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
{c.berragan,A.Calafiore}@liverpool.ac.uk

Abstract. 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 considered to contain flood related information are identified using a deep neural classification model, and evaluated against a more commonly employed rule-based classification.

Keywords: natural language processing · information extraction · twitter · floods

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 [11,14], Paris terror attacks in 2015 [27], and UK flooding events [29,4].

Extreme weather events have become increasingly common [15], a trend that is expected to continue into the future [9], 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 [25,19,14,17]. It is therefore important to be able to extract flood related Tweets, removing the noise that often comes with social media streams [2].

Much of the past work that has used Twitter to study past emergency events has used keywords to identify relevant Tweets [16,4,21]. 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 [28,30], 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 [13,1,28,18]. These studies have utilised a variety of methods, building from classical approaches like Naïve Bayes classification [12,18] and Support Vector Machines (SVMs) [5,28], while more recent work has considered the emerging prevalence of neural networks in text-based classification [6,8,24]. 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 [6]. Work with neural networks has shown that given pre-trained word embeddings, they have the capability to outperform these methods [10,6].

This work considers the retrospective classification of a selection of Tweets from past flooding events in the United Kingdom into relevant Tweets and irrelevant, evaluating the effectiveness of a deep neural classification model against a keyword based approach. This work aims to demonstrate the benefits and costs of the use of new sophisticated methods in natural language processing for this task, while evaluation methods account for over-fitting by selecting a train, validation and test data split. The model utilises a modern neural network architecture called a transformer, which was pre-trained on a large corpus of Tweets, forming what is known as a language model. These models have been demonstrated to outperform past neural network methods by utilising the pre-learned embedded information captured during pre-training, while the architecture itself allows for semantic context to be more heavily utilised during training [32].

This method presents the first-stage for an approach that considers the ability to capture information automatically relevant to flood events, outlining the initial retrieval of Tweets, and filtering to remove Tweets unrelated to floods. Further work is expected to build on this, allowing for information extraction from the relevant Tweets to inform first responders, providing more fine-grained information based on the first-hand experience of individuals like specific property damage, or missing persons, allowing social media to complement existing methods used during flood events [22].

2 Methodology

2.1 Data Collection

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.

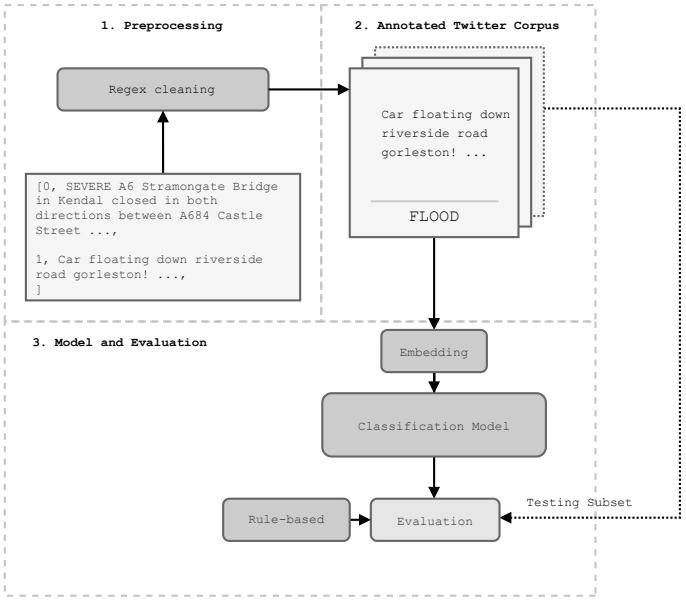

Fig. 1. Overview of the model processing pipeline.

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
- **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 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 [20,7,21]. 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 [26] (Stage 1). A random subset of 2,000 Tweets were then taken from the overall corpus and manually annotated to train the classification model using Doccano [23], with 10% used for model validation (Stage 2).

A separate subset of 504 Tweets were also manually labelled 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, trained using a corpus of 58 million Tweets [3][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:

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

To 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 [31].

3 Results

3.1 Comparison of classification methods

The test corpus contained 151 Tweets that were flood related (30%) and 352 Tweets that were not flood related (70%). Overall the classification model outperformed the rule-based method, achieving an F1 score of 0.884, compared with 0.843. The primary reason for this difference in F_1 score is a lower *specificity* for the rule-based model (0.960 compared with 0.994), indicating a higher number of false-positives.

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.

[1] Available on the Huggingface Model Hub [33]
[2] including words that share a stem, e.g. *flooding*, *raining*
[3] https://github.com/cdpierse/transformers-interpret

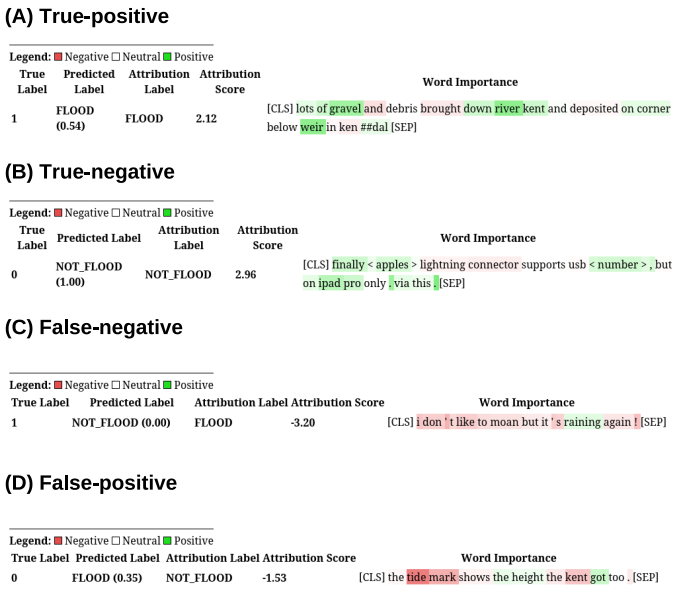

Fig. 2. Attribution levels for selected Tweets classified by the transformer model. Attribution label indicates the human annotated label, predicted label shows assigned label with confidence values. Positive attributions dictate the importance of a feature in the given label prediction.

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 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 [30,1,29]. 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 [6]. 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.

5 Conclusion

This work 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. As demonstrated, the performance appears to be relatively similar to classification through keyword matching, but results may be improved through the use of additional training data.

Further work may consider utilising the entire corpus of 89,864 Tweets extracted relating to UK flood events, instead of the small subset used in this paper. A model trained using the full corpus would be expected to more suitably handle the broader semantic context used when Tweeting in relation to flood events, meaning for future use it may further the gap between itself and the rule-based methods. Additionally, a corpus of relevant Tweets excluding noise presents the opportunity capture additional information, for example the identification of locations through geoparsing, or multi-class classification to identify

Tweets specifically relating to flood related events such as property damage or missing persons.

References

1. Arthur, R., Boulton, C.A., Shotton, H., Williams, H.T.P.: Social sensing of floods in the UK. PLOS ONE **13**(1), e0189327 (Jan 2018). https://doi.org/10.1371/journal.pone.0189327
2. Ashktorab, Z., Brown, C., Nandi, M., Culotta, A.: Tweedr: Mining Twitter to Inform p. 5 (2014)
3. Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L.: TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. arXiv:2010.12421 [cs] (Oct 2020)
4. Brouwer, T., Eilander, D., van Loenen, A., Booij, M.J., Wijnberg, K.M., Verkade, J.S., Wagemaker, J.: Probabilistic flood extent estimates from social media flood observations. Natural Hazards and Earth System Sciences **17**(5), 735–747 (May 2017). https://doi.org/10/gcdh2v
5. Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.W., Mitra, P., Wu, D., Tapia, A.H., Giles, L., Jansen, B.J., Yen, J.: Classifying Text Messages for the Haiti Earthquake p. 10 (2011)
6. Caragea, C., Silvescu, A., Tapia, A.H.: Identifying Informative Messages in Disaster Events using Convolutional Neural Networks p. 8 (2016)
7. Carley, K.M., Malik, M., Landwehr, P.M., Pfeffer, J., Kowalchuck, M.: Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. Safety Science **90**, 48–61 (Dec 2016). https://doi.org/10/gc7q4j
8. de Bruijn, J.A., de Moel, H., Weerts, A.H., de Ruiter, M.C., Basar, E., Eilander, D., Aerts, J.C.: Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. Computers & Geosciences **140**, 104485 (Jul 2020). https://doi.org/10/gk8gzg
9. Forzieri, G., Cescatti, A., e Silva, F.B., Feyen, L.: 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 (Aug 2017). https://doi.org/10/gfz74r
10. Ghafarian, S.H., Yazdi, H.S.: Identifying crisis-related informative tweets using learning on distributions. Information Processing & Management **57**(2), 102145 (Mar 2020). https://doi.org/10/gj8br7
11. Hughes, A.L., St. Denis, L.A.A., Palen, L., Anderson, K.M.: Online public communications by police & fire services during the 2012 Hurricane Sandy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1505–1514. ACM, Toronto Ontario Canada (Apr 2014). https://doi.org/10/gmf7qx
12. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Extracting Information Nuggets from Disaster- Related Messages in Social Media p. 10 (2013)
13. Imran, M., Ofli, F., Caragea, D., Torralba, A.: Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. Information Processing & Management **57**(5), 102261 (Sep 2020). https://doi.org/10/gmf7s4
14. Kim, J., Hastak, M.: Social network analysis: Characteristics of online social networks after a disaster. International Journal of Information Management **38**(1), 86–96 (Feb 2018). https://doi.org/10/gcqdv5

15. Kron, W., Löw, P., Kundzewicz, Z.W.: Changes in risk of extreme weather events in Europe. Environmental Science & Policy **100**, 74–83 (Oct 2019). https://doi.org/10/gmhcpm
16. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M.: Rapid assessment of disaster damage using social media activity. Science Advances **2**(3), e1500779 (Mar 2016). https://doi.org/10/gc5tfp
17. Laylavi, F., Rajabifard, A., Kalantari, M.: A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response. ISPRS International Journal of Geo-Information **5**(5), 56 (May 2016). https://doi.org/10/f8v96g
18. Li, H., Caragea, D., Caragea, C., Herndon, N.: Disaster response aided by tweet classification with a domain adaptation approach. Journal of Contingencies and Crisis Management **26**(1), 16–27 (2018). https://doi.org/10/gc35fc
19. Martínez-Rojas, M., Pardo-Ferreira, M.d.C., Rubio-Romero, J.C.: Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. International Journal of Information Management **43**, 196–208 (Dec 2018). https://doi.org/10/gfmbxd
20. Middleton, S.E., Middleton, L., Modafferi, S.: Real-Time Crisis Mapping of Natural Disasters Using Social Media. IEEE Intelligent Systems **29**(2), 9–17 (Mar 2014). https://doi.org/10/gfv7c6
21. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the Sample Good Enough? comparing Data from Twitter 's Streaming API with Twitter's Firehose p. 9 (2013)
22. Muller, C.L., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., Overeem, A., Leigh, R.R.: Crowdsourcing for climate and atmospheric sciences: Current status and future potential. International Journal of Climatology **35**(11), 3185–3203 (2015). https://doi.org/10/f7qrps
23. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: Doccano: Text annotation for humans (2018)
24. Nguyen, D., Mannai, K.A.A., Joty, S., Sajjad, H., Imran, M., Mitra, P.: 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 (May 2017)
25. Perng, S.Y., Büscher, M., Wood, L., Halvorsrud, R., Stiso, M., Ramirez, L., Al-Akkad, A.: Peripheral Response: Microblogging During the 22/7/2011 Norway Attacks. International Journal of Information Systems for Crisis Response and Management **5**(1), 41–57 (Jan 2013). https://doi.org/10/gmgncq
26. Pota, M., Ventura, M., Catelli, R., Esposito, M.: An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. Sensors **21**(1), 133 (Dec 2020). https://doi.org/10/gmhdqv
27. Reilly, P., Vicari, S.: 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 (Jan 2021). https://doi.org/10/gmdkxv
28. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors p. 10 (2010). https://doi.org/10/b6zm4b
29. Saravanou, A., Valkanas, G., Gunopulos, D., Andrienko, G.: 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. pp. 1233–1238. ACM, Florence Italy (May 2015). https://doi.org/10/ghcxcv
30. Spielhofer, T., Greenlaw, R., Markham, D., Hahne, A.: 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). pp. 1–6. IEEE, Vienna, Austria (Dec 2016). https://doi.org/10/ggwjsv
31. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs] (Jun 2017)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762 [cs] (Dec 2017)
33. 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., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs] (Jul 2020)