Comparing rule-based methods and pre-trained language models to classify flood related Tweets

Cillian Berragan*[1] and Alessia Calafiore†[1]

[1]University of Liverpool, Liverpool, L69 3BX

January 10, 2022

Summary

Social media presents a rich source of real-time information provided by individual users in emergency situations. However, due to its unstructured nature and high volume, it is challenging to extract key information from these continuous data streams. This paper considers the ability to identify relevant flood related Tweets from a Twitter corpus from past flood events, demonstrating the ability to capture this information from a real-time Twitter stream, when initial flood warnings are known. Tweets containing flood related information are identified using a deep neural classification model, and evaluated against a more commonly employed rule-based classification.

1 Introduction

Twitter presents large continuous feed of information regarding emergency events, contributed through individual users, as these events occur. 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 et al., 2019), a trend that is expected to continue (Forzieri et al., 2017), meaning there is an increasing demand to predict and understand how natural disasters develop. Tweets have proved useful in complementing and supporting emergency response in many cases, and often the first reports about emergencies on social media often precede those of mainstream media (Perng et al., 2013; Martínez-Rojas et al., 2018; Kim and Hastak, 2018; Laylavi et al., 2016). It is therefore important to be able to extract flood related Tweets, removing the noise that often comes with social media streams (Ashktorab et al., 2014).

Much of the past work that has used Twitter to study past emergency events has used keywords to

*C.Berragan@liverpool.ac.uk
†A.Calafiore@liverpool.ac.uk

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 et al., 2010; Spielhofer et al., 2016), for example a person may be in *'floods of tears'*. Finally, Tweets relevant to emergency events also do not necessarily contain an obvious keyword (*'Cars are floating down the street!'*), and therefore are unable to be detected. More recent work has considered the ability to use machine learning to classify Tweets into those relevant to emergency events, and those that are irrelevant (Imran et al., 2020; Arthur et al., 2018; Sakaki et al., 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 (SVMs) (Caragea et al., 2011; Sakaki et al., 2010), while more recent work has considered the emerging prevalence of neural networks in text-based classification (Caragea et al., 2016; de Bruijn et al., 2020; Nguyen et al., 2017). Traditional machine learning methods however rely on the use of feature engineering to determine model input, are unable to preserve word order, and have limited capability to use context, often over-fitting based on features selected (Caragea et al., 2016). Work with neural networks has shown that given pre-trained word embeddings, they have the capability to outperform these methods (Ghafarian and Yazdi, 2020; Caragea et al., 2016).

2 Methodology

2.1 Data Collection

2.1.1 Flood Data

A historical dataset containing all *Severe Flood Warnings*, *Flood Warnings*, and *Flood Alerts* issued by the UK flood warning system is available through the UK Government under the Open Government Licence. This data was linked with flood zones from the Environment Agency Real Time Flood-Monitoring API. To reduce the volume of flood events being considered, only *Severe Flood Warnings* occurring after 2010 were selected, leaving a total of 314 individual *Severe Flood Warning* events.

2.1.2 Tweets

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

Geographic information associated with every Tweet was required due to the decision to use bounding boxes to pre-emptively filter Tweets in areas not subject to flooding. 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 et al., 2014;

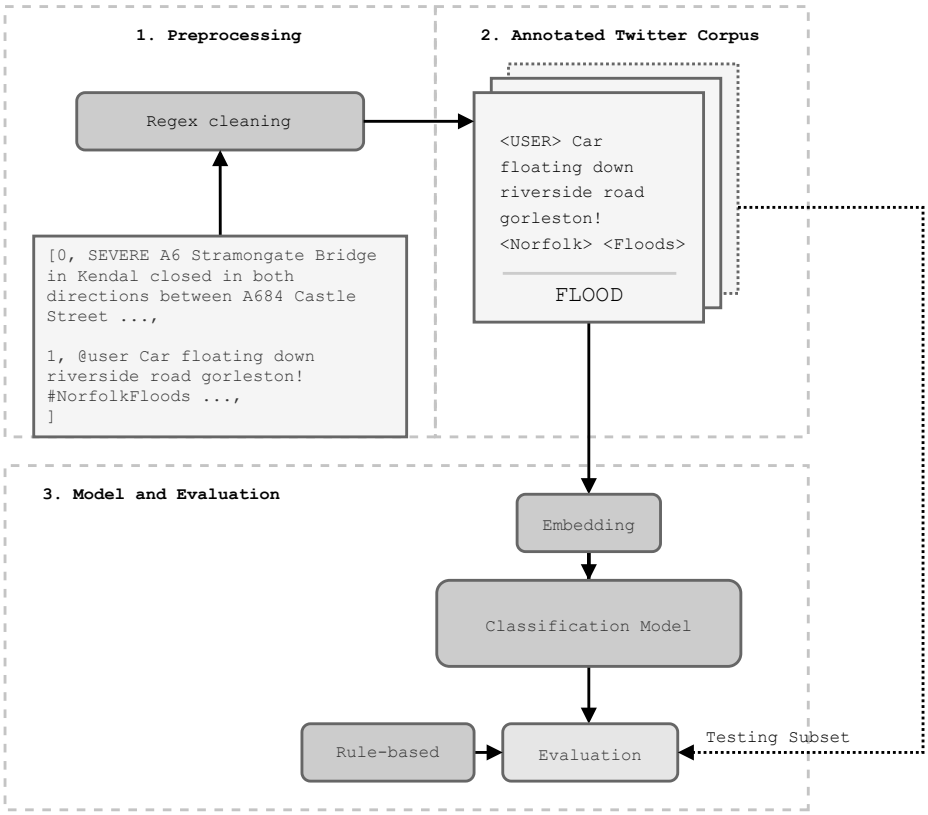

Figure 1: Overview of the model processing pipeline.

Carley et al., 2016; Morstatter 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 of 286 Tweets per flood warning.

2.2 Classification

Figure 1 gives an overview of the classification pipeline used, each Tweet was first pre-processed to normalise usernames and web addresses, and hashtags were parsed to extract words (Pota et al., 2020) (Stage 1). A random subset of ˜2,500 Tweets were then taken from the overall corpus and manually annotated to train the classification model using Doccano (Nakayama et al., 2018), with 10% used for model validation (Stage 2). The validation subset was then used to evaluate model performance in relation to the simple rule-based approach (Stage 3).

The model builds on the established NLP task of sequence classification, taking token sequences ($\mathbf{x} = \{x_0, x_1 \dots x_n\}$), and predicting a single label (y). A pre-trained transformer language model was taken as a base, using the RoBERTa architecture, pre-trained using a corpus of 58 million

Tweets (Barbieri et al., 2020)[1].

To construct a rule-based approach for evaluation against this model, every Tweet retrieved that included a selection of keywords were labelled as being flood related (*FLOOD*), while 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*[2]

For comparative evaluation, the F_1 metric was used, which takes the harmonic mean of the precision and recall, meaning class imbalance is accounted for.

To qualitatively assess the performance of the transformer model, *attributions*[3] for each word in a few selected Tweets were visualised to identify the ability of the model to capture information relevant to flood events, without having to explicitly be fed in keywords (Sundararajan et al., 2017).

3 Results

3.1 Comparison of classification methods

Overall the classification model out-performed the rule-based method on the validation subset, achieving an F_1 score of 0.938, compared with 0.803 for the rule-base approach. The primary reason for this difference in F_1 score is a lower recall for the rule-based model (0.873 compared with 0.952), indicating a higher number of false-negatives.

Figure 2 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 2 (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 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 through feature engineering or keywords.

In the second example on Figure 2 (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* meaning 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 relevant to floods, in addition to *apples* and *ipad pro*, both of which likely appear relatively frequently on Twitter, but rarely in flood related contexts.

The final two sub-figures give examples where the model gives incorrect classifications, but the rule-based method does not. Figure 2 (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

[1]Available on the Huggingface Model Hub (Wolf et al., 2020)

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

[3]`https://github.com/cdpierse/transformers-interpret`

Figure 2: Attribution levels for selected Tweets classified by the transformer model. Attribution label indicates the human annotated label, predicted label shows assigned label with confidence values. Positive attributions dictate the importance of a feature in the given label prediction.

not in reference to a flooding event. This example reflects a potential 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 2 (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.

4 Discussion

While the transformer-based classification model outperforms a rule-based approach, they present different benefits and costs. Supervised classification through a neural network relies heavily on a suitable amount of high quality labelled data, which presents an initial time-cost. Keyword selection is comparatively straightforward, and does not rely on a pre-existing corpus of relevant text. The training and inference for the transformer model also costs both time and resources, while keyword selection may be applied directly during the extraction of Tweets through the Twitter API.

Keywords however are inherently subjective, as demonstrated by past work which found varying selections of keywords to be appropriate the classification of flood related Tweets (Spielhofer et al., 2016; Arthur et al., 2018; Saravanou et al., 2015). Constructing a labelled corpus a broad binary classification of relevant and irrelevant Tweets to train a supervised model is less subjective, as the model itself may use the context provided through the training data to independently learn how to approach the classifications. As classification is not limited to specific keywords, relevant Tweets therefore include a broad range of flood related information that would not typically be captured (See Figure 2 (A)).

The complexity of the transformer architecture itself also presents improvements over past machine learning methods, as word order is preserved, and the pre-trained word embeddings mean no *ad hoc* feature engineering is required, which may have contributed to some bias and over-fitting in past work (Caragea et al., 2016). The semantic context captured by the model is notable on Figure 2 (B), which 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.

References

Arthur, R., Boulton, C. A., Shotton, H., and Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLOS ONE*, 13(1):e0189327.

Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). Tweedr: Mining Twitter to Inform. page 5.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv:2010.12421 [cs]*.

Brouwer, T., Eilander, D., van Loenen, A., Booij, M. J., Wijnberg, K. M., Verkade, J. S., and Wagemaker, J. (2017). Probabilistic flood extent estimates from social media flood observations. *Natural Hazards and Earth System Sciences*, 17(5):735–747.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J., and Yen, J. (2011). Classifying Text Messages for the Haiti Earthquake. page 10.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). Identifying Informative Messages in Disaster Events using Convolutional Neural Networks. page 8.

Carley, K. M., Malik, M., Landwehr, P. M., Pfeffer, J., and Kowalchuck, M. (2016). Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. *Safety Science*, 90:48–61.

de Bruijn, J. A., de Moel, H., Weerts, A. H., de Ruiter, M. C., Basar, E., Eilander, D., and Aerts, J. C. (2020). Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, 140:104485.

Forzieri, G., Cescatti, A., e Silva, F. B., and Feyen, L. (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–e208.

Ghafarian, S. H. and Yazdi, H. S. (2020). Identifying crisis-related informative tweets using learning on distributions. *Information Processing & Management*, 57(2):102145.

Hughes, A. L., St. Denis, L. A. A., Palen, L., and Anderson, K. M. (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*, pages 1505–1514, Toronto Ontario Canada. ACM.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). Extracting Information Nuggets from Disaster- Related Messages in Social Media. page 10.

Imran, M., Ofli, F., Caragea, D., and Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Information Processing & Management*, 57(5):102261.

Kim, J. and Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96.

Kron, W., Löw, P., and Kundzewicz, Z. W. (2019). Changes in risk of extreme weather events in Europe. *Environmental Science & Policy*, 100:74–83.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., and Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3):e1500779.

Laylavi, F., Rajabifard, A., and Kalantari, M. (2016). A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response. *ISPRS International Journal of Geo-Information*, 5(5):56.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Martínez-Rojas, M., Pardo-Ferreira, M. d. C., and Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, 43:196–208.

Middleton, S. E., Middleton, L., and Modafferi, S. (2014). Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems*, 29(2):9–17.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the Sample Good Enough? comparing Data from Twitter 's Streaming API with Twitter's Firehose. page 9.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). Doccano: Text annotation for humans.

Nguyen, D., Mannai, K. A. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (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):632–635.

Perng, S.-Y., Büscher, M., Wood, L., Halvorsrud, R., Stiso, M., Ramirez, L., and Al-Akkad, A. (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.

Pota, M., Ventura, M., Catelli, R., and Esposito, M. (2020). An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. *Sensors*, 21(1):133.

Reilly, P. and Vicari, S. (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.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. page 10.

Saravanou, A., Valkanas, G., Gunopulos, D., and Andrienko, G. (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*, pages 1233–1238, Florence Italy. ACM.

Spielhofer, T., Greenlaw, R., Markham, D., and Hahne, A. (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)*, pages 1–6, Vienna, Austria. IEEE.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*.