EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Sequence Classification of Tweets with Transfer Learning via BERT in the Field of Disaster Management

Sumera Naaz[1,†], Zain Ul Abedin[1,†] and Danish Raza Rizvi[1,*]

[1]Department of Computer Engineering, Jamia Millia Islamia, New Delhi, 110025 India

Abstract

Twitter is extensively used as an information-sharing platform during any kind of emergency like disasters etc. People tweet useful information about disaster-related events such as evacuations, volunteer need, help, warnings etc. This data is sometimes very useful for rescue teams, NGOs, military and various other government and private organisations who are tasked with responsibilities to save lives and provide volunteers. This data can also be used to analyze disaster behaviour. In this paper, we have collected labelled tweets from crisisLexT26 and crisisNLP and classified them into seven labels on the basis of information provided by them. The data was heavily skewed. So to improve the accuracy of classifiers, we have applied various techniques as a result of which we have created two datasets (Imbalanced and Balanced). We have compared the performance of various BERT-based models on these two datasets. For sequence classification, a balanced dataset performs better than an imbalanced dataset. We can improve accuracy of classifiers to great extent by adopting good data preprocessing and data splitting techniques.

Received on 21 June 2020; accepted on 18 March 2021; published on 23 March 2021

Keywords: BERT (Bidirectional Encoder Representation from Transformers), Tweet classification, Balanced Dataset, Imbalanced Dataset, Disaster Management, Natural Language Processing.

doi:10.4108/eai.23-3-2021.169071

1. Introduction

The popularity of social media is increasing rapidly due to which massive volumes of data is generated each day. This massive volume of data provided great opportunities and challenges for natural language processing[1]. Although there is a huge availability of social media data, there is quite a limitation in making sense of this data because of its high velocity, veracity and large volume [2]. So this huge availability and complications of data make it even more prone to research and exploration.

The majority of people choose Twitter when choosing a social media outlet for reliable scientific information and news [3]. Microblogging and social networking sites like Twitter play important role in spreading information during disasters [4]. Recent research in this area has affirmed the potential use of such social media data for various disaster response tasks [2].

Whenever a disaster occurs, there is a shortage of time because the safety of people is in question. So there is a need to act as quickly as possible [5].Different types of information are shared in real-time by victims; by people who wish to help these affected people or by people who need any kind of help [6]. Twitter has helped a lot in spreading news of damages, donation needs, volunteering which also include videos and photos [4].

It is also difficult to identify relevant information about disasters [7], thus it becomes more difficult for disaster-affected communities and rescue teams to act quickly [4].

Recent studies have shown the relevance of social media messages and how these messages contain information that can be used effectively during disasters [4][8][9][10]. These social media messages can be processed by various NLP techniques like

*Corresponding author. Email: Drizvi@jmi.ac.in
†These authors contributed equally

automatic summarization, information classification, named-entity recognition, information extraction etc [6][11]. But most of this data is brief, informal, noisy and contains typographical errors etc [12] which affects the accuracy of the model. Also, real-world data has a severe problem of imbalanced classes. Some classes have a fewer number of instances than other classes. This problem seriously affects the performance of the classifier.

We have collected twitter data from various sources and applied various data preprocessing techniques to make it suitable for processing. We have made two datasets, balanced and imbalanced from the collected tweets. Balanced dataset has an equal distribution of tweets of each label while an imbalanced dataset has an unequal distribution of tweets. We then compared the performance of these datasets on four BERT based models- default BERT, BERT+NL (BERT with non-linear layers), BERT+LSTM (BERT with LSTM) and BERT+CNN2 (BERT with two convolutional layers).

Acknowledgement. The code in this paper is adapted from the Guoqin Ma paper [13].

1.1. Related work

1. Imran, Muhammad, et al. [4] have developed a system that can filter messages that do not contribute to situational awareness. They then classified these filtered relevant messages into labels like caution and advice, casualties and damage, donation of money, goods or services etc.
2. In Nair, Meera R. et al. paper, Twitter messages have been classified using keyword analysis and a comparative study of three machine learning algorithms such as Random Forests, Decision tree and Naive Bayes is carried out. The comparison of all three algorithms is done with the help of weka, an analytical tool. This paper also focuses on identifying the most influential users of Chennai flood [14].
3. Starbird et. al has collected Tweets posted during Red River Flood that occurred in Red River valley in central North America using the keyword redriver. They then categorized these tweets into labels like hopeful, humour, support and fear [15].
4. Case study of Thai floods that occurred in 2011 collected tweets using keyword thaiflood. These tweets are then classified into five categories based on information provided by them. These categories were requests for assistance, announcements for support, Situational Awareness, requests for information and others. Along with this, they also identified the influential users related to Thai Flood, by scrutinizing the sources of the tweets. Most of the top users were from government or non-government organizations who were somehow related to the disaster [16].
5. Clarkson, Kyle, et al. focuses on geolocation inference of twitter users by taking reference of discrete sets of some geographical phenomena(for ex- solar eclipse). They applied this unique model to twitter's data gathered during the solar eclipse of 2017. They decided on the basis of some parameters if a particular feature can be used to decide that a user is viewing the eclipse or not [1].
6. Mendhe, Chetan Harichandra, et al. developed a platform for big data analysis. The platform supports various filters for data and contains a big collection of social media data. They offer a convenient method of collecting and hosting large data sets, implementing state-of-the-art algorithms for preprocessing, ranging from removal operations (e.g., of repeated tweets) to transformations (e.g., of abbreviations, acronyms, and emoticons into fully formed words), and making use of collective intelligence to annotate large collection of tweets. For annotation of large set of tweets, they have combined social media data collection with crowd-sourcing by using Amazon Mechanical Turk to label Twitter data [3].
7. Praznik, Logan, et al. focuses on link prediction of hashtag graphs. They showed how different hashtags can be linked with each other and hence, can belong to the same topic. This can also be helpful for tracking the development of a topic over the time and help in prediction of future course of topic. They mapped twitter data in terms of hashtag graphs, where vertices correspond to hashtags, and edges correspond to co-occurrences of hashtags within the same distinct tweet. Also, the weight of vertex in hashtag graphs corresponds to the number of tweets a hashtag has occurred in, and edges can be weighted with the number of tweets both hashtags have co-occurred in [17].
8. TextAttack is a Python library and a system for executing or building ill-disposed attacks against NLP models. This is profoundly useful in the assessment of the attack strategies and the NLP model's strength. Improving the model performance is one of the most crucial tasks and TextAttack is working in the betterment of the model's performance. TextAttack is quite flexible as it provides the option of customization in the formation of attacks. The four components from which TextAttack builds attacks are : a goal function, a set of constraints, a transformation, and a search method. The attacks can be reused for data augmentation and adversarial training [18].

1.2. Novelty

The dataset we have collected has an imbalanced distribution of tweets among different labels. This imbalanced nature of data causes a lower accuracy of classifiers. So we have applied various techniques for better performance and created two datasets(D1 and D2). The two main tasks of this paper are **1) Apply various data preprocessing techniques to improve the accuracy of classifier, 2) Compare the performance of various models on imbalanced and balanced datasets.** We have developed several BERT-based models and compared their performance. We chose BERT because it has achieved state-of-art performance in many NLP tasks.

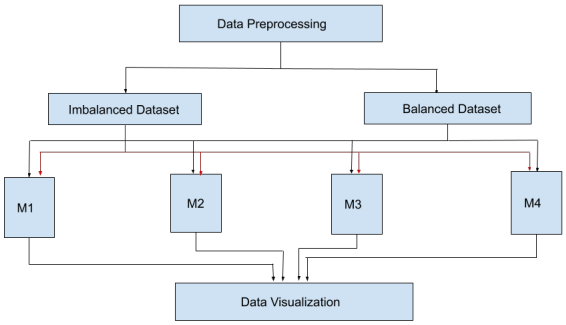


Figure 1. Workflow (M1, M2, M3 and M4 are four different BERT based models)

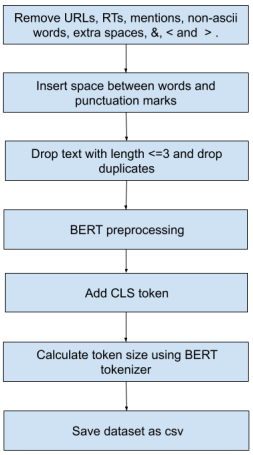


Figure 2. Flow Chart of data preprocessing

2. Approach

The flow chart for the approach is shown in (Figure 1).

2.1. Text Preprocessing

Tweets are converted into lowercase. User mentions do not convey any information, so they are removed. Non-Ascii characters and URLs are also removed. As we are using BERT in all models, an additional [CLS] token is also inserted at the beginning of each tweet (Figure 2). We have not removed stopwords for fluency purposes [13].

2.2. Data Preparation

The combined dataset we have is imbalanced. A dataset is said to be imbalanced if at least one of the classes has significantly fewer annotated instances than the others. The class imbalance problem has been known to hinder the learning performance of classification algorithms [6]. So we have applied several techniques to improve the accuracy of classifiers. We have compared the performance of all models on two datasets.

1. We split this imbalanced dataset(D1) into train, validation and test data in such a way that there is an equal distribution of tweets of each label (imbalanced dataset + equal distribution) [19].
2. We first converted the imbalanced dataset into balanced dataset(D2) and then split into train, validation and test data such that there is the equal distribution of tweets of each label (balanced dataset + equal distribution).

2.3. BERT-based models

BERT has two models, BERT base and BERT large. BERT base has 110 million parameters while BERT large has 345 million parameters. The BASE model is used to compare the performance of different architecture and the LARGE model produces state-of-the-art results as stated in BERT research paper. In all the models, we are using BERT base uncased model which consists of 12 layers, 768 hidden layers and 12 heads [20].

Default BERT. For sequence classification, we have used default BERT. The last layer is the softmax layer and softmax function is a squashing function. In this approach, we adjust the hyperparameters of the Pre-trained BERT model very precisely [13] (Figure 3a).

BERT with Non-Linear Layer. Three fully connected layers are stacked on the BERT model. The activation function used in the first two layers is a leaky rectified linear unit (negative slope=0.01) and softmax is performed by the third layer. In this approach also, we adjust the hyperparameters of the pre-trained BERT model very precisely [13] (Figure 3b).

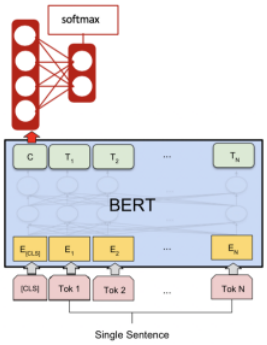
BERT with Long-Short Term Memory. This is a feature-based approach. This model is developed by stacking a bidirectional LSTM on default BERT model. The input to the bidirectional LSTM is provided by the final hidden state of BERT. The last fully-connected layer performs softmax. The bidirectional LSTM is followed by a softmax layer [13] (Figure 3c).

BERT with Convolutional Neural Network. This is a feature-based approach.This model is developed by stacking a CNN model on default BERT model. This model is developed by two convolutional layers followed by a softmax layer. The number of in-channels and out-channels for the first convolutional layer are 12 and 12 respectively. The number of in-channels and out-channels for the second layer are 12 and 192 respectively. The output from the second convolutional layer is fed to the softmax layer (Figure 3d).

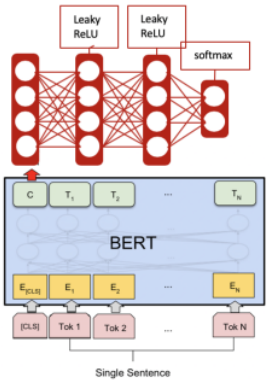
3. Experiment

3.1. Data

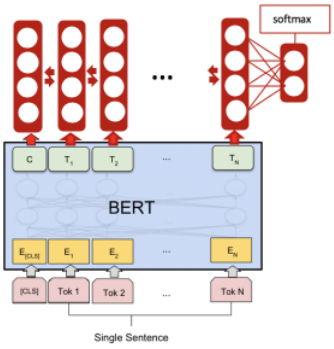
We have collected various small aforementioned datasets from crisisNLP and crisisLexT26 and compiled them into a single large dataset. This dataset basically contains tweets which are posted during various types of disasters across various parts of the world. These tweets are distributed into seven labels. We have followed taxonomy as given in Guoqin Ma paper [13]. The labels are -not related or not informative -other useful information -donations and volunteering -affected individuals -sympathy and emotional support -infrastructure and utilities damage -caution and advice.



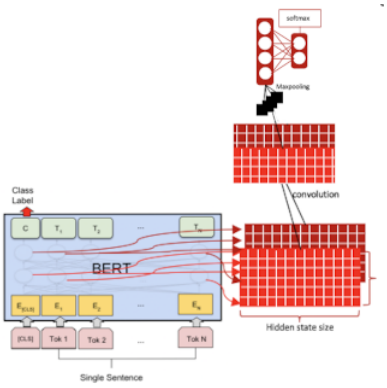
(a) Default BERT



(b) BERT+NL



(c) BERT+LSTM



(d) BERT+CNN2

Figure 3. BERT-based model diagrams are taken and adapted from Guoqin Ma paper [13]

Tweets are categorized on the basis of their information types. For example [6]:

not related or not informative. information and questions which are either not related to disaster or are out of this scope to categorize.

other useful information. information and questions related to disaster.

donations and volunteering. regarding the donation of food, clothes, medicines and other basic stuff. People willing to volunteer to provide help.

affected individuals. information regarding injured or dead people and other victims of the disaster.

sympathy and emotional support. information regarding prayers and well wishes.

infrastructure and utilities damage. information related to damaged buildings, places, things and services.

caution and advice. information regarding warnings, tips and advice by concerned authorities and people.

This dataset has a highly skewed distribution of labels. This imbalanced distribution causes lower accuracy of the classifier. So we have applied various techniques to improve the performance of models. We have trained all the models on two datasets (D1 and D2). The insight for these datasets is shown in Table 1 and Table 2 respectively.

Table 1. Insights of D1

Labels	Count
not related or not informative	25785
other useful information	18877
donations and volunteering	11315
affected individuals	10587
sympathy and emotional support	6100
infrastructure and utilities damage	5468
caution and advice	4301

Table 2. Insights of D2

Labels	Count
not related or not informative	10000
other useful information	10000
donations and volunteering	10000
affected individuals	10000
sympathy and emotional support	10000
infrastructure and utilities damage	10000
caution and advice	10000

3.2. Evaluation Method

Multiple metrics are calculated so that the model is evaluated properly. Accuracy, precision, recall, F1-score and Matthews

correlation coefficient are determined during the evaluation. Macro precision, macro recall, F1-score, Matthews correlation coefficient and accuracy are determined for every model respectively, while recall, precision and F1-score score are determined for every class[13].

3.3. Experimental Details

For both D1 and D2, the train, test and validation set split percentage is the same. The train set is 85%, the test set is 10% and the validation set is 5%. We shuffle the samples in the train set between the epochs. The loss function used is the Cross-Entropy loss function. There are 7 seven labels so we use multiclass classification variation of the loss function.

$$\sum_{c=1}^{M} y_{o,c} log(p_{o,c})$$

Figure 4. M - number of classes, y - binary indicator (0 or 1), p - predicted probability observation of o is of class c

Max epoch is set to 100. For the model training, it uses Adam optimizer. Batch size is 32. Learning rate for BERT parameters is 0.00002 and the learning rate for non-BERT parameters is 0.001. The learning rate decay by 50% when patience hit 5 [13].

3.4. Results

Macro Precision(D1). Default BERT > BERT+NL > BERT+CNN2 > BERT+LSTM

Macro Precision(D2). BERT+LSTM > BERT+CNN2 > Default BERT > BERT+NL

Macro Recall(D1). BERT+CNN2 > Default BERT > BERT+LSTM > BERT+NL

Macro Recall(D2). BERT+CNN2 > BERT+LSTM > Default BERT > BERT+NL

Macro F1-score(D1). Default BERT > BERT+NL > BERT+CNN2 > BERT+LSTM

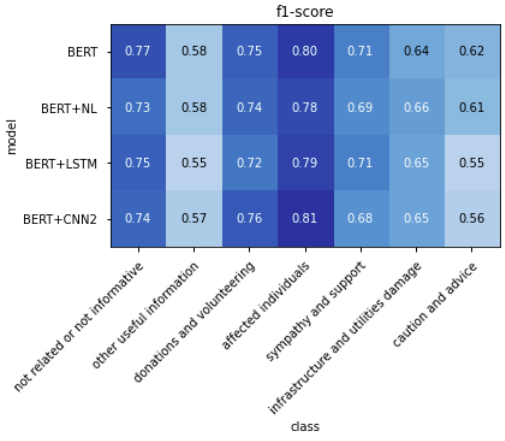
Macro F1-score(D2). BERT+LSTM > BERT+CNN2 > Default BERT > BERT+NL

The evaluation metrics for both datasets on all BERT-based models is shown in table 3. For D1, Default BERT has performed best with an accuracy of 71% whereas for D2, BERT+CNN2 and BERT+LSTM has performed best with an accuracy of 72%.The models in general perform better when trained and tested with balanced dataset.

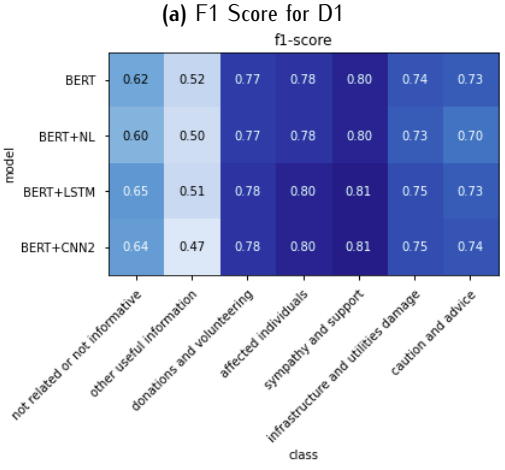
The heatmaps of F1 score for D1 and D2 are shown in Figure 5. The heatmaps of confusion matrix for all BERT-based models on D1 as well as D2 are shown in Figure 6 and Figure 7 respectively. From the confusion matrix of the test data for all the models, misclassification across the labels can be observed. Some of the reasons for misclassifications are ambiguity in the context of the tweet, the presence of special characters like emoji, etc.

Table 3. summarize the evaluation metrics for D1 and D2 on all BERT-based models. {Acc(Accuracy), MCC(Matthews Correlation Coefficient), MP(Macro Precision), MR(Macro Recall), M-F1(Macro F1)}.

Model	Acc.	MCC	MP	MR	M-F1
Dataset-1					
Default BERT	0.71	0.65	67.14	74.14	69.57
BERT+NL	0.69	0.62	66.29	72.29	68.43
BERT+LSTM	0.69	0.63	65.57	73.43	67.43
BERT+CNN2	0.70	0.64	65.86	74.86	68.14
Dataset-2					
Default BERT	0.71	0.66	71.29	71.00	70.86
BERT+NL	0.69	0.65	71.00	69.43	69.71
BERT+LSTM	0.72	0.67	72.14	72.14	71.86
BERT+CNN2	0.72	0.67	71.43	72.29	71.29

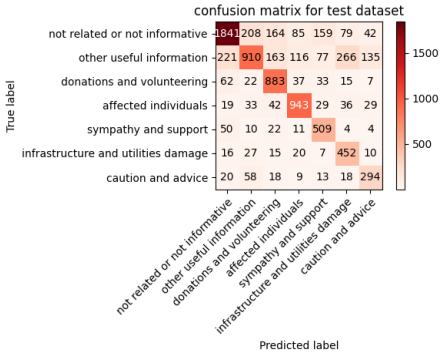


(a) F1 Score for D1

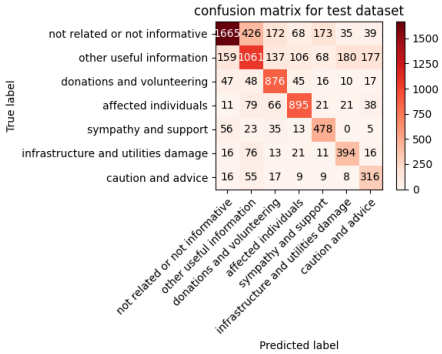


(b) F1 Score for D2

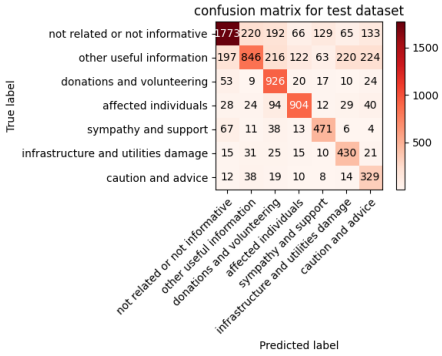
Figure 5. F1 Score for all the BERT-based models



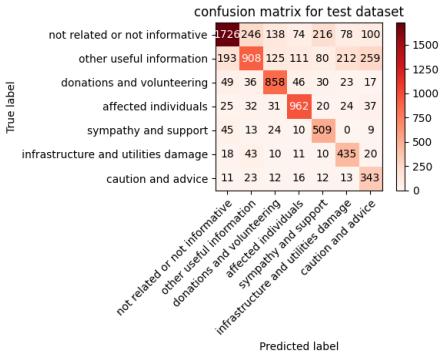
(a) Confusion matrix of Default BERT



(b) Confusion matrix of BERT+NL

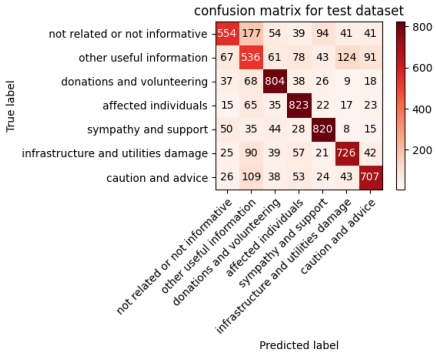


(c) Confusion matrix of BERT+LSTM

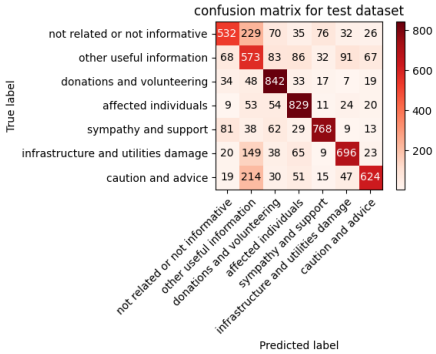


(d) Confusion matrix of BERT+CNN2

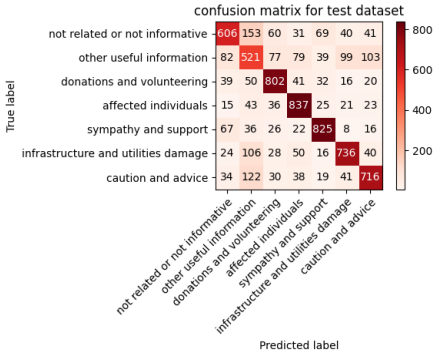
Figure 6. Confusion matrix of test dataset of D1



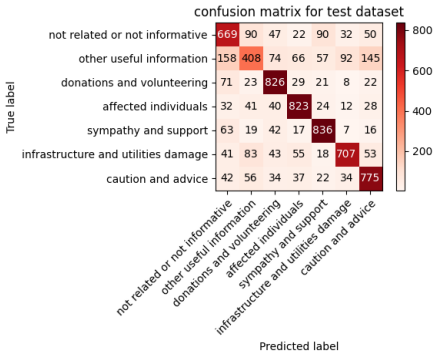
(a) Confusion matrix of Default BERT



(b) Confusion matrix of BERT+NL



(c) Confusion matrix of BERT+LSTM



(d) Confusion matrix of BERT+CNN2

Figure 7. Confusion matrix of test dataset of D2

3.5. Conclusion

The information generated on social media can be utilised in the field of disaster management. The transference of noise from data can lead to better decisions. The data preprocessing is very significant in sequence classification. The way we split the data into train, test and validation sets, also affects the performance of the classifier. The balanced data prove to be better than unbalanced data. The value of accuracy and other evaluation metrics should be up to the mark because the decisions made in the field of disaster management impact lives.

References

[1] Clarkson, K., Srivastava, G., Meawad, F. and Dwivedi, A.D. (2019) Where's@ waldo?: finding users on twitter. In *International Conference on Artificial Intelligence and Soft Computing* (Springer): 338–349.

[2] Kaur, A. (2019) Analyzing twitter feeds to facilitate crises informatics and disaster response during mass emergencies .

[3] Mendhe, C.H., Henderson, N., Srivastava, G. and Mago, V. (2020) A scalable platform to collect, store, visualize, and analyze big data in real time. *IEEE Transactions on Computational Social Systems* .

[4] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P. (2013) Extracting information nuggets from disaster-related messages in social media. In *Iscram*.

[5] Olteanu, A., Vieweg, S. and Castillo, C. (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.

[6] Imran, M., Mitra, P. and Castillo, C. (2016) Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894* .

[7] Stowe, K., Anderson, J., Palmer, M., Palen, L. and Anderson, K.M. (2018) Improving classification of twitter behavior during hurricane events. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*: 67–75.

[8] Walters, T.N. (2008), Ongoing crisis communication: Planning, managing, and responding, wt coombs, sage publications (2007), 207 pp., paper, $45.95.

[9] Purohit, H., Castillo, C., Diaz, F., Sheth, A. and Meier, P. (2014) Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* **19**(1).

[10] Cameron, M.A., Power, R., Robinson, B. and Yin, J. (2012) Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*: 695–698.

[11] Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D. and Aswani, N. (2013) Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*: 83–90.

[12] Han, B., Cook, P. and Baldwin, T. (2013) Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)* **4**(1): 1–27.

[13] Ma, G. Tweets classification with bert in the field of disaster management .

[14] Nair, M.R., Ramya, G. and Sivakumar, P.B. (2017) Usage and analysis of twitter during 2015 chennai flood towards disaster management. *Procedia computer science* **115**: 350–358.

[15] Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*: 241–250.

[16] Kongthon, A., Haruechaiyasak, C., Pailai, J. and Kongyoung, S. (2014) The role of social media during a natural disaster: a case study of the 2011 thai flood. *International Journal of Innovation and Technology Management* **11**(03): 1440012.

[17] Praznik, L., Srivastava, G., Mendhe, C. and Mago, V. (2019) Vertex-weighted measures for link prediction in hashtag graphs. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE): 1034–1041.

[18] Morris, J., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D. and Qi, Y. (2020) Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*: 119–126.

[19] Géron, A. (2019) *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (O'Reilly Media).

[20] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .