classification model is constructed, built on the transformer architecture with a pre-trained language model base. My dissertation demonstrates that this model outperforms both simple rule-based approaches, and past machine learning methodologies, often used in past literature. Following classification, spatio-temporal analysis is performed to observe how information through Tweets develops over emergency flood events.

Keywords: twitter; text classification; flooding

In Partial Fulfillment of the Requirements for the Degree of
Data Analytics and Society PhD

UNIVERSITY OF
LIVERPOOL

Contents

List of Figures

List of Tables

1. Introduction

Twitter presents large continuous feed of information regarding emergency events, contributed through individual users, as these events occur. In some cases, the first reports about emergencies on social media often precede those of mainstream media (Perng et al. 2013; Martínez-Rojas, Pardo-Ferreira, and Rubio-Romero 2018; Kim and Hastak 2018; Laylavi, Rajabifard, and Kalantari 2016). Many emergency events have been studied in relation to Twitter, including hurricanes and floods in the US (Hughes et al. 2014; Kim and Hastak 2018), Paris terror attacks in 2015 (Reilly and Vicari 2021), and UK flooding events (Saravanou et al. 2015; Brouwer et al. 2017).

Extreme weather events have become increasingly common (Kron, Löw, and Kundzewicz 2019), a trend that is expected to continue into the future (Forzieri et al. 2017), meaning there is an increasing demand to predict and understand how natural disasters develop. While environmental emergency events are often predictable through meteorological instruments, they generally provide the most detail in high population areas due their high cost, limiting their real-time availability and spatio-temporal scale in many locations (Muller et al. 2015). This is a particular issue for flooding events, where events are typically a smaller scale and time-frame (Willems et al. 2012; Arnbjerg-Nielsen et al. 2013).

Tweets during emergency events often contain more fine-grained locational information (Grace 2020), at a scale that is unlikely identifiable at real-time through meteorological tools (Mazzoleni 2017; Muller et al. 2015). The ability to accurately identify relevant Tweets during flood events presents the opportunity to provide first responders with first-hand information regarding floods as they develop, while ignoring the large amounts of noise that often overpowers the informative information (Ashktorab et al. 2014).

Much of the past work that has used Twitter to study past emergency events has used keywords to identify relevant Tweets (Kryvasheyeu et al. 2016; Brouwer et al. 2017; Morstatter et al. 2013). This however has several issues; keywords are human selected, meaning they require a pre-existing knowledge of the semantics used to describe targeted events. Certain keywords also do not always relate to these emergency events (Sakaki, Okazaki, and Matsuo 2010; Spielhofer et al. 2016), for example a person may be in *'floods of tears'*. Finally, Tweets relating to emergency events also do not necessarily contain an obvious keyword, and therefore are unable to be detected.

More recent work has considered the ability to use machine learning to classify Tweets into those relating to emergency events, and those that are unrelated (Imran et al. 2020; Arthur et al. 2018; Sakaki, Okazaki, and Matsuo 2010; Li et al. 2018). These studies have utilised a variety of methods, building from classical approaches like Naïve Bayes classification (Imran et al. 2013; Li et al. 2018) and Support Vector Machines (Caragea et al. 2011; Sakaki, Okazaki, and Matsuo 2010), while more recent work has considered the emerging prevalence of neural networks in text-based classification (Caragea, Silvescu, and Tapia 2016, 2016; de Bruijn et al. 2020).

Most analysis of social media data relating to emergency events is retrospective, and work that considers the ability to *predict* emergency events through social media primarily relates to *event detection* in

literature (Castillo 2016). My dissertation considers the ability to target and classify Tweets relating to past *known* flooding events in the United Kingdom. This therefore represents a retrospective approach that may be applied to future events in real-time. Such events are already suitably predictable at a general scale using existing meteorological equipment, suggesting pure *event detection* is not required. This work aims to demonstrate a pipeline that is primarily focussed on the removal of irrelevant information from the Twitter pipeline during ongoing flooding events, providing information that may be of use to emergency responders.

For classification, my work uses a neural network architecture known as a *transformer* to classify Tweets extracted from past known major flooding events within the United Kingdom. This model was pre-trained using a large amount of Twitter data, allowing for relevant embedded information to be captured prior to additional task-specific training. The transformer architecture proceeds more traditional neural models for text classification like Recurrent Neural Networks (RNNs), by using the attention mechanism to more suitably capture textual context (Vaswani et al. 2017).

2. Literature Review

The use of Twitter data as a source of information relating to emergency events may be seen as a type of crowd-sourcing. Arthur et al. (2018) suggest a distinction between 'solicited' crowd-sourcing, where citizens are encouraged to contribute directly to data collection efforts, and 'unsolicited', where communication between users on web platforms allow for the indirect extraction of important information. While several attempts have been made to encourage direct contributions to monitor extreme weather events (Met Office, n.d.; European Severe Storms Laboratory (ESSL), n.d.; Muller et al. 2015), there is a heavy reliance on dedicated volunteers choosing to upload data, which is incomparable to the volume of passively contributed data through social media sources. Other forms of 'unsolicited' crowd-sourcing exist for example through call data records (CDR), used to analyse movement patterns in response to natural disasters (Lu, Bengtsson, and Holme 2012), these however often exist as commercial data products, with increased privacy concerns.

In emergency situations, traditionally there would have been a reliance on formal media sources to disseminate information, of which only one-way communication is possible, with the information itself only provided by official organisations (Schneider and Check 2010). With the introduction of social media, there is now a source of two-way communication between both citizens and official sources, meaning information propagates more quickly, and often through first-hand experiences. As part of best practices aimed towards emergency legislation, Lin et al. (2016) suggest that the use of social media for the two-way communication of information is an important source of information, and two-way communication should be encouraged. As of 2021 there are now over 17 million active Twitter users in the United Kingdom (Statista 2021), with a significant number using mobile devices, meaning there are often multiple accounts that are able to communicate specific information regarding emergencies as they occur. The speed of communication through social media is also much faster than traditional sources, and many studies have found that the first messages regarding emergencies situations often appear on social media before traditional news sources (Kim and Hastak 2018; Laylavi, Rajabifard, and Kalantari 2016; Perng et al. 2013; Martínez-Rojas, Pardo-Ferreira, and Rubio-Romero 2018). Twitter specifically has been noted as a platform that through design allows for a good source of information regarding emergency events due to its speed and ease of use (Simon, Goldberg, and Adini 2015; Williams, Terras, and Warwick 2013), and first-hand accounts of emergency responders have noted Twitter provides them with the information they need to assist during a crisis (Brengarth and Mujkic 2016).

2.1. Studying Emergencies through Twitter

Much of the existing work regarding Twitter and emergencies concentrates on event detection. An event may be described as a significant occurrence of something at a specific time and location (Brants, Chen, and Farahat 2003). In social media these are typically characterised by an increase in volume of messages associated with particular topics or entities such as people or places (Dou et al. 2012). With respect to emergencies, Sakaki, Okazaki, and Matsuo (2010) notes that they are distinguishable by their large scale

in which many users are experiencing a particular event, they influence the daily life of a person, and have both spatial and temporal regions. There is also a distinction between retrospective data analysis of emergencies and live data analysis (Castillo 2016), and often with live data analysis the event itself is unknown (Brants, Chen, and Farahat 2003), which Atefeh and Khreich (2015) considered to be *unspecified* event detection.

While some studies consider the ability to detect *unspecified* events through live data analysis of social media (Sakaki, Okazaki, and Matsuo 2010; Dou et al. 2012; Pekar et al. 2020), Castillo (2016) note that for many natural disaster events, the information provided through meteorological observations is likely to be far more informative to predict the initial temporal and spatial information regarding events. However, it is likely that when an event is known, the information provided through relevant social media posts may provide additional fine-grained information about localised incidents as the event progresses, especially as real-time data at a small scale is often unavailable (Mazzoleni 2017; Muller et al. 2015). The type of information provided through Twitter is also different; rainfall data and flood prediction do not necessarily provide information relating to the direct impact on peoples lives, whilst Twitter presents information based on first-hand experiences (Muller et al. 2015).

For this reason, the following work considers only the retrospective analysis of *specified* past flooding events. Conceptually this work aims to demonstrate the ability to capture information *during* known flood events, rather than demonstrate the ability to predict these events, as meteorological tools already do this with good accuracy. This capture of information first relates to the *classification* of relevant Tweets relating to floods, filtering out large proportion of irrelevant information that simply provides no use to responders. Second, the ability to capture information that is not provided through meteorological tools is explored, considering the fact that events exist primarily in a spatio-temporal dimension. Information extracted through my dissertation therefore primarily concentrates on fine-grained locational information that is often communicated during emergency events (Grace 2020), and which cannot be identified through meteorological tools. Temporal information may be captured through analysis of Tweets classified as relevant, and the information extracted through them as events progress.

2.2. Classification of Flood Tweets

Tweet classification methodologies are grouped broadly into either binary classification, or multi-label classification. Binary classifications generally take a stream of Tweets, identifying those that are *informative* or *relevant*, removing any that do not provide useful information or relate to the emergency event being targeted (Ghafarian and Yazdi 2020). This classification is essential given relevant information during emergencies is often diluted due to the large general volume of social media platforms (Lin et al. 2016). Multi-label classification typically focusses on Tweets that are already considered to be relevant to the targeted emergency event, categorising them into groups that indicate the type of information being captured, for example Tweets relating to *missing persons* or *damage* (Ashktorab et al. 2014).

2.2.1. Building a Corpus

The first stage of Tweet classification is the extraction of the Tweets themselves. Studies that have considered Tweet extraction relating to floods sometimes identify *non-specific* events, extracting general Tweets using keywords with Twitters Streaming API over large time periods. For example Arthur et al. (2018) extracted almost 18 million Tweets over a year using the keywords *'flood, flooding, flooded'*, filtering Tweets that did not use a UK time-zone as an alternative to the reliance on geotags to only retrieve Tweets within the UK. Others have targeted specific major flooding events, Spielhofer et al. (2016) were able to tar-

get major flooding events in the UK over 2015-2016, filtering using the keywords *'flood(ing), heavy rain, stormy weather'*, amounting to over 970,000 Tweets over 50 days.

One issue with these studies is that they opt to ignore any geographic information relating to Tweets, due to the very low proportion of Tweets actually containing a geotag (∼1.5%, Spielhofer et al. 2016). This decision led to many of the Tweets extracted being unrelated to any targeted flooding event, even when containing the correct keywords. Spielhofer et al. (2016) for example found that terms like *'floods of tears'* were present, and that many of the Tweets extracted that did relate to flooding were outside of their target area. The use of keywords to filter for relevant Tweets also means that many flood related Tweets may be removed that do not necessarily contain the keywords selected. The two papers mentioned for example select two distinctly different criteria for keywords, meaning even given the same time-period, their corpora would differ.

Alternatively, Saravanou et al. (2015) specifically targeted Tweets with geotags placing them within a bounding box of the United Kingdom over a 5-day period in 2014 during major flooding events, without filtering for keywords. In total a corpus of 2.3 million geotagged Tweets was built. To extract flood related Tweets, a large collection of 456 flood related keywords was created based on a lexicon derived from an initial subset of the full corpus using a much smaller set of initial keywords.

2.2.2. Machine Learning Classification

While the stage at which Tweets were classified differed in the above studies, either at the time of collection (Arthur et al. 2018; Spielhofer et al. 2016), or following collection, after observing the corpus (Saravanou et al. 2015), all have relied on simple keyword extraction to classify relevant and irrelevant Tweets. Using keywords leads to relatively *ad hoc* classifications, based entirely on the authors decision regarding the keywords selected. Other work has considered using more sophisticated methods, using machine learning to classify Tweets relating to emergency events. This work is again split into binary classification of relevant and irrelevant information, and multi-label classification where multiple labels as assigned to Tweets indicating the type of content they include. Rather than pre-selecting keywords, machine learning enables a trained model to select its own *features* to determine the correct classification of Tweets.

As machine learning models require input variables rather than raw text, there is a reliance on suitable *feature engineering* to create input variables for certain model types. Sakaki, Okazaki, and Matsuo (2010) for example generate features using (1) the number of words in a Tweet and the position of a keyword, (2) words within each Tweet represented as an index value, and (3) context surrounding the keywords. With these features they classify Tweets relating to earthquakes or not using a Support Vector Machine (SVM) model. Their evaluation results report an F_1 score of 73.7% when using all features, with most coming from feature (1), notably precision is much lower than recall.

Alternatively, classification of Tweets may consider a two phase approach, first identifying emergency related Tweets using keywords, then classifying those Tweets as related to flooding events or not, attempting to reduce the number of false positives. Ashktorab et al. (2014) present a pipeline for automatically classifying Tweets extracted using keywords, geographical bounding areas and hashtags relating to past emergency events. For classification, a subset of Tweets were assigned a binary classification, noting whether the content related to the emergency event or not. Several classification algorithms were evaluated in relation to this dataset, taking a standard unigram feature vector as input. Logistic regression gave the best F1 score of ∼0.65, but recall was low across all models, suggesting many flood related Tweets were likely missed during classification.

Caragea et al. (2011) present a comparison between the use of SVMs and keywords for the classifi-

cation of text messages relating to the Haiti earthquake into various topics. Various features were used as input for the SVM, including *feature abstractions*, *bag of words*, *topic words*, and *LDA output*, with *feature abstractions* giving the best results. Notably however, the performance of the SVM machine learning approach only slightly outperformed the results of classification using only keywords.

Ghafarian and Yazdi (2020) suggest that methods that place emphasis on certain keywords as a feature input into SVMs, Naive Bayes Classifiers and other machine learning classifiers to identify relevant Tweets likely miss other relevant information. Ghafarian and Yazdi (2020) instead consider each Tweet to be a *distribution* based on word vectors derived from the `Word2Vec` model, suggesting that flood related Tweets will have similar distributions. Using these distributions alongside models like SVM they were able to identify informative Tweets with better results compared with studies using more traditional input features, with F1 scores often above 80%.

A notable issue with the use of these machine learning models for text classification is the reliance on feature representations as input, including bag of words, or statistical features like the proportion of upper case letters, bigrams, *et cetera*. This approach does not preserve word order, and approaches that take keyword inputs also risk over-fitting (Caragea, Silvescu, and Tapia 2016). Alternatively to these traditional machine learning methods, is the potential for neural networks, which in some cases are able to preserve word order, allowing for more accurate representations of the input text.

Caragea, Silvescu, and Tapia (2016) directly compare the results of an SVM against a Convolutional Neural Network (CNN) for the classification and identification of informative Tweets relating to crisis events. Overall their CNN model outperforms several iterations of SVMs that each use various feature types.

Similarly, Lorini et al. (2019) use 'GloVe' word embeddings from as input to an CNN to classify Tweets relating to flooding events in multiple languages, as an alternative to other methods that use bag of word bigrams or unigrams. They note that the use of word embeddings has the ability to capture language specific semantic information that may be lost when using traditional methods.

Ghafarian and Yazdi (2020) compare the results of an SVM for formative flood Tweet classification against a method proposed in Nguyen et al. (2017) that instead uses a CNN with pre-trained word embeddings. Overall performance of the CNN was notably better than the SVM, however Ghafarian and Yazdi (2020) note that for very small dataset sizes, the SVM model performs better.

2.3. Information Extraction from Tweets

2.3.1. Geographic Entity Recognition

Geographic Information Retrieval (GIR) relates to the broad research area that considers the extraction of geographic information from text. This information primarily exists as place names in text, which are extractable using automated tools built using Named Entity Recognition (NER) models.

Recent progress in GIR research has presented models which target place names explicitly, excluding irrelevant entities which are often found with established corpora, traditionally used to train such models (Tjong Kim Sang and De Meulder 2003; Mani et al. 2010). Using NER to identify place names, rather than keyword matching using a gazetteer enables place names to be extracted that may not formally exist. These exist as nicknames, misspellings, and fine-grained place names, all of which are common on Twitter when referring to emergency events (Grace 2020).

2.3.2. Temporal Information

Past studies have considered the delineation of events into their temporal stages, for example *before*, *during*, and *after* (Iyengar, Finin, and Joshi 2011; Kryvasheyeu et al. 2016), considering how the language shifts between these stages.

Kryvasheyeu et al. (2016) found that the per-capita number of Twitter messages corresponds directly with disaster-inflicted monetary damage, with more prominence following the peak of a disaster. They demonstrate that surface level analysis of social media may provide a suitable preliminary rapid damage assessment immediately following a disaster.

2.4. Natural Language Processing

The first major developments in Natural Language Processing (NLP) derive from word embeddings. Word embeddings are an embedded representation of a word, typically a high dimensional vector of floats, mapped to a location in space using some form of unsupervised training procedure. The first word embeddings were generated by Mikolov et al. (2013) using an algorithm called Word2Vec. These embeddings were created from a large corpus of text using a shallow neural network, which enabled the embeddings of words used in similar contexts have similar representations in their corresponding vector space. Word2Vec however is limited by its speed, meaning it becomes inefficient with larger amounts of data, leading to embeddings that are overall limited in accuracy. This was in part solved by GloVe embeddings which, using an optimised algorithm, was able to speed up the creation of word embeddings, while achieving higher accuracy on word analogy and word similarity tasks through unsupervised training (Pennington, Socher, and Manning 2014).

More recent developments in NLP have come through the use of the *transformer* architecture. Vaswani et al. (2017) demonstrated that the attention mechanism is alone suitable for many NLP tasks, moving models away from the purely sequential architectures built on Recurrent Neural Networks (RNNs). Transformers have several key benefits over previously common architectures like Long Short-Term Memory (LSTMs), as attention allows all tokens in a sequence to be weighted against each other, unlike LSTMs where the sequences are processed in order. This allows contextual information to propagate through each word, meaning embedded information relates to that token in its particular context, and not the non-specific learned embeddings derived from general text corpora from pre-trained algorithms like GloVe. This feature also means that transformers can be highly parallelised, allowing for much faster training times on GPUs, despite their larger size. Parallelisation has enabled pre-trained transformer models to be built, often known as *Language Models* which describes their pre-training procedure. The most famous is BERT (Devlin et al. 2019), which used unsupervised training to learn word embeddings as model weights, using the full English language Wikipedia corpus and the Common Crawl. These language models are often fine-tuned to specific tasks like Named Entity Recognition (NER), where weights are updated on a task-specific labelled dataset, modifying the embedded weights that were learned during the pre-training procedure.

My dissertation proposes using the Twitter API v2 to extract Tweets that occurred during past severe flooding events in the UK, within the bounding box for the affected flood zone. This corpus emulates the expected baseline information that would be extractable automatically from Twitter during such events. Tweets will then be classified into *flood related* and *unrelated*, removing noise that is present when extracting information from social media. This flood related corpus is then analysed with respect to its spatial and temporal information. This dissertation aims to demonstrate the ability to automatically extract *relevant* flood related Tweets during known flood events in real-time to assist with emergency response.

3. Methodology

The following section details the methodology involved in the extraction, classification and analysis of Tweets relating to historical *Severe* flooding events in England. The first section, Data Collection, outlines the extraction and cleaning of historic flood related data and Tweets. Sequence Classification describes the construction of a neural network transformer model to identify relevant Tweets from the corpus extracted from Twitter. Spatio-Temporal Analysis of Tweets outlines analysis performed on the extracted Tweets, realising the information extractable that may be useful for emergency response.

3.1. Data Collection

3.1.1. Flood Data

A historical dataset containing all *Severe Flood Warnings*, *Flood Warnings*, and *Flood Alerts* issued by the UK flood warning system is available here under the Open Government Licence. Table 3.1 gives an example entry from this dataset.

For each of these flood warnings, the `CODE` column was linked with *Flood Areas* using the Environment Agency Real Time Flood-Monitoring API. These flood areas are large polygons representing catchment flood zones, and are the smallest resolution geographic area that may be associated with the *Historic Flood Warnings* dataset. The following API query was used to obtain a full dataset of *Flood Areas*:

```
http://environment.data.gov.uk/flood-monitoring/id/floodAreas?_limit=10000
```

To reduce the volume of flood events being considered, only *Severe Flood Warnings* occurring after 2010 were selected This left a total of 314 individual *Severe Flood Warning* events. Figure 3.1 gives an overview of all flood zones, and the flood zones that were identified as having a *Severe Flood Warning* within the historic flood warnings dataset.

3.1.2. Twitter

The Twitter API v2 was used to extract Tweets from the full historic Tweet archive. For each flood warning the query was constructed using several requirements:

- **Time-frame:** 7 days before to 7 days after flood warning

Table 3.1: *Entry from the Historic Flood Warnings Dataset.*

DATE	2020-01-15 07:30:32
AREA	Yorkshire - North and East
CODE	122WAF933
WARNING / ALERT AREA NAME	Lower River Nidd
TYPE	Flood Alert

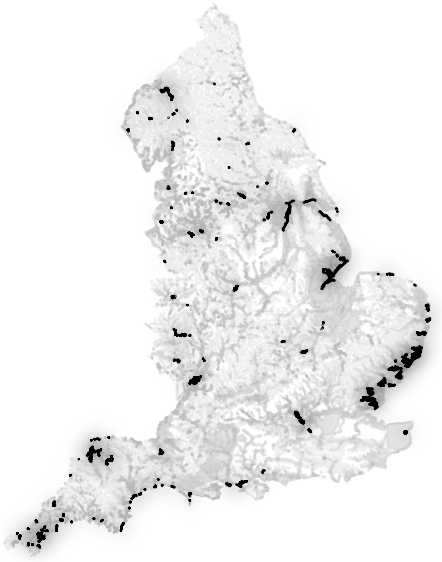

Fig. 3.1: *All flood zones and flood zones relevant to study highlighted in black.*

- **Bounds:** Bounding box of the relevant flood area
- **Parameters:** has *geography*, exclude retweets, exclude replies, exclude quotes

The following gives an example of a query based on the outlined parameters:

```
{
    "endpoint": "https://api.twitter.com/2/tweets/search/all",
    "request_parameters": {
        "query": "has:geo bounding_box:[-5.21508 50.06655 -4.68482 50.36412]
            -is:retweet -is:reply -is:quote",
        "start_time": "2014-01-26T14:28:00Z",
        "end_time": "2014-02-09T14:28:00Z",
        "tweet.fields": "created_at",
        "user.fields": "location",
        "place.fields": "contained_within,country,country_code,full_name,geo,id,name,place_type"
    },
    "max_tweets": 500
}
```

To evaluate the effectiveness of using keywords to identify flood related Tweets from Twitter, every Tweet retrieved that included a selection of keywords were labelled as being flood related (*FLOOD*). Similarly, for all Tweets that did not contain this selection of keywords were labelled as *NOT_FLOOD*. The following keywords were used:

flood, *rain*, *storm*, *thunder*, *lightning*[1]

Geographic information associated with every Tweet was required due to the decision to use bounding boxes to filter out irrelevant Tweets. The new Twitter API now uses a combination of factors to associate geographic coordinates with Tweets which overcomes the issues with limited availability of geotags found with many previous studies (Middleton, Middleton, and Modafferi 2014; Carley et al. 2016; Morstatter

[1] including words that share a stem, e.g. *flooding*, *raining*

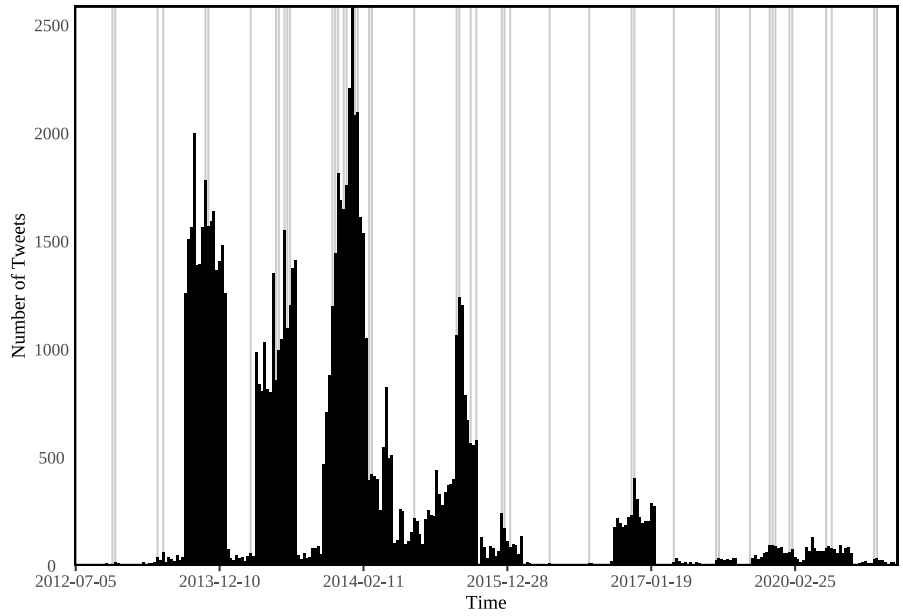

Fig. 3.2: *Timeline showing number of Tweets per day, excluding days with no Tweets retrieved. Vertical lines indicate a day with a severe flood warning.*

et al. 2013). Geography associated with a Tweet may now include either *geotags*, *user profile location* or *locations mentioned in Tweet*.

The total number of Tweets extracted was 89,864, with an average 286 Tweets per flood warning. Tweets were extracted over a period of 14 hours. Figure 3.2 shows the daily Tweet counts extracted for all days with at least one Tweet, with vertical lines indicating days with an issued flood warning.

3.2. Sequence Classification

Figure 3.3 gives an overview of the pipeline for classifying the Tweet corpus. The first stage consists of pre-processing the text in each Tweet, removing properties of the Tweets that may be difficult for the model to understand. Following pre-processing, a subset of Tweets were labelled for use in the model training procedure. Section 3.2.1 outlines the pre-processing procedure for the full Tweet corpora, Section 3.2.2 outlines the architecture of the model used, finally analysis of the model is presented in 3.2.3.

3.2.1. Pre-processing & Labelling

Tweets are typically noisy, with informal language, misspellings and Twitter specific features like user-names prefixed with '@', hashtags '#', and URLs. The following pre-processing procedure is used to remove noise from Tweets, allowing for the subsequent model to focus on the important information conveyed through the remaining body of text.

Usernames tagged in Tweets likely convey some information regarding the Tweet content, but due to the number of unique usernames on Twitter, a model is unlikely to have sufficient information to associate any properties with individual unique usernames. For this reason usernames were normalised, converting any word prepended by '@' into the special string '[USER]'.

```
(@[A-Za-z0-9]+) → [USER]
```

Similarly URLs often do not contain any information as a string, but may provide some context to the

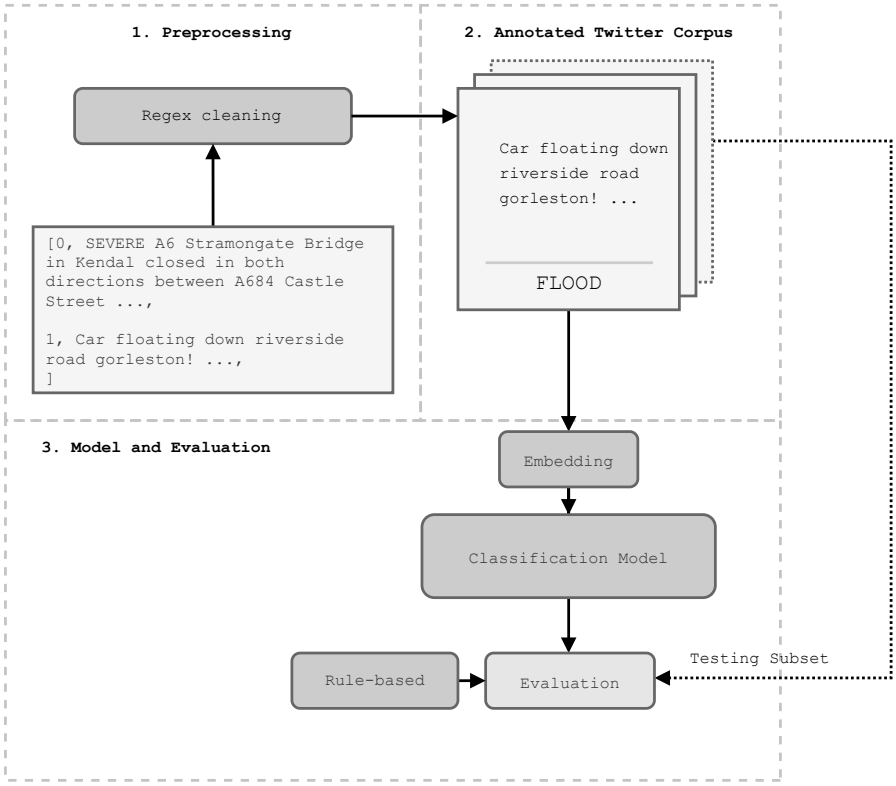

Fig. 3.3: *Overview of the model processing pipeline.*

Tweet, any token starting with 'http' was converted to the special string '[HTTP]'.

```
(http[A-Za-z0-9]+) → [HTTP]
```

Unlike usernames however, hashtags do often provide key information that would be useful for the model to identify. For example a flood Tweet may contain the hashtag #CumbriaFloods. In order for the model to more easily identify words from hashtags, the '#' was removed, and each hashtag was split based on their capitalisation and surrounded in special characters '<' and '>', which allows words from a single hashtag to be grouped, while separating it from the rest of the Tweet (Pota et al. 2020).

```
~#~<(?=[A-Z])>

#CumbriaFloods → <Cumbria Floods>
```

Often Tweets may be very short, or contain no text excluding a URL, therefore Tweets below 10 characters were removed.

For use in a flood Tweet classification model, a subset of 2,000 Tweets were taken to be manually labelled. From this 10% was kept back for validation. Additionally, a testing dataset was kept back for model testing following training, and further analysis with 503 Tweets. Manual labelling was assisted through the open source annotation tool Doccano (Nakayama et al. 2018). When labelling, anything directly related to the flood event was considered to be 'relevant', and therefore given the positive label of *FLOOD*. This follows similar rules noted by de Bruijn et al. (2020), including all rescue and support operations, mentions of flood water movement, and related extreme weather.

3.2.2. Model Architecture

Sequence classification is an established NLP task in which a sequence of tokens ($\mathbf{x} = \{x_0, x_1 \dots x_n\}$) are associated with a single classification (y). In this case, the model constructed uses a pre-trained RoBERTa transformer model as a base. Unlike general use transformers which use a variety of corpora, this models pre-training procedure used only a corpus of 58 million Tweets, as part of the *TweetEval* benchmark task (Barbieri et al. 2020). This model is available on the Huggingface Model Hub (Wolf et al. 2020). To use for sequence classification, this model was extended by pooling the output from the final transformer layer, feeding that into a linear layer with two outputs, one for each label (*FLOOD*, *NON_FLOOD*). This model was developed using PyTorch (Paszke et al. 2019), with the PyTorch Lightning library (Falcon 2019).

All weights in the model were updated each epoch during training to minimise the training loss. Performance was evaluated during training using the F_1 metric on the separate validation data;

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

To prevent training from continuing after performance had stopped improving, early stopping was used, completing training once the validation F_1 score stopped improving. Once training had completed, the model weights that gave the best validation F_1 score were selected to use in the model analysis section.

3.2.3. Model Analysis

Once trained, the classification model was evaluated on the manually labelled test corpus, against the rule-based method which uses keywords to predict flood related Tweets with reference to the F_1 score metric. Whichever performed best was selected to label flood related Tweets for the full corpus extracted for each flood event. Confusion-matrices for both are visualised to demonstrate the distribution of false positive and negatives for both models.

Following this, the transformer model *attributions* for each word in a few selected Tweets are visualised to identify the models ability to capture information relating to flood events, without having to explicitly be fed in key words. For this the Python library `transformers_interpret` was used. Attributions essentially relate the prediction of a deep neural network to its input features, an in-depth description on how this is achieved is given in (Sundararajan, Taly, and Yan 2017).

3.3. Spatio-temporal Analysis of Tweets

3.3.1. Extracting Geographic Entities

Useful geographic information from text is usually present as place names. While many methods for place name identification choose to select place names based on formal corpora, the *Geographic Entity Recognition* model available on the HuggingFace Model Hub considers the ability to identify any place name from text, regardless of whether it appears within formal datasets. This is useful as Twitter users may often use place names that are spelled incorrectly, are fine-grained locations that do not appear in general purpose datasets, or are informal, local names. As with the above sequence classification model, this model uses a pre-trained transformer base, fine-tuned with a set of labelled data. In this case the model was trained on annotated Wikipedia data relating to locations within the United Kingdom.

Named Entity recognition is a subset of token classification, meaning a sequence of tokens ($\mathbf{x} = \{x_0, x_1 \dots x_n\}$) are associted with their most likely sequence of tags of the same length ($\mathbf{y} = \{y_0, y_1 \dots y_n\}$),

from a set number of tags, in this case *PLACE* and *OTHER*. As each token embedding is classified with a tag, this model differs from the sequence classification model in that no pooling is required.

3.3.2. Temporal Analysis

This analysis considers the time-frame for each event as the 7 day run up to the date where the major flood warning was made, to the following 7 days after this warning was made. General analysis is performed, observing how the frequency of Tweets both relevant and irrelevant varies over time and how place name usage varies over time.

4. Results

4.1. Classification Evaluation

Figures 4.1 (A) and (B) show the loss and F_1 scores relating to the training and validation data as the model trained over 5 epochs (starting at epoch 0). Notably on Figure 4.1 (A) the training loss immediately drops to near zero, only slightly decreasing further over subsequent epochs, while the validation loss remains relatively uniform over all epochs. As the training loss drops so quickly, the validation loss never crosses above the training loss value, normally a stage at which model training is considered to be over-fit.

Table 4.1 outlines various evaluation metrics for the Rule-based model against the fine-tuned Transformer model on the separate corpus of manually labelled Tweets for testing. Results overall are relatively similar, with a slight advantage across each metric by the transformer model, excluding the *Sensitivity*. The same value for this metric suggests that while the transformer model is able to more accurately identify when a Tweet relates to a flooding event, they perform similarly when attempting to correctly classify Tweets that are unrelated to flooding events.

Most surprising is the performance of the transformer model on this testing data, relative to its performance on the validation data. In theory both are unseen datasets, so performance should be relatively similar, with only a slight favour towards the validation F1 due to the model selected being based on the highest validation F1 value. The discrepancy between a validation F1 score of around 0.98 and 0.88 for the testing data suggests there may be issues with the training corpus.

A confusion matrix is given on Figure 4.2 showing both the rule-based and transformer model incorrectly classify 31 and 30 Tweets respectively as not flood related (false-negatives). The main difference between the two methods is that false-positives are more prevalent when using a rule-based approach, with 2 false-positives for the transformer model against 14 for the rule-based model. Figure 4.2 also demonstrates the large imbalance in the dataset, as while only 6% of all Tweets are incorrectly classified as not flood related, this reflects 19.7% of all the flood related Tweets in the corpus.

Figure 4.3 further explores the decisions made by the transformer model, using four example Tweets to demonstrate the *attribution* given to each token when assigning a label. Figure 4.3 (A) first gives an example Tweet that is correctly identified as being flood related by the transformer, but does not contain any selected flood related keywords. In this example three keywords are highlighted as important by the

Table 4.1: *Rule-based against Transformer model evaluation metrics on testing corpus.*

	Transformer	**Rule-based**
Accuracy	0.94	0.91
F1	0.88	0.84
Sensitivity	0.80	0.80
Specificity	0.99	0.96

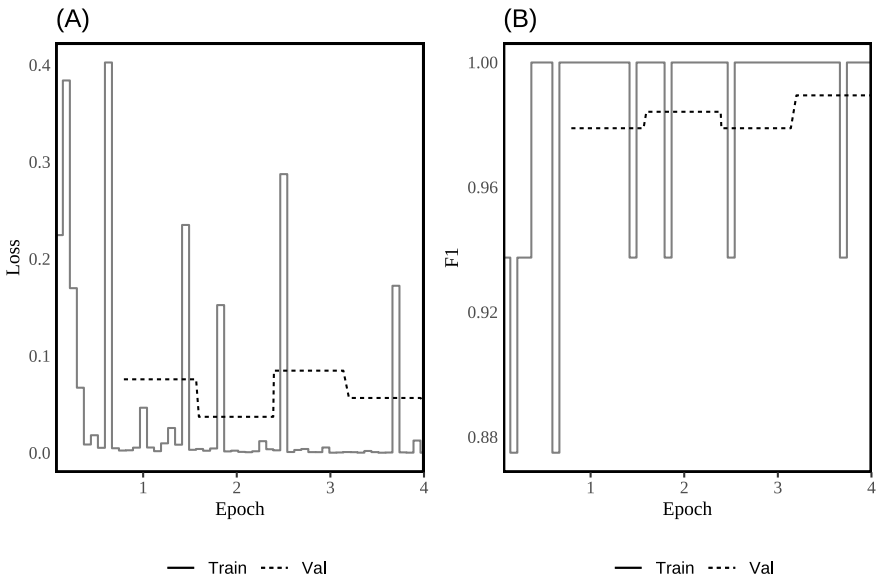

Fig. 4.1: *Evaluation metrics for transformer model during training, showing (A) Loss and (B) F1 Score.*

model for its correct classification *gravel, river* and *wier*. This suggests that the model is able to infer from context that these words relate to flooding, rather than having to be explicitly told.

In the second example on Figure 4.3 (B), an example is chosen where the model was able to correctly identify the Tweet as being unrelated to flooding, but contains the keyword *lightning* which means the rule-based method incorrectly identified it as flood related. Several keywords again appear important for this correct classification, *finally* which is unlikely to appear in Tweets relating to floods, in addition to *apples* and *ipad pro*, both of which likely appear relatively frequently on Twitter, but never in flood related contexts.

The final two figures give examples where the model gives incorrect classifications, but the rule-based method does not. Figure 4.3 (C) shows that while the model realises that *raining* is a word positively associated with flooding, the rest of the sentence implies that the overall Tweet is likely not in reference to a flooding event. This example reflects the issue with selecting a broad annotation scheme, which considered mentions of weather that may relate to flooding events to be a positive match. A Tweet like this is relatively borderline, even for human annotation, meaning it is unsurprising that the model struggles to make a correct decision. This issue is also reflected in Figure 4.3 (D), the words *tide, mark* and *kent* are all identified as flood related words, which is likely true and the label reflects an issue with human annotation.

Table 4.2 gives an overview of the average attribution given to words found within the full corpus of Tweets labelled as relating to flood events. The top two positive attributions are interesting, presenting the words *cyclone* and *tornado*, both of which have strong semantic links with natural emergency events but were not considered for keywords.

4.2. Analysis of Tweets

Figure 3.2 gives an overview of all Tweets extracted from the Twitter API per day, ignoring days when no Tweets were extracted, grey bars indicate days in which a severe flooding event was recorded. Notably, Tweeting frequency appears to drop off quickly after 2015.

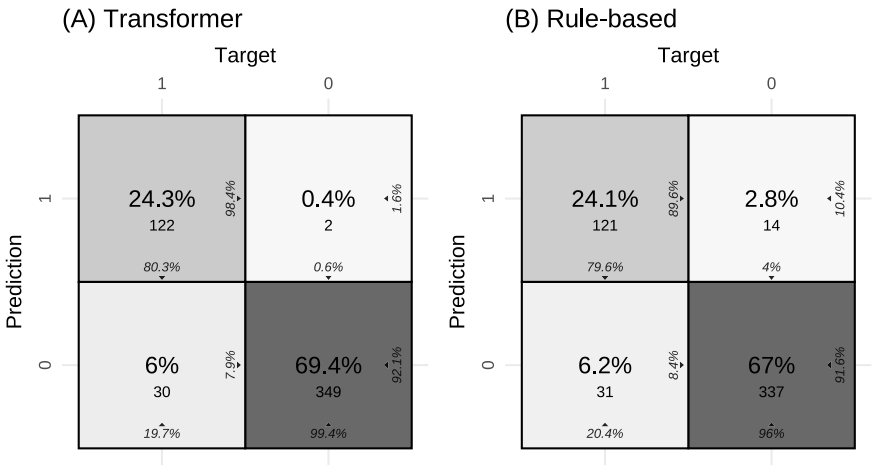

Fig. 4.2: *Confusion matrices comparing the (A) Transformer model and the (B) Rule-based method. Central values represent the normalised count (overall percentage) and the count. Bottom values show the column percentage and right values show the row percentages.*

Table 4.2: *Top 10 positive and negative attributions relating to Tweets predicted as flood related from full corpus.*

Positive		Negative	
word	**attribution**	**word**	**attribution**
cyclone	0.99	orange	-0.50
tornado	0.97	experiencing	-0.22
rains	0.97	sewer	-0.21
storm	0.90	matrix	-0.20
flood	0.89	sundays	-0.18
flooding	0.85	studio	-0.18
floods	0.84	dunes	-0.17
hurricane	0.84	towards	-0.16
thunder	0.78	pushed	-0.16
flooded	0.77	sometimes	-0.16

Figure 4.4 shows temporal changes Tweets, starting 7 days before and ending 7 days following the date the Severe Flood Warning was issued. Figure 4.4 (A) shows total Tweet numbers, with flood related Tweets automatically labelled. Unusually each day appears to have at least some flood related Tweets, only increasing notably in numbers on the date that the severe flood warning is issue, slowly tapering off over the next several days. Figure 4.4 (B) shows how the number of places mentioned in Tweets varies over time. While the number of places mentioned in non-flood related Tweets says relatively uniform over time, there is a very large increase in the number of places mentioned in flood related Tweets on the date relating to severe flood warnings, and several days after. Notably, comparing these two graphs, the increase in place mentions is much larger relative to the increase in number of Tweets relating to floods, suggesting that flood related Tweets contain a much higher proportion of place names.

It should be noted that there is a large variance in the total number of Tweets extracted through Twitter over time as demonstrated on Figure 3.2. Due to this the total number of Tweets extracted from each event varies significantly, for example the mean number of Tweets extracted from flooding events was 298, with a median of 42 a standard deviation of 793, maximum of 6453 and minimum of 1.

(A)

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	FLOOD (0.54)	FLOOD	2.12	[CLS] lots of gravel and debris brought down river kent and deposited on corner below weir in ken ##dal [SEP]

(B)

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	NOT_FLOOD (1.00)	NOT_FLOOD	2.96	[CLS] finally < apples > lightning connector supports usb < number > , but on ipad pro only . via this . [SEP]

(C)

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	NOT_FLOOD (0.00)	FLOOD	-3.20	[CLS] i don ' t like to moan but it ' s raining again ! [SEP]

(D)

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	FLOOD (0.35)	NOT_FLOOD	-1.53	[CLS] the tide mark shows the height the kent got too . [SEP]

Fig. 4.3: *Attribution levels for selected Tweets classified by the transformer model. Attribution label indicates the human annotated label, while predicted label shows the assigned label with a confidence level. Positive attributions indicate tokens that were used to assign the label given by the model, while negative does the opposite.*

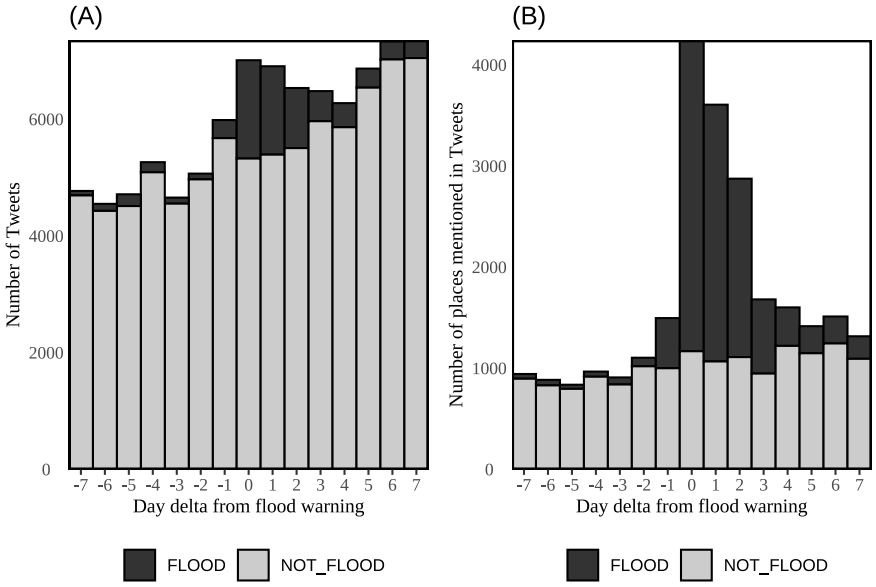

Fig. 4.4: *(A) Total number of Tweets classified as flood related and not, starting 7 days before, to 7 days after severe flood warnings. (B) Total number of place names mentioned in both flood related and unrelated Tweets, starting 7 days before, to 7 days after severe flood warnings.*

5. Discussion

5.1. Classification Evaluation

The training loss dropping quickly below the validation loss on Figure 4.1 (A) likely identifies a potential issue with the network, suggesting that even at a very low learning rate, within one epoch the model is slightly over-fit to the training data. This is again emphasised by the F1 score of 1.0 shown on Figure 4.1 (B), that is both above the validation F1 and suggests the model perfectly predicts flood related Tweets, a strong indication that the model is over-fit. Despite this however, the high validation F1 score does suggest that despite the model being over-fit, it is likely that it will perform well, generalising well with unseen data.

One particular issue with this model is that despite the good performance on unseen validation data, the metrics shown on Table 4.1 and Figure 4.2 are far lower. This issue suggests that it is likely that there is simply not enough labelled data. Tweets themselves are very variable, using unstructured language which often contains misspellings and inconsistent capitalisation. With only a selection of around 500 Tweets selected for testing, 1800 for training, and 200 for validation, the model is unlikely to capture enough information to generalise well enough to perform consistently on unseen data. Despite this however, it should be noted that the model still outperforms the rule-based approach, meaning even with such a small amount of labelled data, the model does capture relevant semantic information automatically. The final issue that is highlighted by this discrepancy is a general issue with supervised learning, especially when considering text-based information. A distinct difference in testing performance against a validation performance suggests that labelling requirements may have been subtly different. This reflects the ambiguity that is often encountered when attempting to assign binary labels to short text. Qualitatively this was observed during annotation, for example the Tweet:

> 'Why do people leave their brains behind when they get behind the wheel of a car? surely the police have enough to do https://t.co/PnWooKQzVF'

At first glance, this Tweet appears unrelated to flooding and as such would be given a negative classification. However, knowing that this Tweet occurred during a flooding event, within the boundaries of the flood zone affected, it then seems more likely than not that police involvement is very likely related to the flood event itself. This Tweet also highlights another issue with the use of a purely text-based model for this task. Viewing the image attached to the Tweet reveals a news article that describing motorists ignoring police barricades to prevent flooded roads from being accessed. This appears to reflect a more emotive response that often comes with Tweets relating to emergency events. While people do often Tweet more when experiencing incidents, their language is more often ambiguous and in a more complex form, which is more difficult to quantitatively assess (Umihara and Nishikitani 2013).

This leads to the primary issue with this classification task. Binary annotation of Tweets into flood related and unrelated is often complicated by ambiguity. This has been noted in past related work, for

example, Olteanu, Vieweg, and Castillo (2015) note that attempting to classify 'informative' Tweets against 'non-informative' Tweets, is a difficult task, even for human annotators. This is also emphasised on Twitter due to the large amount of informal language, which leads to ambiguity when semantics are unclear (Chen, Vorvoreanu, and Madhavan 2014).

There are two considerations that may improve this problem, firstly, it is likely beneficial to consider a more structured definition when determining whether a Tweet is flood related or not. Studies that have considered multi-class classification may find this simpler, selecting classes that are less ambiguous, for example *missing persons*, *property damage*. This would mean that both human annotators, and the model, have more defined targets for labelling, with less ambiguity when a Tweet relates to a flood in a more tangential way. Issues however form if classes are too dilute, limiting the training information available to the model, which would lead to over-fitting. In this dissertation for example, there were no Tweets observed that related to missing persons, most are simply exclamations about the weather, often noting the broad areas affected. The other consideration is the use of multiple annotators, which is common in NLP tasks (Beck et al. 2020). This would highlight instances where ambiguity is common, making them easier to resolve, or identify adjustments to be made to annotation guidelines.

After these considerations, the major question is whether to select a *supervised* classification model or use *unsupervised* keyword extraction. I have demonstrated that it is possible to gain superior results by choosing a supervised methodology, meaning with tweaks to training data, it may be possible to produce a model that significantly outperforms a rule-based approach.

The model presented also appears to outperform past machine learning methods from many past studies. This is likely a result of both the pre-training procedure used by the Twitter specific language model, enabling semantic links between relevant flood related words to be observed (See Figure 4.3), and the complexity of the transformer model itself, which enables weight updates to quickly learn the appropriate embeddings to match the annotated training data. Notably, 4.3 (B) indicates that while lightning is likely considered by the model in most contexts to be associated with floods, the model is able to consider this instance independently, understanding that in this context the word *'lightning'* is not weather related. The attributions on Table 4.2 mirror keyword selection, showing the model automatically highlights keywords that are important for classification, additionally giving them weightings. This mirrors the manual, time-consuming methods of Saravanou et al. (2015) who identify many keywords through observation of a subset of their full corpus.

Overall however, there is still a relatively small difference in the model performance against the rule-based approach, relative to the amount of time required to annotate data for supervised training, and the time required for training. This is also reflected in previous work that considered this comparison, only gaining small improvements using machine learning over rule-based methods for a similar task (Caragea et al. 2011).

Due to the variability and ambiguity of this task, it is tempting to suggest there is a higher influence from outside factors when determining performance metrics. Tweets from other flooding events from past studies vary significantly to those observed in my dissertation, keywords from Saravanou et al. (2015) for example are often event specific. This influence also comes from the annotation of training data, it is likely that given two separate annotators, the results of the model are likely to vary widely. This may also be applied to keyword classification methods, as past studies have selected a variety of keywords, without a method to formally justify their choices.

The dataset imbalance indicated by Figure 4.2 also contributes to some potentially misleading evaluation metrics reported by similar work. Notably Caragea, Silvescu, and Tapia (2016) demonstrate that

using a *Naïve approach* to classification of flood related Tweets, i.e. classifying every Tweet as the majority class, gives results approaching an accuracy of just below 70%, and in some cases only slightly below the results of their SVM models. This also reflects the requirement of using an F1 score as a performance metric over accuracy, as F1 scores are more capable of taking into account class imbalances.

It has been generally demonstrated that the use of deep learning neural networks that take in pre-trained word embeddings enables models to infer their own rules based purely on the raw text and derived semantic contexts. This comes from both the short passage itself, and the data used during the pre-training procedure. This presents an alternative to the *ad hoc* rules created when selecting keywords, or when engineering features for input into more traditional machine learning models.

5.2. Spatio-temporal Analysis

The analysis regarding place names mentioned in Tweets over the flood events on Figure 4.4 (B) appears to correlate with past research that has suggested there is often a much larger proportion of geographic information present in Tweets during emergency events (Grace 2020). The slight up-tick in flood related Tweets mentioned prior to the flood warning day on 4.4 (A) also suggests that there may be Tweets that report flooding emergencies prior to official warnings, as found with past research (Perng et al. 2013; Martínez-Rojas, Pardo-Ferreira, and Rubio-Romero 2018). Further analysis would be required to determine whether particular Tweets do explicitly report flooding before it is reported formally. Given *flood related* Tweets appear even 7 days prior to events, they are not all likely predictive of flooding at that stage.

Further work may consider geo-coding place names from Tweets, to observe the spatial distribution of Tweets during each event (Lorini et al. 2019). However, many gazetteers, popular for geo-coding work, are likely to exclude place names that are more localised or are colloquial (Twaroch, Jones, and Abdelmoty 2008; Gao et al. 2017).

5.3. Conclusion

My dissertation demonstrates the use of a Twitter-based pre-trained transformer language model to classify Tweets relating to flood events as relevant or irrelevant. This model requires no *ad hoc* feature engineering or keyword selection, meaning outputs are less likely to demonstrate bias derived from this selection of features. However, as demonstrated, the performance appears to be relatively similar to classification through keyword matching, suggesting that use-cases for such models may be situation dependant.

Future work when considering deep learning for Tweet classification should consider the ambiguity that is often present when selecting binary labels. There is a balance between the use of multi-label classification, which in some cases may allow the model to group Tweets that contain a similar context, but could present problems if the categories become too diluted. Labelling ambiguity may be partially resolved by the use of multiple annotators (Beck et al. 2020), but given the informal nature of Tweets, this problem is likely to persist.

Recent work has considered the prominent issue with human annotated labelling accuracy in a variety of supervised machine learning methods, which appears on many major datasets (Northcutt, Athalye, and Mueller 2021)[1]. Future work may consider the edge cases that occur with low model confidence, and incorrect labels to determine whether there are clear labelling issues.

[1] See https://labelerrors.com/

Bibliography

Arnbjerg-Nielsen, K., P. Willems, J. Olsson, S. Beecham, A. Pathirana, I. Bülow Gregersen, H. Madsen, and V.-T.-V. Nguyen. 2013. "Impacts of Climate Change on Rainfall Extremes and Urban Drainage Systems: A Review." *Water Science and Technology* 68 (1): 16–28. https://doi.org/f44bgx.

Arthur, Rudy, Chris A. Boulton, Humphrey Shotton, and Hywel T. P. Williams. 2018. "Social Sensing of Floods in the UK." Edited by Guy J-P. Schumann. *PLOS ONE* 13 (1): e0189327. https://doi.org/10.1371/journal.pone.0189327.

Ashktorab, Zahra, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. "Tweedr: Mining Twitter to Inform," 5.

Atefeh, Farzindar, and Wael Khreich. 2015. "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence* 31 (1): 132–64. https://doi.org/f62chq.

Barbieri, Francesco, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification." *arXiv:2010.12421 [Cs]*, October. https://arxiv.org/abs/2010.12421.

Beck, Christin, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. "Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias," 14.

Brants, Thorsten, Francine Chen, and Ayman Farahat. 2003. "A System for New Event Detection," 8. https://doi.org/fcmb5q.

Brengarth, Lauren Bacon, and Edin Mujkic. 2016. "WEB 2.0: How Social Media Applications Leverage Nonprofit Responses During a Wildfire Crisis." *Computers in Human Behavior* 54 (January): 589–96. https://doi.org/gmkgqg.

Brouwer, Tom, Dirk Eilander, Arnejan van Loenen, Martijn J. Booij, Kathelijne M. Wijnberg, Jan S. Verkade, and Jurjen Wagemaker. 2017. "Probabilistic Flood Extent Estimates from Social Media Flood Observations." *Natural Hazards and Earth System Sciences* 17 (5): 735–47. https://doi.org/gcdh2v.

Caragea, Cornelia, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, et al. 2011. "Classifying Text Messages for the Haiti Earthquake," 10.

Caragea, Cornelia, Adrian Silvescu, and Andrea H Tapia. 2016. "Identifying Informative Messages in Disaster Events Using Convolutional Neural Networks," 8.

Carley, Kathleen M., Momin Malik, Peter M. Landwehr, Jürgen Pfeffer, and Michael Kowalchuck. 2016. "Crowd Sourcing Disaster Management: The Complex Nature of Twitter Usage in Padang Indonesia." *Safety Science* 90 (December): 48–61. https://doi.org/gc7q4j.

Castillo, Carlos. 2016. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. 2014. "Mining Social Media Data for Understanding Students' Learning Experiences." *IEEE Transactions on Learning Technologies* 7 (3): 246–59. https://doi.org/f6swvp.

de Bruijn, Jens A., Hans de Moel, Albrecht H. Weerts, Marleen C. de Ruiter, Erkan Basar, Dirk Eilander, and Jeroen C. J. H. Aerts. 2020. "Improving the Classification of Flood Tweets with Contextual Hydrological Information in a Multimodal Neural Network." *Computers & Geosciences* 140 (July): 104485. https://doi.org/gk8gzg.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [Cs]*, May. https://arxiv.org/abs/1810.04805.

Dou, Wenwen, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou. 2012. "LeadLine: Interactive Visual Analysis of Text Data Through Event Identification and Exploration." In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 93–102. https://doi.org/ggd95t.

European Severe Storms Laboratory (ESSL). n.d. "European Severe Weather Database." https://www.eswd.eu/.

Falcon, et al., WA. 2019. "PyTorch Lightning." *GitHub. Note: Https://Github.com/PyTorchLightning/Pytorch-Lightning* 3.

Forzieri, Giovanni, Alessandro Cescatti, Filipe Batista e Silva, and Luc Feyen. 2017. "Increasing Risk over Time of Weather-Related Hazards to the European Population: A Data-Driven Prognostic Study." *The Lancet Planetary Health* 1 (5): e200–208. https://doi.org/gfz74r.

Gao, Song, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. 2017. "Constructing Gazetteers from Volunteered Big Geo-Data Based on Hadoop." *Computers, Environment and Urban Systems* 61 (January): 172–86. https://doi.org/f9jhdk.

Ghafarian, Seyed Hossein, and Hadi Sadoghi Yazdi. 2020. "Identifying Crisis-Related Informative Tweets Using Learning on Distributions." *Information Processing & Management* 57 (2): 102145. https://doi.org/gj8br7.

Grace, Rob. 2020. "Toponym Usage in Social Media in Emergencies." *International Journal of Disaster Risk Reduction*, October, 101923. https://doi.org/ghjp44.

Hughes, Amanda L., Lise A. A. St. Denis, Leysia Palen, and Kenneth M. Anderson. 2014. "Online Public Communications by Police & Fire Services During the 2012 Hurricane Sandy." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1505–14. Toronto Ontario Canada: ACM. https://doi.org/gmf7qx.

Imran, Muhammad, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. "Extracting Information Nuggets from Disaster- Related Messages in Social Media," 10.

Imran, Muhammad, Ferda Ofli, Doina Caragea, and Antonio Torralba. 2020. "Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions." *Information Processing & Management* 57 (5): 102261. https://doi.org/gmf7s4.

Iyengar, Akshaya, Tim Finin, and Anupam Joshi. 2011. "Content-Based Prediction of Temporal Boundaries for Events in Twitter." In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011*

IEEE Third Int'l Conference on Social Computing, 186–91. Boston, MA, USA: IEEE. https://doi.org/fzz5ns.

Kim, Jooho, and Makarand Hastak. 2018. "Social Network Analysis: Characteristics of Online Social Networks After a Disaster." *International Journal of Information Management* 38 (1): 86–96. https://doi.org/gcqdv5.

Kron, Wolfgang, Petra Löw, and Zbigniew W. Kundzewicz. 2019. "Changes in Risk of Extreme Weather Events in Europe." *Environmental Science & Policy* 100 (October): 74–83. https://doi.org/gmhcpm.

Kryvasheyeu, Yury, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. 2016. "Rapid Assessment of Disaster Damage Using Social Media Activity." *Science Advances* 2 (3): e1500779. https://doi.org/gc5tfp.

Laylavi, Farhad, Abbas Rajabifard, and Mohsen Kalantari. 2016. "A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response." *ISPRS International Journal of Geo-Information* 5 (5): 56. https://doi.org/f8v96g.

Li, Hongmin, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. "Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach." *Journal of Contingencies and Crisis Management* 26 (1): 16–27. https://doi.org/gc35fc.

Lin, Xialing, Patric R. Spence, Timothy L. Sellnow, and Kenneth A. Lachlan. 2016. "Crisis Communication, Learning and Responding: Best Practices in Social Media." *Computers in Human Behavior* 65 (December): 601–5. https://doi.org/ggff7z.

Lorini, V., C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon. 2019. "Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach." *arXiv:1904.10876 [Cs]*, April. https://arxiv.org/abs/1904.10876.

Lu, X., L. Bengtsson, and P. Holme. 2012. "Predictability of Population Displacement After the 2010 Haiti Earthquake." *Proceedings of the National Academy of Sciences* 109 (29): 11576–81. https://doi.org/f2z47x.

Mani, Inderjeet, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. "SpatialML: Annotation Scheme, Resources, and Evaluation." *Language Resources and Evaluation* 44 (3): 263–80. https://doi.org/cnmp8m.

Martínez-Rojas, María, María del Carmen Pardo-Ferreira, and Juan Carlos Rubio-Romero. 2018. "Twitter as a Tool for the Management and Analysis of Emergency Situations: A Systematic Literature Review." *International Journal of Information Management* 43 (December): 196–208. https://doi.org/gfmbxd.

Mazzoleni, Maurizio. 2017. *Improving Flood Prediction Assimilating Uncertain Crowdsourced Data into Hydrologic and Hydraulic Models*.

Met Office. n.d. "Met Office WOW." https://wow.metoffice.gov.uk/.

Middleton, Stuart E., Lee Middleton, and Stefano Modafferi. 2014. "Real-Time Crisis Mapping of Natural Disasters Using Social Media." *IEEE Intelligent Systems* 29 (2): 9–17. https://doi.org/gfv7c6.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv Preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781.

Morstatter, Fred, Juergen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," 9.

Muller, C. L., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. R. Leigh. 2015. "Crowdsourcing for Climate and Atmospheric Sciences: Current Status and Future Potential." *International Journal of Climatology* 35 (11): 3185–3203. https://doi.org/f7qrps.

Nakayama, Hiroki, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. "Doccano: Text Annotation for Humans."

Nguyen, Dat, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. "Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks." *Proceedings of the International AAAI Conference on Web and Social Media* 11 (1): 632–35.

Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. 2021. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." *arXiv:2103.14749 [Cs, Stat]*, April. https://arxiv.org/abs/2103.14749.

Olteanu, Alexandra, Sarah Vieweg, and Carlos Castillo. 2015. "What to Expect When the Unexpected Happens: Social Media Communications Across Crises." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 994–1009. Vancouver BC Canada: ACM. https://doi.org/gmmfbh.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 8024–35. Curran Associates, Inc.

Pekar, Viktor, Jane Binner, Hossein Najafi, Chris Hale, and Vincent Schmidt. 2020. "Early Detection of Heterogeneous Disaster Events Using Social Media." *Journal of the Association for Information Science and Technology* 71 (1): 43–54. https://doi.org/ggwkh5.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. Doha, Qatar: Association for Computational Linguistics. https://doi.org/gfshwg.

Perng, Sung-Yueh, Monika Büscher, Lisa Wood, Ragnhild Halvorsrud, Michael Stiso, Leonardo Ramirez, and Amro Al-Akkad. 2013. "Peripheral Response: Microblogging During the 22/7/2011 Norway Attacks." *International Journal of Information Systems for Crisis Response and Management* 5 (1): 41–57. https://doi.org/gmgncq.

Pota, Marco, Mirko Ventura, Rosario Catelli, and Massimo Esposito. 2020. "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian." *Sensors* 21 (1): 133. https://doi.org/gmhdqv.

Reilly, Paul, and Stefania Vicari. 2021. "Organizational Hashtags During Times of Crisis: Analyzing the Broadcasting and Gatekeeping Dynamics of #PorteOuverte During the November 2015 Paris Terror Attacks." *Social Media + Society* 7 (1): 205630512199578. https://doi.org/gmdkxv.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," 10. https://doi.org/b6zm4b.

Saravanou, Antonia, George Valkanas, Dimitrios Gunopulos, and Gennady Andrienko. 2015. "Twitter Floods When It Rains: A Case Study of the UK Floods in Early 2014." In *Proceedings of the 24th International Conference on World Wide Web*, 1233–38. Florence Italy: ACM. https://doi.org/ghcxcv.

Schneider, Scott, and Pietra Check. 2010. "Read All about It: The Role of the Media in Improving Construction Safety and Health." *Journal of Safety Research*, Special Topic: Construction Safety, 41 (3): 283–87. https://doi.org/bh8jrq.

Simon, Tomer, Avishay Goldberg, and Bruria Adini. 2015. "Socializing in Emergencies—A Review of the Use of Social Media in Emergency Situations." *International Journal of Information Management* 35 (5): 609–19. https://doi.org/gmkgpb.

Spielhofer, Thomas, Reynold Greenlaw, Deborah Markham, and Anna Hahne. 2016. "Data Mining Twitter During the UK Floods: Investigating the Potential Use of Social Media in Emergency Management." In *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 1–6. Vienna, Austria: IEEE. https://doi.org/ggwjsv.

Statista. 2021. "Twitter: Most Users by Country." *Statista*. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." *arXiv:1703.01365 [Cs]*, June. https://arxiv.org/abs/1703.01365.

Tjong Kim Sang, Erik F., and Fien De Meulder. 2003. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–47. https://doi.org/d8qpkd.

Twaroch, Florian A., Christopher B. Jones, and Alia I. Abdelmoty. 2008. "Acquisition of a Vernacular Gazetteer from Web Sources." In *Proceedings of the First International Workshop on Location and the Web - LOCWEB '08*, 61–64. Beijing, China: ACM Press. https://doi.org/10.1145/1367798.1367808.

Umihara, Junko, and Mariko Nishikitani. 2013. "Emergent Use of Twitter in the 2011 Tohoku Earthquake." *Prehospital and Disaster Medicine* 28 (5): 434–40. https://doi.org/f5jv86.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *arXiv:1706.03762 [Cs]*, December. https://arxiv.org/abs/1706.03762.

Willems, Patrick, Peter Molnar, Thomas Einfalt, Karsten Arnbjerg-Nielsen, Christian Onof, Van-Thanh-Van Nguyen, and Paolo Burlando. 2012. "Rainfall in the Urban Context: Forecasting, Risk and Climate Change." *Atmospheric Research* 103 (January): 1–3. https://doi.org/bv2qqn.

Williams, S. A., M. M. Terras, and C. Warwick. 2013. "What Do People Study When They Study Twitter? Classifying Twitter Related Academic Papers." *Journal of Documentation* 69 (3): 384–410. https://doi.org/f42xw4.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." *arXiv:1910.03771 [Cs]*, July. https://arxiv.org/abs/1910.03771.