A Comparative Study of Deep Learning Methods for the Detection and Classification of Natural Disasters from Social Media

Spyros Fontalis, Alexandros Zamichos, Maria Tsourma, Anastasis Drosou and Dimitrios Tzovaras
Information Technologies Institute, Centre for Research and Technology Hellas (CERTH),

Keywords: Disaster Management, Twitter, Preprocessing, Bias Mitigation, Deep Learning.

Abstract: Disaster Management, defined as a coordinated social effort to successfully prepare for and respond to disasters, can benefit greatly as an industrial process from modern Deep Learning methods. Disaster prevention organizations can benefit greatly from the processing of disaster response data. In an attempt to detect and subsequently categorise disaster-related information from tweets via tweet text analysis, a Feedforward Neural Network (FNN), a Convolutional Neural Network, a Bi-directional Long Short-Term Memory (BLSTM), as well as several Transformer-based network architectures, namely BERT, DistilBERT, Albert, RoBERTa and DeBERTa, are employed. The two defined main tasks of the work presented in this paper are: (1) distinguishing tweets into disaster related and non relevant ones, and (2) categorising already labeled disaster tweets into eight predefined natural disaster categories. These supported types of natural disasters are earthquakes, floods, hurricanes, wildfires, tornadoes, explosions, volcano eruptions and general disasters. To achieve this goal, several accessible related datasets are collected and combined to suit the two tasks. In addition, the combination of preprocessing tasks that is most beneficial for inference is investigated. Finally, experiments have been conducted using bias mitigation techniques.

1 INTRODUCTION

Over the last decade, social media networks have entered in people's everyday lives, allowing them to post and share any information that one considers important. Daily, the number of people using social media, such as Twitter, is growing, leading to an increased flow of information (Chaffey, 2016). The importance of this information lies in the fact that users can post it from anywhere and also that they can post any information instantly without any barrier. This allows for third parties, bearing the property of developers and research scientists, to collect and analyse this information aiming to extract general features such as public opinion upon one topic (Neri et al., 2012), or create an evacuation plan in case of an occurring natural disaster.

On this basis, two types of text classifiers based on Deep Learning models have been developed. The first classifier plays the role of a real natural disaster detector, whose goal is the binary classification of tweets into those that relate to a real natural disaster, and those that are irrelevant. The second classifier classifies tweets that are already known to refer to natural disasters based on predefined natural disaster types.

Several previous attempts provide an evaluation of methods for classifying tweet disaster, such as this evaluation of machine learning techniques (Kumar et al., 2019) and this BERT-like model evaluation (Zhou et al., 2022) to this task. However, none of the above approaches combine a clear evaluation of a wide range of machine and deep learning models on a large and diverse dataset.

Our contributions to this paper are:

- Evaluating and comparing several Deep Learning classifiers for two separate disaster tweet classification tasks.
- Experimenting with the combination of preprocessing steps that (1) maximises the efficiency of the two classifiers on the above downstream tasks and (2) mitigates the pre-existing bias of our collected training datasets.

2 RELATED WORK

There are a number of notable previous attempts to classify disaster tweets, using a variety of methods. Previous attempts include a number of deep learning techniques, such as convolutional neural networks

Fontalis, S., Zamichos, A., Tsourma, M., Drosou, A. and Tzovaras, D.
A Comparative Study of Deep Learning Methods for the Detection and Classification of Natural Disasters from Social Media.
DOI: 10.5220/0011666500003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 320-327
ISBN: 978-989-758-626-2; ISSN: 2184-4313

(Nguyen et al., 2017) and recurrent neural networks (Nikolov and Radivchev, 2019), as well as conventional machine learning algorithms (Huang and Xiao, 2015). Of particular note is the introduction of a robust transformer for crisis classification and contextual crisis embedding (Liu et al., 2021). An interesting domain adaptation technique is also used by (Li et al., 2018), which learns classifiers from unlabelled target data, in addition to labelled source data.

Concerning the latest advances in the field of Natural Language Processing, the current state-of-the art architecture is the Transformer architecture (Vaswani et al., 2017). The first model to apply this architecture to language modelling in an encoder-decoder context is BERT (Devlin et al., 2018). BERT has achieved state-of-the-art results in a large number of NLP benchmark downstream tasks. The effectiveness of the architecture proposed by BERT has paved the way for a lot of suchlike attempts, providing the inspiration for various kinds of modifications and improvements.

An important category of BERT variations is dealing with size reduction. During pre-training, BERT sets to adjust millions of parameters. This compels the process to sometimes be exclusionary for many researchers or small companies to implement (Schwartz et al., 2020). An important aspect of this characteristic is the consideration of the environmental impact that the training process entails (Strubell et al., 2019). In this context, DistilBERT (Sanh et al., 2019) is a successful transfer-learning operation, which demonstrates a reduction on the size of the original BERT model by 40%, while retaining 97% of its language understanding capabilities. Similarly, ALBERT (Lan et al., 2019) is a smart approach to performing novel distillation techniques on the base BERT model.

One of the most successful BERT variants is the RoBERTa model (Liu et al., 2019), which has exposed a lot of BERT's main weaknesses (Cortiz, 2021) and proved that it is severely under trained from reaching its full potential. XLM-RoBERTa (Conneau et al., 2019) is trained on one hundred languages and demonstrates that multilingual language modelling is not necessarily associated with performance degradation. Finally, the DeBERTa (He et al., 2020) architecture improves the BERT and RoBERTa models using two novel techniques, that of disentangled attention and the use of an enhanced mask decoder.

3 DATASETS - PREPROCESSING

This section presents the datasets used for training, along with their analytical synthesis composition.

In total, four final datasets are used for the experiments in this paper. These are the *Kaggle* dataset which is first analysed in Section 3.1, the *Synthetic binary dataset* which is analyzed in Section 3.2, the *Multi-class binary classification dataset* which is analysed in Section 5.4 and finally the *Synthetic multi-class dataset* which is analysed in Section 3.3. The first three are part of the first text classification task and the last one is part of the second text classification task. All the above datasets are divided into 80% train, 10% validation and 10% test sets.

3.1 Data Sources

Listed here are all the various independent sources we have combined to create our datasets. These are the following:

1. *CrisisLex*: Crisis-Related Social Media Data and Tools (Olteanu et al., 2014).
2. *HumAID*: Human-Annotated Disaster Incidents Data from Twitter by CRISISNLP (Alam et al., 2021).
3. *Disaster Eyewitness Tweets* (Zahra et al., 2020).
4. *Kaggle* [1]. This dataset is provided by the relevant Kaggle competition "Natural Language Processing with Disaster Tweets".
5. *Volcano Eruptions Tweets*. This dataset contains 2516 tweets collected using the Twitter API and are referring to two volcano eruptions (i.e. Honga Tonga & La Palma volcanoes)

3.2 Binary Classification of Tweets into "disaster" and "non relevant"

Three distinct datasets are used for this task:

1. For the binary classification task of decoupling the disaster tweets from the non-disaster ones, the *Kaggle* dataset is primarily used.
2. To achieve more diverse representation of the non-relevant class, 5000 random disaster unrelated tweets were extracted from the CrisisLex dataset and combined with the Kaggle dataset. The final result is a merged dataset with a total of 12373 tweets, with 4535 tweets referring to disasters and 7838 being non-relevant. This dataset henceforth referred to as *Synthetic binary* dataset.
3. The *Multi-class binary classification* dataset is comprised of 46672 tweets referring to disasters or not. Its structure is analysed at Section 5.4.

[1] www.kaggle.com

Table 1: Detailed synthesis of the *Synthetic multi-class* dataset.

Disaster categories	Source datasets					
	Disaster eyewitness	HumAID	CrisisLex	Volcano er.	Kaggle	Total
earthquake	3980	2015	-	-	-	5980
flood	3980	1302	-	-	-	5282
hurricane	3940	1654	-	-	-	5594
wildfire	1964	3757	-	-	-	5721
tornado	-	-	4172	-	-	4172
explosion	-	-	4239	-	-	4239
volcano eruption	-	-	-	2516	-	2516
general disasters	-	-	-	-	3271	3271
Total	13864	8728	8411	2516	3271	**36775**

3.3 Multi-label Classification of Tweets into Predefined Disaster Categories

All the data sources that were combined to create the new ensemble dataset for this task are listed in Section 3.1. The final dataset contains 36775 tweets and is henceforth referred to as *Synthetic multi-class* dataset. The full synthesis of this dataset can be explored in Table 1 and observed visually in Figure 1:

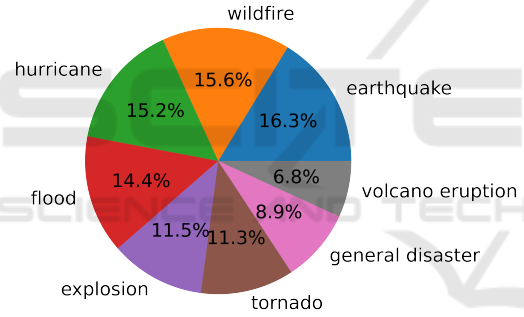

Figure 1: Visual synthesis of the *Synthetic multi-class* dataset.

Some tweet categories refer to specific natural disaster cases. An interesting class addition is the general disaster class, which was filled with the tweets from the Kaggle dataset that are labelled as disasters. This choice was made because most of them do not correspond to the other predefined categories. Those that do refer to predefined disaster categories, or those that refer to non-natural disasters, such as shootings, are regarded as noise that can help with model generalisation.

3.4 Preprocessing

Most textual data extracted from social media is unstructured, as it typically contains colloquialisms, html tags, emojis, scripts, hashtags, links and advertisements, which makes it difficult to decontextualize the main text from all this peripheral noise (Baldwin et al., 2013). Generally, allowing the presence of these non-word entities in the text increases the dimensionality of the unseen vocabulary. This leads to an excessive complication in text classification, as each non-word entity is considered an individual dimension by the machine (Kumar and Dhinesh Babu, 2019). Therefore, a typical list of preprocessing tasks includes the following: link removal, html tag removal, URL removal, emojis removal, mention removal, named entities removal, removing of stop words, lemmatization, stemming, lowercasing and punctuation removal (Anandarajan et al., 2019). Also, due to the idiosyncratic textual nature of tweets, the pipeline also includes the removal of links, html tags, URLs, emojis and hashtags.

However, it is advocated (Uysal and Gunal, 2014) that carefully choosing appropriate combinations of preprocessing tasks, rather than enabling or disabling them all, can potentially provide a boost to the effectiveness of classification depending on the domain and language. For this reason, some preliminary experiments were conducted (see Section 5), so as to capture the combination of preprocessing tasks that improves the results of the classification tasks the most.

3.5 Bias Mitigation

Due to the nature of the information collected from social media platforms and the scope of the task, the collected data includes position biases as it contains information on the location of the disastrous event. With this in mind, and knowing that the models that are trained on this dataset internalise biases with respect to certain case specific words or expressions (Garrido-Muñoz et al., 2021), an additional experiment was conducted, concerning the application of bias mitigation methods on the synthetic dataset in order to evaluate the outcomes of the multi-class classification task.

Positional bias concerns the inclusion of location's information within a text. For example, all the tweets referring to the tornado predefined disaster category correspond to a single tornado case, that of the 2013 Oklahoma tornado. To avoid the potential bias problem, there needs to be a disassociation of special case specific terms with the target disaster classes. Inspired by previous attempts at bias mitigation (Dixon et al., 2018) (Murayama et al., 2021), a relevant technique is applied to the input data before feeding it into the models.

The most important named entities are recognised and replaced by special code tokens within the text with the use of spaCy. The entity types that are replaced are the following: people, nationalities, buildings and facilities, companies, agencies, institutions and locations.

4 METHODS

This section presents the methodology behind all the experiments performed in detail.

The method of **Logistic Regression** is used to establish a reasonable machine learning baseline for the results of the rest of the deeper methods. The TF-IDF method is employed for token vectorization.

Next, the three custom shallow networks evaluated are presented. These are a feedforward neural network, a convolutional neural network and a Bi-directional Long Short-term Memory network. First, as input to the networks, the tweet word sequences get tokenized by a Keras text vectorization function which creates a vocabulary of 12000 words from the dataset For each network, a broad hyperparameter grid search is performed by trying a variety of hyperparameter configurations and recording the best final result. The shifting variables of the hyperparameter search grid are the input embedding dimensions (50-100-200), the total number of layers (2-3-4), the layer dimensions (16-32-64 for LSTM, 10-50-100 for dense) and the dropout rate (0-0.1-0.2). The invariant hyper-parameters for all the training instances are: 100 maximum sequence length, Adam optimiser, categorical cross entropy / binary cross entropy as loss function.

First, for the **Feedforward Neural Network**, the whole embedding input is passed through either a flattening layer or a global Max Pooling operation, and then to a series of one or two of standard dense layers with dropout rate. Following the same workflow as before, a custom network whose structural core is based on **Convolutional layer(s)** is tried. The structure of the network after the embedding input is completed with a sequence of one or two convolutional and global max pooling layers of dimensions followed by a dense layer between the output of the last max pooling layer and the final output layer. The convolution dimensions are 128 5×5 filters with stride 2 in width and height. Additionally, a custom network whose structural core is based on a **Recurrent Neural Network** and more specifically, long short-term memory layer, is tried. The structure of the network after the embedding input is completed with one or two iterations of BILSTM layers with a dropout, followed by a dense layer between the output of the last max pooling layer and the last output layer.

For the **Transformer models**, a variety of transformer models and variations are tried, all of which are presented at 2. All the Transformer models were trained for 5 epochs with a 1e-5 learning rate and Adam as optimiser, with an early stopping strategy.

5 EXPERIMENTAL RESULTS

In this section, all our experimental results are displayed, along with some useful short comments and analysis.

5.1 Preprocessing

The aim of these experiments is to find out which preprocessing tasks perform the best for each model category. To this end, three models were trained on data preprocessed in different ways and the average accuracy from 3 different experiments was measured. The results can be seen on Table 2. The standard pipeline refers to all the steps mentioned in Section 3.4.

It seems that the set of preprocessing tasks that performs the best differs for each model. Therefore, the following experiments all adhere to the appropriate preprocessing procedure for the trained model. More specifically, the input data of the custom CNN network is preprocessed by the same pipeline indicated by the custom RNN. In similar manner, the input data of all the Transformer models are preprocessed by the pipeline indicated by BERT base.

5.2 Bias Mitigation

Some examples of the output of a reference Transformer model - in this case BERT base - with and without the use of bias mitigation are shown in Table 3. The sentences include a specific case-specific bias present, while semantically each sentence is referring to a disaster type not corresponding to this bias. For

Table 2: Effect of applying different combination of preprocessing steps on the results (mean accuracy from 3 identical experiments) of the binary classification task.

Preprocessing tasks	**Models**		
	Log. Regression	**Custom RNN**	**BERT base**
standard pipeline	**0.7946**	0.7812	0.8293
no stemming	0.7855	0.7931	**0.8328**
no lemmatization/stemming	0.7839	**0.8084**	0.8263
no stopwords removal	0.7841	0.7788	0.8216
no lemmatization/stemming/stopwords removal	0.7783	0.7876	0.8208

Table 3: Effect of Bias Mitigation on disaster type classification prediction using BERT model.

Test sentences			**predicted class (normalized logit)**	
	Active biases	**Ground truth**	**No bias mitigation**	**Bias mitigation**
Pity such beautiful nature was destroyed by the fire *#LaPalma*	volcano eruption	wildfire	volcano eruption (0.379)	wildfire (0.345)
Oklahoma will stand strong after the explosion	tornado	explosion	tornado (0.648)	explosion (0.255)
Huge earthquake in *Alberta!* our homes are destroyed	wildfire	earthquake	earthquake (0.268)	earthquake (0.292)
Greeces' tourism at an all time high despite tragic earthquake	wildfire	earthquake	earthquake (0.213)	earthquake (0.241)

example, the second sentence has an active bias towards the tornado class, because in the input dataset, the word "Oklahoma" is encountered only in tweets referring to tornado cases. Also, this particular sentence matches a different class, namely the tornado class, and thus poses a challenge to the model.

The model trained without bias mitigation is heavily affected by the biased words. Performing bias mitigation leads to correct predictions in the first two cases, while it raises the probability of the correct class in the last two. Performing this bias mitigation technique seems to steer the model in the right direction by undermining the effect of problematic biases in the dataset. For this reason, all the following experiments are performed in datasets which have been processed with this technique.

5.3 Classification Tasks

The classification results from all the methods tried for the binary classification of Tweets into disaster and non relevant can be seen in Table 4 for all related datasets. The rows where the *Multi-class binary* classification dataset is inscribed, represents a series of experiments that is explained in subsection 5.4 below. The classification results from all the methods tried on the classification of tweets into predefined disaster categories can be seen at Table 5. In order to mitigate statistical randomness, all the experiments were run 3 separate times, and their average results are shown.

5.4 Merging the Classification Tasks

Given the superior results of the transformers models on the downstream task of classification into multiple classes, it is interesting to find out whether these models can perform binary classification into disaster and non relevant tweets accurately, even though they have been trained for another task. This way, a conclusion can be drawn about whether it is worth embedding the two tasks together and solving both without having to train the model for both tasks separately.

More specifically, a new class of 9897 tweets non relevant to disasters is added to the *synthetic multi-class* dataset under the newly found class non relevant. The pre-existing eight disaster type classes are going to correspond to the disaster class. Thus, along with the disaster type classification results, we examine if this class correspondence can lead accurately to simultaneous binary classification of tweets into the disaster and non relevant classes. Their binary classification scores are in the last rows of Table 4.

6 DISCUSSION

In this section, observations and conclusions from the two tables of results are commented on and analysed.

Table 4: Results from the binary classification of tweets into 'disaster' and 'non relevant'.

Model/Method	**Data-set**	**Results**					
		Acc.	**Precision**	**Recall**	**F1 micro**	**F1 macro**	**F1 weighted**
Log. regression	Kaggle	0.7946	0.8220	0.6662	0.7946	0.7839	0.7909
Custom FFN	Kaggle	0.7832	0.7734	0.7893	0.7832	0.7757	0.7793
Custom CNN	Kaggle	0.7975	0.7785	0.7791	0.7975	0.7784	0.7841
Custom RNN	Kaggle	0.8084	0.7955	0.7936	0.8084	0.7848	0.7953
BERT base	Kaggle	0.8328	0.8036	**0.8066**	0.8328	0.8294	0.8329
Albert	Kaggle	0.8334	0.8131	0.7931	0.8334	0.8293	0.8331
DistilBERT	Kaggle	0.8143	0.8567	0.6822	0.8143	0.8042	0.8104
RoBERTa base	Kaggle	0.8355	**0.8711**	0.7688	0.8355	**0.8299**	0.8351
RoBERTa large	Kaggle	**0.8374**	0.8591	0.7764	**0.8374**	0.8283	**0.8363**
XLM-RoBERTa	Kaggle	0.8292	0.8236	0.7647	0.8292	0.8238	0.8282
DeBERTa base	Kaggle	0.8341	0.8156	0.7921	0.8341	0.8231	0.8277
BERT base	Synthetic binary	0.8293	0.8099	0.8036	0.8293	0.8270	0.8272
Albert	Synthetic binary	0.8353	0.8183	0.7977	0.8353	0.8250	0.8295
DistilBERT	Synthetic binary	0.8341	**0.9165**	0.7133	0.8341	0.8256	0.8287
RoBERTa base	Synthetic binary	0.8373	0.8731	0.7964	0.8373	0.8367	**0.8397**
RoBERTa large	Synthetic binary	0.8399	0.8678	0.8049	0.8399	0.8359	0.8381
DeBERTa base	Synthetic binary	**0.8404**	0.8532	**0.8093**	**0.8404**	**0.8326**	0.8308
XLM-RoBERTa	Synthetic binary	0.8232	0.8256	0.7122	0.8232	0.8194	0.8235
BERT base	Multi-class binary	0.7033	0.6999	0.7012	0.7033	0.7021	0.7029
Albert	Multi-class binary	0.6911	0.6910	0.6921	0.6911	0.6896	0.6908
DistilBERT	Multi-class binary	0.6989	0.6943	0.6948	0.6989	0.6965	0.6971
RoBERTa base	Multi-class binary	**0.7067**	0.7048	**0.7055**	**0.7067**	**0.7056**	0.7060
RoBERTa large	Multi-class binary	0.7061	**0.7056**	0.7051	0.7061	0.7046	**0.7062**
XLM-RoBERTa	Multi-class binary	0.7053	0.7031	0.7022	0.7053	0.7038	0.7041

Table 5: Results from classification of tweets into predefined disaster categories on the *Synthetic multi-class* dataset.

Model/Method	**Results**					
	Accuracy	**Precision**	**Recall**	**F1 micro**	**F1 macro**	**F1 weighted**
Custom FFN	0.9061	0.8883	0.8912	0.9061	0.8903	0.8908
Custom CNN	0.9039	0.9045	0.9039	0.9039	0.9013	0.9033
Custom RNN	0.9074	0.9054	0.9088	0.9074	0.9066	0.9057
BERT base	0.9222	0.9234	0.9215	0.9222	0.9205	0.9210
Albert	0.9191	0.9195	0.9188	0.9191	0.9175	0.9178
DistilBERT	0.9176	0.9199	0.9167	0.9176	0.9184	0.9180
RoBERTa base	**0.9271**	**0.9274**	0.9267	**0.9271**	**0.9273**	**0.9269**
XLM-RoBERTa	0.9252	0.9237	**0.9283**	0.9252	0.9250	0.9251
DeBERTa base	0.9243	0.9174	0.9212	0.9243	0.9251	0.9254

6.1 Binary Classification of Tweets Into 'disaster' and 'non relevant'

A lot of interesting observations can be made from Table 4. For the machine learning method - *Logistic regression* - the recall value of 0.66 is poor compared to the others, meaning that this method often classifies disaster tweets as non-relevant. The rest of the scores are generally lower than the other methods, although very comparable. The overall decent performance of such a shallow method compared to the others could mean that the semantic and grammatical language features that distinguish disaster tweets from non-relevant ones are apparent enough to be able to be analyzed by simpler methods well enough. The *custom networks* all perform slightly worse than the Transformer models. As for the *Transformer models*, the qualitative difference in their results compared to the other methods is significant. The performance of all the Transformer models is comparable. It seems that the DeBERTa model generally performs better than the rest.

Finally, the binary classification results when including the non-disaster class in the multiclass setting show a significant drop in performance in comparison with the previous methods. According to post-experiment analysis, it seems that all the models generally have a problem distinguishing the non-relevant and the general disaster tweets. For reference, from approximately 36% of the non-related tweets that are misclassified by BERT base, 54% of them are assigned as general disaster tweets. Following the same principle, approximately 35% of general disaster tweets that are misclassified by BERT base, 43% of them are classified as non-relevant.

6.2 Classification of Tweets Into Predefined Disaster Categories

The results from the *custom networks* differ only slightly from those of the deeper Transformer based methods. The results of the *Transformers* outperform the previously tested custom neural networks as expected. The best transformers model in terms of results is the Roberta base model.

7 CONCLUSION

In this paper we perform an evaluation of Deep Learning methods on two text classification tasks. The first task classifies tweets into disaster related and non relevant classes and the second classifies disaster tweets into predefined disaster categories. The combination of preprocessing steps that enables each model to learn better is identified, showing that sometimes the omission of certain typical preprocessing steps can lead to better downstream classification. Also, it is shown that mitigating the bias through named entity substitution in the input datasets is an effective strategy when data sources are limited. The three shallow custom neural networks - feedforward, convolutional and recurrent - perform well on both tasks. As expected, the Transformer models outperform the previous methods by a considerable margin. The best overall results are achieved by the DeBERTa model for the first task and the RoBERTa base for the second. Finally, embedding the two tasks together is not a fruitful idea, as it gives very poor results.

The results obtained from the experiments have the potential to be used in practice, showing capacity to effectively perform automatic disaster detection from social media in service of disaster relief organizations. Future work could include the acquisition of more diverse datasets through manual annotation, and the application of more sophisticated bias mitigation and bias measurement techniques, as demonstrated in (Dixon et al., 2018).

ACKNOWLEDGEMENTS

This research was supported by grants from Horizon 2020, the European Union's Programme for Research and Innovation under grant agreement No. 870373 - SnapEarth, grant agreement No. 101004594 - ETAPAS and grant agreement No. 101037648 - SOCIOBEE. This paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Alam, F., Qazi, U., Imran, M., and Ofli, F. (2021). Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *ICWSM*, pages 933–942.

Anandarajan, M., Hill, C., and Nolan, T. (2019). Text preprocessing. In *Practical Text Analytics*, pages 45–59. Springer.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Chaffey, D. (2016). Global social media research summary 2016. *Smart Insights: Social Media Marketing*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. *arXiv preprint arXiv:2104.02041*.

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*.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A. (2021). A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Huang, Q. and Xiao, Y. (2015). Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4(3):1549–1568.

Kumar, A., Singh, J. P., and Saumya, S. (2019). A comparative analysis of machine learning techniques for disaster-related tweet classification. In *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)*, pages 222–227. IEEE.

Kumar, P. and Dhinesh Babu, L. (2019). Novel text preprocessing framework for sentiment analysis. In *Smart intelligent computing and applications*, pages 309–317. Springer.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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.

Liu, J., Singhal, T., Blessing, L. T., Wood, K. L., and Lim, K. H. (2021). Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 133–141.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Murayama, T., Wakamiya, S., and Aramaki, E. (2021). Mitigation of diachronic bias in fake news detection dataset. *arXiv preprint arXiv:2108.12601*.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., and By, T. (2012). Sentiment analysis on social media. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 919–926. IEEE.

Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh international AAAI conference on web and social media*.

Nikolov, A. and Radivchev, V. (2019). Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 691–695.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth international AAAI conference on weblogs and social media*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12):54–63.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Zahra, K., Imran, M., and Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, 57(1):102107.

Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., and Mandal, D. (2022). Victimfinder: Harvesting rescue requests in disaster response from social media with bert. *Computers, Environment and Urban Systems*, 95:101824.