Multimodal Classifier for Disaster Response

Saed Alqaraleh[1](✉) and Hatice Sirin[2]

[1] Computer Engineering Department, Hasan Kalyoncu University, Gaziantep, Turkey
saed.alqaraleh@hku.edu.tr
[2] Software Engineering Department, Hasan Kalyoncu University, Gaziantep, Turkey

Abstract. Data obtained from social media has a massive effect on making correct decisions in time-critical situations and natural disasters. Social media content generally consists of messages, images, and videos. In situations of disasters, using multimedia files such as images can significantly help in understanding the damage caused by disasters compared to using text only. In other words, the exact situation and the effect of disaster are better understood using visual data.

So far, researchers widely use text datasets for building efficient disaster management systems, and a limited number of studies have focused on using other content, such as images and videos. This is due to the lack of available multimodal datasets. We addressed this limitation in this work by introducing a new Turkish multimodal dataset. This dataset was created by collecting disaster-related Turkish texts and their related images from Twitter. Then, by three evaluators and the majority voting, each sample was annotated as a disaster or not a disaster.

Next, multimodal classification studies were carried out with the late fusion technique. The BERT embedding approach and a pre-trained LSTM model are used to classify the text, and a pre-trained CNN model is used for the visual content (images). Overall, concatenating both inputs in a multimodal learning architecture using late fusion achieved an accuracy of 91.87% compared to early fusion, which achieved 86.72%.

Keywords: Multimodal Classifier · Disaster Management · Tweet Text Classification · Image Classification · Turkish language

1 Introduction

A vast number of images and texts captured during most of our daily events are uploaded to social media platforms worldwide. This large-scale data shared on social media can be classified using visual recognition and textual understanding. Since natural disasters are time-critical, a practical classification of data published on social networks is extremely useful for people in charge and humanitarian organizations to make plans and correct decisions on time. It is worth mentioning that, unfortunately, sometimes, during important events such as disasters, people share irrelevant information with disaster hashtags to ensure that more readers see their tweets.

Disasters are events that negatively affect people, the environment, and societies due to the life losses and damages that occur during disasters. Recently, messages and photographs have been highly used to describe the situations of people and the environment

A. Ortis et al. (Eds.): ICAETA 2023, CCIS 1983, pp. 1–13, 2024.
https://doi.org/10.1007/978-3-031-50920-9_1

during natural disasters such as earthquakes, floods, fires, etc. Social media platforms, where information and news can be accessed and used in real-time, are considered one of the most widely used tools for communication and its purposes. However, in cases of misuse, it can create a chaotic environment and causes various harms. Hence, an automated system that can find the most valuable and relevant information before, during, or after disasters is vital.

Due to advances in deep learning, the performance of both text and image classification methods has increased significantly in recent years. This increases the interest in multimodal deep-learning classification systems. It demonstrates that deep representations of image and text data can be transferred to a new field by performing common deep-learning representations for different data types. However, most of these studies working on introducing efficient systems that automatically classify English only [1, 2]. In contrast, unfortunately, only a few works have been done in this field related to other languages, such as Turkish, and all of them focus only on the text [3–5]. In other words, no Turkish study uses the available text and images of disaster-related information to build an efficient multimodal classifier for disaster response.

Information on emergency management is typically time-sensitive, subject to constant change, and critical to society's readiness to respond to emergencies and disasters. Emergency managers are trying to allow people to report critical situations through all available channels, such as phones, TV, and the Internet (websites like social media). These channels also inform and guide the public before/during, and after disasters. For example, in 2011, a magnitude 5.8 earthquake occurred in the United States; authorities contacted the public via Twitter to report the disaster damage in their regions and inform them what to do. In this earthquake, where calls could not be made due to equipment disruptions, Twitter was used to reach people's relatives and get information from public institutions [6]. Also, during the Van Earthquake on October 23, 2011, in Turkey, people in Van and surrounding places organized social media campaigns to aid, support, and rescue activities. Also, it has been reported that some injured under the rubble of the Van Earthquake used social media to request help.

Overall, it is crystal clear that such platforms can effectively help crisis management. However, due to the immense amount of shared multimodal data, removing redundant and irrelevant information is essential to assist decision-makers in making the most suitable actions during such events.

The manuscript is organized as follows: The recent works and literature related to this study are summarized in Sect. 2. Section 3 describes our proposed method and architecture. Experimental setups, results, and analysis are presented in Sect. 4. Section 5 demonstrates the conclusions of this research.

2 Literature Review

The recent works and literature related to this study are summarized below. In [3], a new social media data analysis framework was proposed. This framework uses deep bidirectional neural networks trained on earthquakes, floods, and extreme flood datasets. It first works on learning from discrete handcrafted features and then fine-tuning the deep bidirectional transformer neural networks. Overall, the developed multiclass classifier

integrated with support vector machines provides a precision of 0.83 and 0.79 for both random and original splits, respectively. While integrating Bernoulli naïve Bayes can achieve 0.59 and 0.76, multinomial naïve Bayes achieved 0.79 and 0.91.

The work presented in [4] aims to create a system that can detect crises that require assistance by making an effective classification for Turkish text tweets using KNN, SVM, and CNN algorithms. Their results indicated that the proposed model could achieve around 94% accuracy. Next, in [5], the work of [4] was further improved, and a text Turkish tweet dataset for crisis response and a deep-learning Turkish tweet classification system for crisis response was presented.

In [7], distinctive features were proposed to classify tweets as contextual and non-contextual. After classifying the tweets, situational tweets summarization techniques were used to convey awareness to government agencies. Then, the system of [7] was experimented with using the Uttarakhand floods and the Nepal earthquakes datasets, which contain tweets in English and Hindi. Results demonstrate that the domain-independent classifier outperforms the domain-dependent technique for English and Hindi tweets.

In [8], a method to classify tweets about damage detection that combines statistical and illuminating features was developed. This method uses Random Forest and AdaBoost classifiers. The results of experimental work in [8] showed that the proposed method outperformed the baseline SVM with the Bag-of-Words model.

Related to multimodal systems, there are two basic approaches to automatically merging different models, in our case, text and images, early fusion and late fusion [9]. Early fusion works on finding the best features from each multi-data (text and image), and the features of both text and image are combined in a single vector [10]. Then, a classifier is trained on top of the standardized vector. Related to the late fusion (voting): each modality, such as the image and text, is propagated through their classifier, and the output probabilities of both models are averaged, where the class is selected using the maximum value [11]. It is worth mentioning that a third type, known as Hybrid Fusion (GMU), exists and can support multiple languages (input text belongs to multiple languages). This type is usually trained using the best features per modality.

In the following, the recent studies related to multimodal systems are summarized. In [12], a simple feature augmentation approach that can leverage the text-image relationship to improve the classification of emergency tweets was presented. Also, a new multimodal dataset containing 4600 tweets (image + text) collected during the 2017 USA disasters and manually annotated was introduced. The model was tested on two categories, i.e., Humanitarian and Damage Assessment, and the observations indicated increased performance.

In [13], another multimodal classification model for mining social media disaster information was introduced. This model uses the Late Dirichlet Allocation (LDA) to identify subject information. Bert embedding and VGG-16 are used to analyze the multimodal data, where text and image data are classified separately. The Weibo data collected during the 2021 Henan heavy storm was used. Their results showed that an improvement of 12% can be achieved using the proposed model compared to "KGE-MMSLDA", a topic-based event classification model.

In [14], the integration of disaster data provided by text and image was investigated. Then, the attention mechanism was used to build a multimodal deep learning model called CAMM. This model was compared with "MUTAN' and "BLOCK" unimodal models and outperformed both by 6.31% and 5.91%, respectively.

Another multimodal disaster identification system was presented in [15]. This model combines the visual features with word features to classify each input tweet. In more detail, the visual features are extracted using a pre-trained convolutional neural network, while the textual features are extracted using a bidirectional long-term memory (BiLSTM) network with an attention mechanism. Next, a feature fusion and the Softmax classifier are used to decide on the class. Overall, the system of [15] outperformed some baselines unimodal and multimodal models by 1% and 7%, respectively.

Until this study, no multimodal crisis management systems were built specifically to classify Turkish language text and images. This work introduces the first-ever Turkish language multimodal crisis management system. We first collected Turkish text and visual tweets related to natural disasters to achieve our goal. Then, intensive experiments were performed to produce an efficient multimodal classification system that processes text and image data shared on Twitter before, during, and after natural disasters and informs authorities about relevant and critical disaster-related information.

3 The Proposed Multimodal Learning Approach

In this work, three evaluators annotated the collected samples as relevant to or irrelevant to natural disasters. Next, a new automated and multimodal classifier that supports the Turkish language was proposed.

As shown in Sect. 4, and based on intensive investigations, an LSTM was selected for text classification. At the same time, an in-depth feature extracted from a fully connected layer of AlexNet-based CNN achieved the best performance for image classification. In addition, after comparing the performance of BERT, Glove, and Word2Vec embedding systems, we selected BERT as it showed superior performance. Finally, we perform the late fusion of classification scores. As shown in Fig. 1, we verify the integration of visual and textual methods with multimodal techniques.

3.1 The Collected Sample

The collected Twitter data consisted of Turkish texts and their related images. The data will be used to analyze whether it is relevant to natural disasters or not. To collect the samples, deprem (earthquake), yangın (fire), trafik kazası (traffic accident), müsilaj (sea saliva), and sağanak (downpour) disaster-related keywords were used. Around 17 thousand samples belonging to five separate sub-datasets with text and image data are created. However, as data may contain meaningless words and symbols, the cleanup is performed afterward. Then, three annotators read each tweet, viewed the related image(s), and independently judged whether each sample was related to the disaster. The majority voting result was applied for the final labeling of each sample.

Note that, as mentioned before, sometimes people share irrelevant tweets during disasters with keywords and hashtags related to the disaster to increase the number of views and shares. Hence, it is crucial to distinguish tweets about the disaster. A sample of the collected relevant and irrelevant tweets (text and image) for different natural disasters are shown in Fig. 2. Overall, the used keywords and the total number of collected samples for each keyword are shown in Table 1.

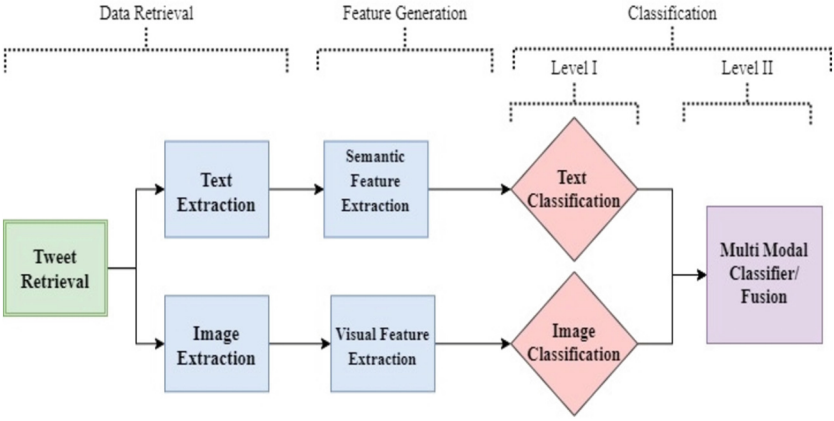

Fig. 1. The workflow of the proposed multi-modal classification system.

Fig. 2. Sample of the collected tweets (images along with their text).

Table 1. The number of samples for each sub-dataset.

Disaster	Datasets	# Relevant Tweets	#Irrelevant Tweets
Deprem	Dataset 1	2596	2400
Yangın	Dataset 2	1450	1448
Trafik Kazası	Dataset 3	1718	1692
Müsilaj	Dataset 4	800	780
Sağanak	Dataset 5	2029	2100
Total number of samples		8593	8420

3.2 The Proposed Text Classification Model

Text classification generally determines whether each sample belongs to one of the predefined classes. In other words, most candidate class is selected for each input. Until the last decade, the text classification process was challenging as computers can only process numbers. This process used traditional techniques, i.e., Bag of word approaches, such as TF-IDF. However, with the recent improvements in text's feature extraction and representation using word embeddings and also the impressive performance of deep learning, we can build not just a text classification system but also other text systems such as document summarization, customer relationship management, web mining, emotion analysis, etc. that can achieve human level. In the following, we summarize the steps of text classification.

Preprocessing
Significantly, the textual data may contain irrelevant and useless terms such as spaces, punctuation, stop words, and repetitive words. Such data will be removed. Also, Turkish has some specific characters, i.e., "ç, ğ, ı, ö, ş, ü" where users in general use the equivalent English character while writing, especially in informal writing. Note that it is an essential preprocessing step to convert back such character to its equivalent Turkish one (this process is named Deascification). Case folding is another preprocessing step. Here, all text is converted to lowercase. Other important text preprocessing are 1) Tokenize: which refers to dividing the string into tokens; 2) Stop word filtering works on eliminating common words. 3) Stemming filtering: Reduces each word to its root by removing prefixes or semed attachments.

Vector Representation of Texts
Fasttext, Word2vec and Glove are successful examples of the first generation of embedding approaches. Although they are easy to develop and use, their main weakness is that each word will always get the exact vector space representation. However, in real life, words can have multiple meanings and may have different meanings and contexts based on their surrounding words. The problem has been overcome by recently developed

embeddings such as BERT, ELMO, and XLNET. In this study, based on our preliminary performance investigations, Word2vec, Glove, and BERT were selected, and their performance was investigated.

Text Classifier

Convolutional Neural Networks (CNN or ConvNet), a category of deep learning neural networks, have performed superhumanly in many areas, such as image recognition, object recognition, automatic video classification, and computer vision.

On the other hand, the recurrent neural network (RNN) is another well-known category of deep learning networks. The output of some of its nodes can affect the subsequent input to the same nodes through cycle connections between nodes. RNNs can use their memory (internal state) to process inputs with different length sequences. Overall, RNN is widely used in Natural language processing (NLP) and can predict the latter word from the former words given in a content. Like traditional neural networks, RNN uses a reverse propagation algorithm. During this backward spread, gradients prone to zero can occur, which is called the reset gradient problem. Long Short-Term Memory Networks (LSTM) can be considered an improved RNN architecture introduced to solve the mentioned problem [16, 17]. LSTM also solves the problems of Memorization-overlap and reset gradient (vanishing gradient), two main problems in deep neural networks applications.

Figure 3 shows the architecture of the LSTM model utilized in this work, which is used to determine whether the text of the input tweet is related to a natural disaster or not.

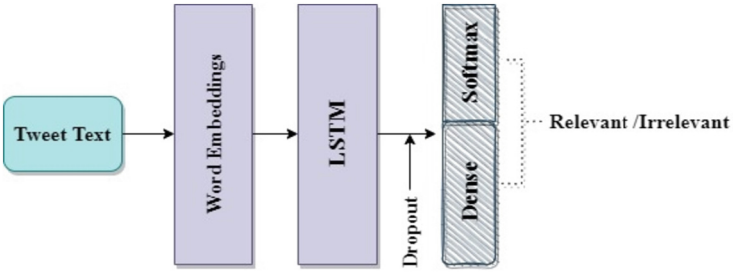

Fig. 3. LSTM network model architecture.

3.3 Image-Based Classification

When people look at an image(s), they can effortlessly differentiate between the color, size, similarity, types of items, etc. When it comes to computers, it is more challenging as computer process the numerical value of the image's pixels. However, recently CNN models have been able to achieve outstanding performance. In general, CNN processes these matrices using some hidden layers that detect the image properties/features to distinguish its objects. Thanks to the proposed multi-model dataset, where each sample was manually annotated to one of the predefined classes (disaster/not a disaster), we

developed and trained a supervised CNN classification model. The main layers of the used model as summarized below.

Input Layer
This layer constitutes the first layer of CNN. While designing the model, selecting the correct input size for the images in this layer is critical and essential. For example, when the selected size is large, it requires more memory, training, and testing time. On the other hand, if the size is small, the network's performance may be low as we may lose the quality of the images while the memory requirement and training time are reduced. In other words, an appropriate input image size is needed for network success regarding network depth and hardware cost when performing image analysis.

Convolution Layer
In convolution, which is a customized linear process, the primary purpose is to extract designating properties for each input image. This layer performs convolution rather than matrix product [18]. In other words, the input image is represented by a matrix of its pixels read in the input layer. The convolution process is to scan this matrix using its filters to extract descriptive properties for images and texts.

Filters generate output data by implementing the convolution process on the previous layer's output. As a result of this convolution process, the activation map is created. Activation maps are regions where characteristics specific to each filter are found. During the training of CNNs, the coefficients of these filters modify with each learning loop in the training set. Thus, the network identifies which input regions are significant as designating properties.

Dropout Layer
CNN sometimes memorizes the data (samples). This layer prevents the network from memorizing and overfitting [19]. The basic logic implemented on this layer is the unloading of some nodes of the network. In other words, dropout works by temporarily ignoring some randomly chosen neurons' incoming and outgoing connections.

Activation Layer
This layer is also known as the rectifier unit (ReLU) layer. The ReLU function generally exchanges the negative input by 0 while it takes its value for positive entries. Hence, all negative values will be replaced by zero. Note that the output of the mathematical operations that are carried out on the convolution layer produces linear results, and this layer will make the deep network a nonlinear structure. With the use of this layer, the network learns faster.

Pooling Layer
Pooling, called "Down Sampling", is usually positioned after the ReLU layer. Its primary purpose is to reduce the size of ReLU output. The processes performed on this layer are working on representing the data in fewer values while still having efficient features. As a result, it will create less transactional load for the following layers and decrease the chance of the model overfitting (memorization).

Like the convolution process, certain filters are defined and applied here. These filters are routed around the image according to a particular step-by-step value and placed in the output matrix by taking the maximum values (maximum pooling) or the average of the values (average pooling). Based on our investigation, maximum pooling was selected as it outperformed others.

Fully Connected Layer
This layer connects all nodes of the previous layer. The entire matrix is given a single class vector with a size of 1 * 1 * 4096. This layer will be followed by another fully connected layer, which is explained below.

Classification Layer (Fully Connected Layer with Softmax)
Classification is carried out in this layer of deep learning models. Based on the previous layer's output, a weight matrix of 4096 * 2 is obtained for the classification layer. The output value of this layer is equal to the most candidate class of the two predefined classes, i.e., disaster and not a disaster. Different functions, such as Softmax and sigmoid, can be used to make the final decision, which is used in this layer. Here, Softmax was favored based on our observations.

4 Experiments

In this section, multiple investigational experiments that demonstrate the performance of proposed classification models vs. some state of art ones are done. First, we perform a comparison when models are used to classify image and text separately (we call it Unimodal Classification) and then when used to classify input samples with text and images (we call it Multimodal Classification). In other words, performance measurements were carried out with three separate classifications: tweet text, tweet image, and tweet text and image together. The performance of the trained models is measured by Accuracy, Precision, Recall, and F1 Score evaluation matrices.

To provide quality and robustness of the achieved results, all samples and sub-datasets presented in this paper were used in the experimental work.

Our experiments were conducted on a machine equipped with the Nvidia GTX1650 GPU with Intel(R) Core (TM) i7-10750H CPU and 16 GB of RAM running the Windows 10 Enterprise operating system. All codes were implemented using Python and its libraries. Google Colab environment is used for implementing the code.

4.1 Unimodal Classification

In this section, Unimodal image and text approaches were trained separately for each natural disaster (deprem-earthquake, yangın-fire, trafik kazası-traffic accident, müsilaj-mucilaj, sağanak-downpour).

Experiment 1. Performance of Image Classification
The main aim of this experiment is to choose the best architecture that will be later adapted and used for multimodal fusion. For the image unimodal experiment, we used

five state-of-the-art CNN architectures. Table 2 reports the results for the tested CNN models. As a result, although all models achieved almost similar performance, "Mouzannar's CNN model1" [20] performed marginally better than other models. It was also the fastest model for training and forecasting.

Table 2. Performance of the selected CNN models when used for image classification (unimodal).

Dataset #	Matrix	CNN Models				
		CNN_Model1 [20]	CNN_Model2 [1]	VGG19	ResNet50	Inception ResnetV2
1st	Acc	0.8548	0.8321	0.7183	0.7333	0.8313
	F1	0.852	0.8309	0.7169	0.7315	0.8316
2nd	Acc	0.8522	0.8295	0.717	0.7319	0.8291
	F1	0.8501	0.8287	0.7157	0.7302	0.8275
3rd	Acc	0.8578	0.8342	0.7219	0.7385	0.8342
	F1	0.8552	0.8321	0.7203	0.7364	0.8327
4th	Acc	0.8601	0.8344	0.7238	0.7407	0.8369
	F1	0.8579	0.8323	0.7217	0.7386	0.8342
5th	Acc	0.8512	0.8301	0.7141	0.7311	0.8301
	F1	0.8479	0.8279	0.7119	0.7381	0.8279

Experiment 2. Performance of Text Classification

In this section, we first compared the performance of the pre-trained Glove, Word2Vec, and BERT word embeddings using the proposed LSTM model and the CNN architecture of [21, 22]. Table 3 summarizes the performance results of CNN and an LSTM model. Overall, Word2Vec outperformed GloVe by around 1%, but BERT marginally outperformed both Word2Vec and GloVe.

Table 3. Performance of the classifiers using multiple embedding approaches.

Classifier	Word Embedding	Accuracy	Precision	Recall	F1-Score
Proposed Model_LSTM	Glove	0.8644	0.8581	0.8704	0.8642
	Word2Vec	0.8721	0.8643	0.8769	0.8701
	BERT	**0.8795**	**0.8702**	**0.8845**	**0.8772**
CNN [21, 22]	Glove	0.8498	0.8421	0.8546	0.8483
	Word2Vec	0.8547	0.8495	0.8592	0.8543
	BERT	**0.8576**	**0.8531**	**0.8607**	**0.8568**

4.2 Multimodal Classification

Experiment 3. Performance of the Late Fusion vs Early Fusion
Based on the results of the above experiments, the developed multimodal consisted of BERT as a text feature extraction system, the LSTM used for text classification, and the CNN model of [20] used for image classification. Then, the late fusion approach is used to make the final decision on the class, whereas, for rule-based, the weighted maximum decision rule is implemented. To finalize the experimental work, we have compared the performance of early fusion and late fusion, as well as the text-only unimodal and image-only unimodal. Overall, as shown in Table 4, it is obvious that late fusion outperforms early fusion. Also, multimodal models provide a further performance improvement compared to both Text-only and image-only unimodal.

Table 4. Performance comparison of different modalities.

Training Modal	Modality	Accuracy	Precision	Recall	F1 Score
Unimodal	Image	85.48	84.87	86.55	85.20
	Text	87.95	87.02	88.45	87.92
Multimodal (Text + Image)	Early Fusion	86.72	85.95	87.50	86.56
	Late Fusion	**91.87**	**90.34**	**92.25**	**91.28**

Experiment 4. Performance of the Developed Model vs Some State of Art Deep Learning Models
In this experiment, the performance of the models of [20] and [23] and the proposed model (LSTM(BERT)–CNN) were compared. Table 5 shows the performance comparisons of the mentioned models. Based on the accuracy, precision, recall, and F1 score, as shown in Table 5, the proposed model has significantly outperformed the others.

Table 5. Performance comparisons of the state-of-art deep learning models.

Models	Datasets	Precision	Recall	F1 Score	Accuracy
CNN (Word2Vec) – CNN [23]	Dataset1	85.93	87.13	86.52	86.47
	Dataset2	84.42	85.82	85.11	85.12
	Dataset3	85.08	86.14	85.60	85.62
	Dataset4	84.63	85.94	85.27	85.44
	Dataset5	85.15	86.36	85.75	85.93
	Average	**85.04**	**86.28**	**85.65**	**85.72**

(*continued*)

Table 5. (*continued*)

Models	Datasets	Precision	Recall	F1 Score	Accuracy
LSTM (Word2Vec) – CNN [20]	Dataset1	90.72	91.95	91.33	91.46
	Dataset2	87.45	88.93	88.18	88.15
	Dataset3	89.77	91.02	90.39	90.38
	Dataset4	89.69	90.94	90.31	90.47
	Dataset5	90.29	91.71	90.99	91.07
	Average	**89.58**	**90.91**	**90.24**	**90.31**
LSTM (BERT) - CNN Proposed Model	Dataset1	90.34	92.25	91.28	91.87
	Dataset2	81.81	90.73	87.80	88.46
	Dataset3	89.12	91.16	88.89	90.87
	Dataset4	89.63	91.28	90.44	91.07
	Dataset5	90.03	92.11	91.05	91.23
	Average	**89.79**	**91.51**	**9.89**	**90.70**

5 Conclusion and Future Works

In this work, we have collected tweets (text) and their related images published before, during, or after some natural disasters. These samples were manually prepared and annotated using three evaluators. Then, a new multimodal classification system was presented after some intensive experimental work to ensure the efficiency and robustness of the proposed model. The late fusion was used to achieve multimodal classification; Also, a pre-trained BERT - LSTM model was used for processing text while a pre-trained CNN model was used for visual modal (images).

The experiment section indicated that the developed multimodal achieved an accuracy of 91.87%, while early fusion achieved 86.72%. Hence, such a model can improve disaster events' classification accuracy and help authorities make the most suitable timely decisions.

As future work, more intensive work on developing a new CNN model for images and using more advanced feature extraction methods for images can be applied. Another direction is to investigate the possibility of supporting multiple languages.

References

1. Alam, F., Ofli, F., Imran, M.: Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. Behav. Inf. Technol. **39**(3), 288–318 (2020)
2. Ponce-López, V., Spataru, C.: Social media data analysis framework for disaster response. Discov. Artif. Intell. **2**(1), 1–14 (2022)
3. Tas, F., Cakir, M.: Nurses' knowledge levels and preparedness for disasters: a systematic review. Int. J. Disast. Risk Reduct. 103230 (2022)
4. Alqaraleh, S., Işik, M.: Efficient Turkish tweet classification system for crisis response. Turk. J. Electr. Eng. Comput. Sci. **28**(6), 3168–3182 (2020)

5. Alqaraleh, S.: Efficient Turkish text classification approach for crisis management systems. Gazi Univ. J. Sci. 1 (2021)
6. Soydan, E., Alpaslan, N.: Medyanin Doğal Afetlerdeki İşlevi. İstanbul J. Soc. Sci. Summer **7**, 53–64 (2014)
7. Rudra, K., Ganguly, N., Goyal, P., Ghosh, S.: Extracting and summarizing situational information from the Twitter social media during disasters. ACM Trans. Web (TWEB) **12**(3), 17 (2018)
8. Madichetty, S., Sridevi, M.: Disaster damage assessment from the tweets using the combination of statistical features and informative words. Soc. Netw. Anal. Min. **9**(1), 42 (2019)
9. Huang, F., Zhang, X., Zhao, Z., Xu, J., Li, Z.: Image–text sentiment analysis via deep multimodal attentive fusion. Knowl.-Based Syst. **167**, 26–37 (2019)
10. Yu, S., Cheng, Y., Xie, L., Luo, Z., Huang, M., Li, S.: A novel recurrent hybrid network for feature fusion in action recognition. J. Vis. Commun. Image Represent. **49**, 192–203 (2017)
11. Guo, D., Zhou, W., Li, H., Wang, M.: Online early-late fusion based on adaptive hmm for sign language recognition. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1), 1–18 (2017)
12. Sosea, T., Sirbu, I., Caragea, C., Caragea, D., Rebedea, T.: Using the image-text relationship to improve multimodal disaster tweet classification. In: The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021) (2021)
13. Zhang, M., Huang, Q., Liu, H.: A multimodal data analysis approach to social media during natural disasters. Sustainability **14**(9), 5536 (2022)
14. Khattar, A., Quadri, S.M.K.: CAMM: cross-attention multimodal classification of disaster-related tweets. IEEE Access **10**, 92889–92902 (2022)
15. Hossain, E., Hoque, M.M., Hoque, E., Islam, M.S.: A deep attentive multimodal learning approach for disaster identification from social media posts. IEEE Access **10**, 46538–46551 (2022)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
18. Rodriguez, R., Gonzalez, C.I., Martinez, G.E., Melin, P.: An improved convolutional neural network based on a parameter modification of the convolution layer. In: Castillo, O., Melin, P. (eds.) Fuzzy Logic Hybrid Extensions of Neural and Optimization Algorithms: Theory and Applications. SCI, vol. 940, pp. 125–147. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68776-2_8
19. Loodos. loodos/bert-base-turkish-uncased hugging face (2022). https://github.com/Loodos/turkish-languagemodels
20. Mouzannar, H., Rizk, Y., Awad, M.: Damage identification in social media posts using multimodal deep learning. In: Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, pp. 529–543 (2018)
21. Chen, Y.: Convolutional neural network for sentence classification. Master's thesis, University of Waterloo (2015)
22. Yoon, K.: Convolutional neural networks for sentence classification [OL]. arXiv Preprint (2014)
23. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 569–576 (2017)