Afetle İlgili Tweetleri Belirleme ve Sınıflandırma

Toygar Tanyel
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{l1118094}@std.yildiz.edu.tr

***Özetçe* —Bu çalışmada, tweetleri kullanarak afet analizi yapmak için rekabetçi bir ardışık derin öğrenme modeli öneriyoruz. Eğitim aşamasındaki sorunları azaltmak için bağlam bağımlı ve bağımsız iki ardışık modelimiz incelenmiş ve gerçeklenmiştir. Bu model, evrişimsel sinir ağları (CNN) ve iki yönlü uzun-kısa vadeli bellek(BiLSTM) ardışık olarak farklı tipteki (GloVe & BERT) kelime vektörleri ile etkili sonuç verebilecek şekilde tasarlanmıştır. Modelimizi beslemek için sosyal medya verilerini kullandık. Günümüzde, özellikle afet kaynaklı kriz anlarında sosyal medyanın önemli bir yeri vardır. Sosyal medyanın ürettiği yüksek miktardaki veri, afet analizi için eşsiz bir fırsat sunmaktadır. Özellikle Twitter gibi sınırlı sayıdaki karakterle bilgi veren platformlar, afet sırasında eyleme geçirilebilir ve taktiksel bilgi edinimi için gerçek zamanlı bir bilgi çıkarım platformu olarak birçok araştırmacı tarafından aktif olarak kullanılıyor. Anlık bilgi alınabilme özelliği nedeniyle, daha fazla kurum, hızlı bir kurtarma planı yapmak için afet olaylarını izlerken Twitter'ı kullanıyor. Ancak, uzunluk sınırı nedeniyle yeterli bağlamdan yoksun olabilecek felaket Tweetlerini belirlemek için doğru bir tahmine dayalı model oluşturmak zordur. Bununla birlikte bu modeli destekleyecek veri seti oluşturmak da aynı oranda zorlu bir görevdir. Afet analizi için önerdiğimiz model ile umut verici sonuçlar elde ettik.**

Anahtar Kelimeler—Sosyal medya, keşifsel veri analizi, sınıflandırma, derin öğrenme, BERT, GloVe, CNN, BiLSTM.

***Abstract*—In this study we propose a competitive deep learning model pipeline for disaster analysis using tweets. Our context dependent and independent pipelines are developed to decrease problems in training phase, and to solve these problems, our model pipeline constructed as the combination of convolutional (CNN) and sequence processing (BiLSTM) models will be evaluated with different type of word embeddings (GloVe & BERT). To feed our model we utilized the social media data. Social media has an important role in today's world especially at the moments of disaster related crisis. Social media generates an enormous amount of data and provides a special opportunity to classify disaster Tweets from regular ones. In particular microblogging platforms such as Twitter, are becoming approved among many researchers as a real-time scraping platform for getting actionable and tactical information during disasters. Due to its instantaneous speciality, more and more agencies are using Twitter for monitoring disaster events to plan a quick rescue. However, building an accurate predictive model to identify disaster Tweets quite challenging. The lack of sufficient context due to the length limit reveals many other problems. For disaster analysis task, we acquired promising results with proposed model pipeline.**

Keywords—Social media, exploratory data analysis, text preprocessing, classification, deep learning, BERT, GloVe, CNN, BiLSTM.

I. Introduction

In this study we provide a model pipeline which includes combination of deep learning models. The model offers high-quality linguistic features extraction even in deficient datasets. Our deep learning model pipeline developed by 3 different modules, an embedding layer (BERT or GloVe), a 1D convolutional layer, and a bidirectional lstm layer.

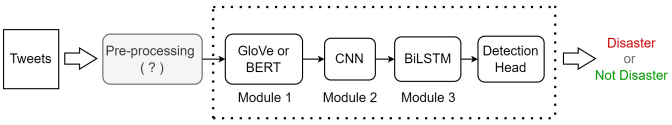

Figure 1 The proposed learning pipeline. (?) indicates whether the dataset has been preprocessed or not. Module 1 is embedding layer, Module 2 is a convolutional layer, and Module 3 is a bidirectional lstm layer. Detection head determines the result.

We utilized natural language processing methods and models to accelerate progress of improvements on the task. Pre-processing step is one of our study to show its effects with different model pipelines. Therefore, this step both performed and unperformed. The GloVe and BERT embeddings are chosen for extracting, respectively, statistical and sentimental features from texts to show difference between them. CNN is employed as feature extractor of the model pipeline while BiLSTM is keeping the chronological order between words in given texts. The last layer which named detection head shows up the result.

We offer classifying disaster tweets from regular ones, and creating a system where provides locations or essential information to relevant institutions and organizations is significant necessity during disasters such as earthquakes, floods and forest fires. In this case, social media platforms which can be considered as instantaneous information sources are remarkable to utilize for good purpose.

During a time of crisis, people tend to use social media platforms to post situational updates, look for useful information, and ask for help. Twitter has reached enormous popularity since its beginning. The latest usage statistics of Twitter data show that as of May 2020, on average, around 6,000 tweets are sent in every second. Society generates large volumes of social data that is used by many high-level analytics practices to create additional value. Moreover, many studies have utilized Twitter data to implement natural language processing practices such as sentiment analysis, named entity recognition, and topic modelling.

In addition to its social function, Twitter increase its popularity as being real time platform for tracking events, including disasters, accidents and emergencies, and collecting useful and legal analytical data for policy or marketing. Information becomes knowledge, especially among the new generations which its majority have a smartphone where allows everyone to share an emergency Tweet instantly to be seen. More and more agencies such as disaster relief organizations and news agencies are starting to realize that there are new ways to reach people to channel resources due to the convenience of social media interaction. Organizations constantly monitor Twitter. Thus, first responders can be deployed and rescue plans can be drawn up as soon as possible.

For instance, as given in [1] the first report of the Westgate Mall attack in Nairobi, Kenya in 2013 was published on Twitter, almost 33 minutes before a local TV channel reported the event. Likewise, the news about the Boston bombing incident appeared on Twitter before any other news channel reported the event. Similarly, in the case of the California earthquake it was observed that the first half dozen tweets were recorded by Twitter about a minute earlier than the recorded time of the event according to the USGS. These direct reports come from witnesses and spectators, those who are directly observing what is occurred.

However, the automation of the process requires an robust and quite accurate classifier to distinguish disaster-related tweets from regular ones. The disaster prediction based on tweets is difficult due to people who can use metaphorical words to describe other things.

The rest of this study is organized as follows: Section 2 remarks relevant studies; Section 3 contains the dataset description and the technical details of the proposed learning models; Section 4 provides experimental validation with result analysis; Section 5 summarizes my work and points out future directions.

II. Related Work

Twitter is becoming popular platform which allows researchers to utilize the social data for emergency and disaster analysis in last few years [2], [3], [4]. Various methods are studied and implemented for extraction of key phrases for general purpose [5], disaster-relevant [6], and knowledge based events [7].

The latest works show that in disaster classification using fine-tuned BERT embeddings [8] is outperforming the old methods such as GloVe, Word2Vec and FastText [9],[10]. Moreover, intuitive reason of outperforming is the GloVe like models generate embeddings that are context-independent, whereas BERT embeddings are context-dependent [11], however, it may depend on the dataset [12]. Meanwhile, Bi-LSTM, CNN and GRU can be considered as some of the most common machine learning models in classification tasks. Recently, some studies focuses on creating a hybrid models that include CNN-BiLSTM pipeline with using different embeddings [8], [13] to propose state-of-art results. Besides of these, [10] propose a valuable priority scheduling algorithms to plan rescues with remarkable results.

III. Material & Methods

Twitter data is utilized to train proposed pipeline. We offer various methods in process of the system.

A. Dataset

The chosen dataset contains Twitter data, and available as a Kaggle competition which is still in progress at here (accessed data on 28 Nov 2021). The dataset contain 10.876 samples. There are 7613 rows and 5 columns in train, and 3263 rows and 4 columns in test. The reason of difference with columns is that test set has no labels. The training set includes 3271 real and 4342 not real labels in terms of class distribution. Due to test set has no label, the train set is used as train-val-test set which divided to 6165 samples as train, 762 samples as dev, and 686 samples as test set to obtain results.

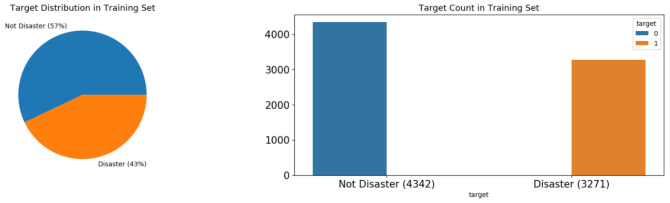

Figure 2 Training set distribution and target count.

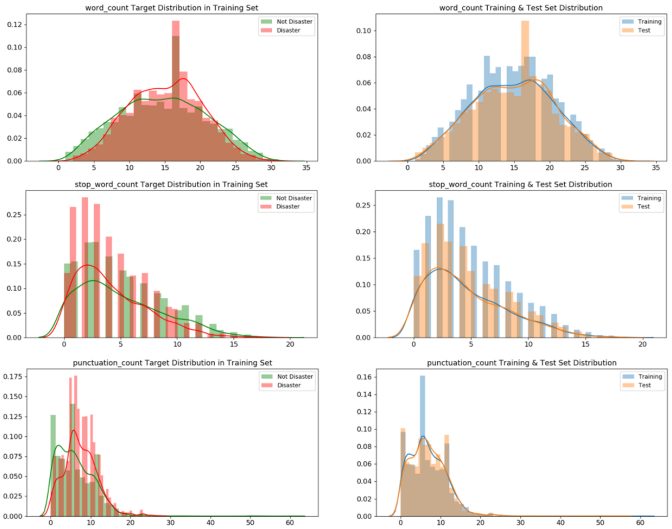

Figure 3 Stats of Tweets in the dataset. Histograms of word count, stop words count and punctuation count sequencially, plotted for disaster and non-disaster Tweets.

The distribution of class and record meta features helps identify disaster tweets. Most disaster tweets come from the news agencies. Therefore, disaster tweets seem to be more formalized in longer words than non-disaster tweets. Disaster tweets contain less typos than non-disastrous ones which comes from individuals. The following notations

are used in the figures as meta features; word count, unique word count, stop word count, url count, mean word length, char count, punctuation count, hashtag count, mention count. All of the meta features have very similar distributions in training and test set which also proves that training and test set are taken from the same sample. Furthermore, meta features have information about target. However, some of them are not sufficient enough to give intuition such as url count, hashtag count and mention count. On the other hand, word count, unique word count, stop word count, mean word length, char count, punctuation count have very different distributions for disaster and non-disaster tweets. Those features might be useful in models.

B. Overview of the Proposed Models

Figure 4 presents the proposed GloVe-CNN-BiLSTM model pipeline, which includes of three consecutive modules. GloVe produces competitive results. Besides that, we will discuss about BERT-CNN-BiLSTM model pipeline, which also includes three consecutive modules. BERT shows its strength in our task, and can be considered as the new electricity of natural language processing tasks such as sentiment analysis, named entity recognition, and topic modeling. A particular combination design is required to be able to use strengths of both CNN and BiLSTM models:

- In text processing, 1D CNN layers are known for their capability to extract as many features as possible.
- BiLSTM is keeping the chronological order between words in given texts, meanwhile, it can ignore the unnecessary words by utilizing the delete gate.

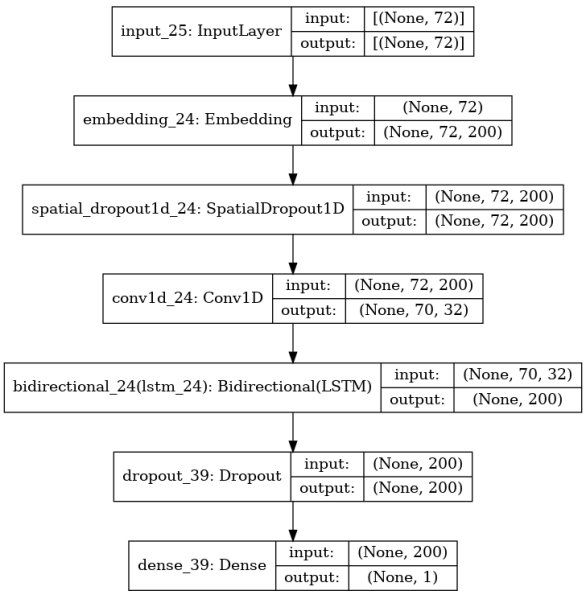

Figure 4 GloVe-CNN-BiLSTM learning pipeline with extended visualization.

Combining CNN and BiLSTM has a particular purpose. We want to create a model pipeline that leverages the strengths of both to capture features extracted using CNN and use them as BiLSTM inputs. Thus, we proposed a model that meets this goal. Word embedding part which includes BERT and GloVe is used as the input of convolutional neural network. Then takes features as input of BiLSTM layers. Afterwards, the output of the BiLSTM is fed to a detection head to produce the final results, in other saying, disaster or not disaster.

1) Global Vectors for Word Representation: GloVe is the unsupervised algorithm is based on data statistics. Models such as Skipgram and CBOW collect semantic information, but do not use co-occurrence statistics. Although matrix decomposition methods use these statistics, they cannot capture semantic relationships. There is no semantics in such models. The "GloVe" model proposed by Pennington et al. [14] aims to solve this problem by constructing a new objective function using probability statistics.

Mathematical perspective of GloVe:

$$J = \sum_{i,j=1}^{V} f(X_{i,j})(w_i^T w_j + b_i + b_j - logX_{i,j})^2$$

2) Bidirectional Encoder Representations from Transformers: BERT [15] is a recent paper published by researchers at Google AI Language. It has caused a stir in the Deep Learning community by presenting state-of-the-art results in a wide variety of NLP tasks. BERT also employs a mask language model (MLM) technique and a next sentence prediction (NSP) task in training to capture word-level and sentence-level contextual information. At this point, to make things clearer it is important to understand the special tokens that BERT authors used for fine-tuning and specific task training.

3) Bidirectional Long Short-term Memory: Original LSTM [16] (Long Short-term Memory), previously published at 1997 to solve problem which is called back-propagation through time. To be more clear, learning to store information over extended time intervals by recurrent back-propagation takes a very long time, mostly because of insufficient, decaying error back-flow. BiLSTM is a sequence processing model that consists of two LSTM, one taking the input in a forward direction, and the other in a backwards direction. Bi-LSTM effectively increase the amount of information available to the network, improving the context available to the algorithm.

4) Convolutional Neural Networks: CNN is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle) & use a variation of multilayer perceptrons designed to require minimal preprocessing. [17] shows that a simple CNN with little hyperparameter tuning and static vectors achieves remarkable results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance.

5) GloVe-CNN-BiLSTM: GloVe provides the ability to derive semantic relationships between words from co-occurrence matrices. Given a corpus having V words, the

co-occurrence matrix X will be a V x V matrix, where the i^{th} row and j^{th} column of X, X_{ij} denotes how many times word i has co-occurred with word j. GloVe gives rest of the model a statistical word embeddings to utilize dataset.

6) BERT-CNN-BiLSTM: BERT is used to transform word tokens from text data to contextual word embeddings. BERT produces competitive prediction results against the models like GloVe. This is because every word under GloVe has a fixed expression regardless of the context in which the word appears, however, BERT produces a word expression that is dynamically influenced by the surrounding words. This situation creates a difference in pre-processing step, and will be explained in section IV-D.

IV. Results and Observations

For performance evaluation we use the disaster Tweet dataset discussed in Section 4.4. First, we will discuss about the performance metrics, afterward, report of the experimental results. We will present an analysis of errors and performances of different pipelines that might give intuition on further improvement of the model.

A. Evaluation Metrics

Precision, Recall, and F1 scores are used to evaluate the performance of models. Positive and negative samples are unbalanced. Thus, F1 score gives better results in this case than accuracy. Precision and recall are also significant as evaluation metrics. The first one represents the number of false positives. If the accuracy increases, the false positives decreases. The second shows the number of wrong predicted positive samples. If the recall increases, the missed disaster tweets decreases. Meanwhile, precision and recall gap should be small. Hence, we can consider single metric is not directing the model. In our case, F1 score needs to be optimized. Let TP, TN, and FP represent the number of true positives, true negatives, and false positives, in order, and we can compute precision in equation (1), recall in equation (2), and F1 in equation (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

B. Preperation for Training and Evaluation

The dataset was separated into training, validation and test sets with the ratio of (9:1):1, we validate our model with, 6165 training, 762 validation, and 686 test samples. For GloVe embedding dimension was 200 and learning rate was 1e-4 *glove.twitter.27B.200d.txt* is used as base embeddings. For BERT embeddings "*bert-base-cased*" is used which includes 12-layer, 768-hidden, 12-heads. For CNN module 32 dimensions are used with 3 filter size. BiLSTM is set to 100 layers with 0.2 dropout on both way. For the overall architecture, learning rate 1e-4, Adam optimizer is used, and experimented with batch size 32 and training epoch 3. All experiments were realized with Python 3.9.7 and Tensorflow 2.7.0 on Kaggle Notebook with TPU v3-8.

C. Analysing of Results

Table 1 presents ten samples. It contains five positive and five negative samples, which are misclassified by the BERT-CNN-BiLSTM model. The first five samples (1-5) are marked as disaster, and for the next five samples (6-10), none of them represent a common sense disaster.

Table 1 Missclassified examples. A "+" represents a disaster, and a "-" represents a not disaster sample.

ID	Text	Label	Predicted
1	Some people are really natural disaster too	+	-
2	Breaking news! Unconfirmed! I just heard a loud bang nearby. in what appears to be a blast of wind from my neighbour's ass.	+	-
3	ALERT! Sandy Hook Elementary School Evacuated After Bomb Threat	+	-
4	We have different moral systems. Mine rejects the mass murder of innocents yours explicitly endorses such behavior.	+	-
5	So much shit has happened today wtf idk how I survive thruuu it all	+	-
6	Our garbage truck really caught on fire lmfao.	-	+
7	My man runs into burning buildings for a living but is scared to hit up a girl. I don't get it.	-	+
8	Google Alert: Emergency units simulate a chemical explosion at NU	-	+
9	?? your last retweet you would think the lion saved people from a burning buildings it's not that deep	-	+
10	China detains seven Christians trying to protect their church's cross from demolition	-	+

1) Mislabeled Texts in Raw Data: We observed that there are some mislabeled samples in the dataset. This situation affects performances of the models. Sample 1 clearly mislabeled in dataset, it is a general sentence which talks about evil of people from perspective of the user, but we cannot classify this sentence as disaster. In sample 5, sentence is clearly mislabeled, general vibe of the sentence with following word "thruuu" is daily issue of the user.

2) Word Ambiguity: Word ambiguity is one of the important problem when it comes to open ideas of people in social media. Therefore, we observed that word ambiguity is an actual problem while classifying. Sample 2 has word ambiguity, everything related with the disaster such as "Breaking News!", "loud bang", and "blast of wind" so it can be disaster, but there might a chance the user tries to make a joke to mess with his/her neighbour.

3) Actual Misclassified Texts by Proposed Model: To improve model we need to understand why the model misclassified the proper texts. Therefore, we cover the actual misclassified texts in the following cases. Sample 3 is an actual alert; sample 4 includes phase "mass murder" and obvious explanation of violence. In sample 6 the word "lmfao" almost every time used in entertaining moments;

sample 7 talks about courage of user's friend; sample 8 is quite close to an actual disaster, however, the word "simulate" changes the semantic meaning to its reverse; sample 9 includes phase "burning buildings", but the rest of sentence irrelevant with the any disaster; sample 10 has bad meaning religiously, but cannot considered as disaster.

These samples are randomly chosen from all error predictions as misclassifed examples which can be seen in the table. The tweet length limit has pros and cons to train a disaster tweet classifier. The good thing is that every user has to write short, meaningful words to explain their ideas due to character restriction by Twitter. The negative side is that many short Tweets just mean nothing. It is quite difficult to give a meaning due to lack of additional contextual information. This is one of the most significant problem with building an accurate model pipeline from Twitter data.

D. Performance Analysis

We performed experiments to evaluate a set of modules, and presented as performance comparison of all evaluated modules in the table 2. The set of models CNN, BiLSTM, and CNN-BiLSTM forms an ablation study with GloVe and BERT embeddings, thus, we can evaluate the many situational effect case on the performance such as existence of pre-processing, and type of embeddings for each individual model and the combinations. We can see that the single-layer CNN model performs unsatisfying results with GloVe embeddings and cannot learn any contextual information, however, a single-layer CNN can learn more contextual information using BERT embeddings. BERT presents an obvious improvement as against GloVe embeddings. BERT utilizes raw dataset better than GloVe embeddings, because as shown in [18] BERT is sensitive to punctuations and cases while creating contextual matrix. On the other hand, GloVe requires data cleaning, because stopwords, punctuations etc. drive context to statistically wrong way. We can easily see the difference on models along with (AP) After Pre-processing in performance table 2. BERT loses its contextual structure which depends on the punctuations and cases while GloVe gains remarkable score when we eliminate the unnecessary punctuations and cases. Overall, CNN-BiLSTM shows its strength in results. Our final pipeline, BERT-CNN-BiLSTM is above all other models, and shows the power of combining the strength of each building block.

Table 2 A performance comparison of models on raw dataset. If pre-processed dataset is used, determined as *(AP) After Pre-process*.

Model	F1 Score	Precision	Recall
GloVe-CNN	63.49	71.65	57.08
GloVe-BiLSTM	73.97	86.19	65.06
BERT-CNN-BiLSTM *(AP)*	77.32	85.28	70.92
BERT-BiLSTM	77.44	**87.61**	69.49
BERT-CNN	77.63	86.99	70.17
GloVe-CNN-BiLSTM	78.13	79.22	77.73
GloVe-CNN-BiLSTM *(AP)*	78.71	78.83	**79.10**
BERT-CNN-BiLSTM	**79.10**	84.74	74.23

V. Conclusion

Classifying disaster tweets is closely related to people's daily lives, and research efforts in this field have increased in recent years. Disaster prediction studies help raise awareness, improve state rescue mechanisms, and plan charitable activities. This study analyzes a new model pipeline for classifying disaster tweets. Our last model pipeline, BERT-CNN-BiLSTM, uses a BERT encoder, a 1D convolutional, and a BiLSTM layer to extract high quality linguistic features for classifying the disaster tweets from the regular ones. This model pipeline was validated against competitors through extensive experimentation which makes it a promising model pipeline for applying to real time disaster detection systems. Although the proposed model pipeline was trained and validated on an English dataset, it can be applied to datasets in other languages.

References

[1] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," & *Management*, vol. 57, no. 1, p. 102107, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457319303590

[2] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages," *CoRR*, vol. abs/1605.05894, 2016. [Online]. Available: http://arxiv.org/abs/1605.05894

[3] R. R. Arinta and E. Andi W.R., "Natural disaster application on big data and machine learning: A review," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 249–254.

[4] H. Shekhar and S. Gangisetty, "Disaster analysis through tweets," 08 2015.

[5] J. R. Chowdhury, C. Caragea, and D. Caragea, "Keyphrase extraction from disaster-related tweets," *CoRR*, vol. abs/1910.07897, 2019. [Online]. Available: http://arxiv.org/abs/1910.07897

[6] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Practical extraction of disaster-relevant information from social media," 05 2013.

[7] S. Zong, A. Baheti, W. Xu, and A. Ritter, "Extracting COVID-19 events from twitter," *CoRR*, vol. abs/2006.02567, 2020. [Online]. Available: https://arxiv.org/abs/2006.02567

[8] G. Song and D. Huang, "A sentiment-aware contextual model for real-time disaster prediction using twitter data," *Future Internet*, vol. 13, no. 7, 2021. [Online]. Available: https://www.mdpi.com/1999-5903/13/7/163

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

[10] M. Y. Kabir and S. Madria, "A deep learning approach for tweet classification and rescue scheduling for effective disaster management," *CoRR*, vol. abs/1908.01456, 2019. [Online]. Available: http://arxiv.org/abs/1908.01456

[11] A. Khatri, P. P, and A. K. M, "Sarcasm detection in tweets with BERT and glove embeddings," *CoRR*, vol. abs/2006.11512, 2020. [Online]. Available: https://arxiv.org/abs/2006.11512

[12] S. Madichetty and S. Muthukumarasamy, "Detection of situational information from twitter during disaster using deep learning models," *Sādhanā*, vol. 45, 12 2020.

[13] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5841

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," pp. 1532–1543, 01 2014.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[17] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[18] A. Ek, J.-P. Bernardy, and S. Chatzikyriakidis, "How does punctuation affect neural models in natural language inference," in *Proceedings of the Probability and Meaning Conference (PaM 2020)*. Gothenburg: Association for Computational Linguistics, Jun. 2020, pp. 109–116. [Online]. Available: https://aclanthology.org/2020.pam-1.15