Article

Disaster Image Classification by Fusing Multimodal Social Media Data

Zhiqiang Zou [1,2,*], Hongyu Gan [1], Qunying Huang [3], Tianhui Cai [4] and Kai Cao [5]

[1] College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 1220045001@njupt.edu.cn
[2] Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing 210023, China
[3] Spatial Computing and Data Mining Lab, Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Science Hall, Madison, WI 53706, USA; qhuang46@wisc.edu
[4] College of Liberal Arts & Sciences, University of Illinois at Urbana Champaign, Champaign, IL 61820, USA; tianhui2@illinois.edu
[5] School of Geographic Sciences, East China Normal University, Shanghai 200241, China; kcao@geo.ecnu.edu.cn
* Correspondence: zouzq@njupt.edu.cn

Abstract: Social media datasets have been widely used in disaster assessment and management. When a disaster occurs, many users post messages in a variety of formats, e.g., image and text, on social media platforms. Useful information could be mined from these multimodal data to enable situational awareness and to support decision making during disasters. However, the multimodal data collected from social media contain a lot of irrelevant and misleading content that needs to be filtered out. Existing work has mostly used unimodal methods to classify disaster messages. In other words, these methods treated the image and textual features separately. While a few methods adopted multimodality to deal with the data, their accuracy cannot be guaranteed. This research seamlessly integrates image and text information by developing a multimodal fusion approach to identify useful disaster images collected from social media platforms. In particular, a deep learning method is used to extract the visual features from social media, and a FastText framework is then used to extract the textual features. Next, a novel data fusion model is developed to combine both visual and textual features to classify relevant disaster images. Experiments on a real-world disaster dataset, CrisisMMD, are performed, and the validation results demonstrate that the method consistently and significantly outperforms the previously published state-of-the-art work by over 3%, with a performance improvement from 84.4% to 87.6%.

Keywords: feature extraction; image classification; text classification; FastText; visual geometry group network; multimodal fusion

check for updates

Citation: Zou, Z.; Gan, H.; Huang, Q.; Cai, T.; Cao, K. Disaster Image Classification by Fusing Multimodal Social Media Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 636. https://doi.org/10.3390/ijgi10100636

Academic Editor: Wolfgang Kainz

Received: 29 July 2021
Accepted: 21 September 2021
Published: 24 September 2021

1. Introduction

With the rapid development of the Internet, people are more willing to share their lives and personal experiences on social media (e.g., Twitter, Facebook, and Instagram). Most of the posts on social media are multimodal and include text, images, video, etc. When natural disasters or crises happen, many multimodal messages informing others the condition and situation of the disasters are posted on social media. These posts usually contain critical information such as infrastructure damage, casualties, and help requests. If governments and humanitarian organizations can fully utilize this information from the Internet, they can evaluate and respond to disasters or emergencies more quickly and efficiently. Therefore, a novel method using this multimodal data needs to be further researched in order to discover and classify disaster-related images from social media, which always contain a lot of irrelevant and misleading information [1].

To retrieve valuable information from the data, novel methods are needed. However, most researchers only use unimodal data for analysis [2], and they mainly choose text or image data. Specifically, veracity analysis of social media information [3–7], sentiment analysis [8–12], and the detection of cyberbullying [13,14] mostly use text messages exclusively. For example, Zubiaga et al. [6] introduced a novel method that is able detect rumors by learning from the sequential dynamics of text reports. Sailunazand and Alhajj [12] used the text from Twitter to analyze emotion and sentiment. Chen et al. [14] proposed a new model to detect offensive content and potential offensive users. Furthermore, there have been some studies that have used images to categorize emotions [15]. Rao et al. [16] proposed a multi-level region-based convolutional neural network (CNN) framework to classify image emotion. Recently, it has been proven that systems with multimodal data have better performance than those using solely unimodal data [17]. Multimodal machine learning aims to build models that can process and relate information from multiple modalities. Multimodal machine learning is a vibrant multidisciplinary field of increasing importance and that has extraordinary potential. It has been widely applied in various scenarios, such as audio-visual speech recognition [18], image captioning [19], automatic shot-boundary detection [20], video summarization [21], and social media data mining [22]. For instance, Hodosh et al. [19] proposed a model that is able to combine an image with a natural language sentence to retrieve specific images. Evangelopoulos et al. [21] designed a model fusing the aural, visual, and text streams of videos to create dynamic movie summarization by means of a content-independent algorithm. However, limited multimodal fusion methods integrate images and text information and are applied to classification tasks [17]. Moreover, few researchers have used multimodal data to classify disaster information [23], which fuse images and text messages to gain situational awareness (SA) of natural hazards. However, there is still more than enough room to further improve model architecture and accuracy.

In this research, we adopted a multimodal fusion method to classify disaster images. This method contained three modules: an image feature extractor, a textual feature extractor, and a multimodal fusion module. The image feature extractor extracts image features based on a visual geometry group network(VGG) [24] framework. The textual feature extractor extracts textual features using the FastText [25] framework and by choosing word embedding. In the multimodal fusion module, three fully connected layers and one SoftMax layer were used to complete the final classification task.

To sum up, the main contributions of the present work can be summarized as follows:

- A novel multimodal fusion model was proposed to efficiently extract useful disaster information from massive social media data.
- An optimized model architecture was adopted to process disaster images smaller parameter sizes.
- The accuracy of the disaster image classification on the representative real-world disaster datasets, generated from different disaster events (e.g., earthquakes, and hurricanes), was further improved.
- The code of the project was released to researchers in order to reproduce research and for conducting further research. The code is available at https://github.com/GanHY97/Classification-by-Fusing-Multimodal-Data(accessed on 2 July 2021).

The paper is organized as follows: Section 2 presents previously published work related to this topic; Section 3 shows the datasets and the models; Section 4 summarizes the experiments and the results; Section 5 discusses this study, and the conclusions are presented in Section 6.

2. Related Work

Social media data have been used to conduct various disaster-related studies [26]. Establishing SA is one of the most important purposes of these studies [27]. It is essential for managers to gather disaster-related information as quickly as possible and to categorize it according to their subject. This will facilitate the implementation of different disaster relief methods, such as providing medical services for injured people, repairing damaged roads, and providing relief supplies for victims [26]. According to the data types, the disaster data classification methods can be divided into two categories: (1) using unimodal data and (2) fusing multimodal data. Approaches based on unimodal data only use one type of data for analysis, which can be further divided into two types: text-based and image-based methods. Meanwhile, multimodal data fusion methods usually integrate multiple data for analysis.

Text-based methods collect text data and mine useful information from social media after a disaster, such as negative crowd sentiment, demands for rescue materials, and emergent rescue requirements among victims [26]. Text-based methods are commonly used to analyze information reliability [7], [28], to classify sentiment [29], and to classify content [30–32]. For example, Bai and Yu [30] proposed a method using the distribution representation of words to filter out disaster-related messages from massive and noisy Weibo data and to sort out negative sentiment messages from all of the disaster-related messages. Alternatively, J. Ragini et al. [31] proposed a model to collect disaster data from social media and to classify them according to disaster management needs. Wu et al. [32] designed a model to evaluate rainstorm and flood disaster vulnerability by combining the text data from social media with temporal and spatial data, such as the land use data during the specific time period. Despite the success of text classification in disaster SA, there is a great deal of non-text-based information on social media. For instance, social media users may simply post an image of a flood or damage scenes without any textual description. As such, purely text-based approaches run the risk of losing this useful information. To avoid this situation, researchers have attempted to use image data to obtain disaster SA based on state-of-the-art machine learning, specifically deep learning, methods [33].

With the rapid development of deep learning, image-based classification algorithms have become popular and have been applied in various disaster response systems, including in sentiment classification systems [15] and in content classification [34]. In 2017, Alam et al. [33] presented Image4Act, an end-to-end social media image processing system, to collect, de-noise, and categorize image content posted on social media platforms to help humanitarian organizations gain SA and to launch relief operations. In 2020, Zohaib et al. [35] proposed a visual sentiment model to analyze disaster-related images, in which emotions could be explored from the images. Although the image classification achieved relatively satisfying results, this algorithm only used image data without considering other types of useful information (e.g., text). To address this limitation, researchers then began to use multimodal data for SA establishment and information extraction.

In fact, a great deal of progress has been made in multimodality fusion and mining for social media data processing and analytics. For example, Dao et al. [36] showed a context-aware data-fusion method for disaster image retrieval from social media where the system combined the image with text. In 2019, using CNN, VGG, and long short-term memory(LSTM), Gautam et al. [37] designed models with multimodal data to categorize the information found on Twitter, which could further improve the accuracy of the classification task. In 2020, Ofli et al. [23] exploited both text and image modalities from social media and mined the useful disaster information from them.

To the best of our knowledge, few studies have investigated multimodality methods to analyze the disaster conditions [23]. However, those existing model architectures were either complex, or their classification accuracy was unsatisfied. Therefore, a model with a simple model architecture and high accuracy that also uses multimodality data is needed.

3. Dataset and Models

This work used the CrisisMMD [38] dataset for the experiments and evaluation. Within our data fusion model framework, VGG16 [24] was adopted to classify images, and FastText [25] was chosen to classify the text. Then, we used pretrained VGG16 and FastText to extract features from images and text, respectively. Specifically, the image features were retrieved, and an image classification task was performed by the last two layers of the fully connected layer of the VGG16 network. Meanwhile, the textual features were extracted from the word embeddings that had been trained by FastText. Then, the above two features were fused by means of a concatenating operation. The final classification was accomplished by three fully connected layers and one SoftMax layer. The dataset that we are using is described in detail in Section 3.1. We introduce the multimodal fusion classification model in Section 3.2.

3.1. Dataset

CrisisMMD, a multimodal Twitter dataset with spatiotemporal features, consists of several thousands of manually annotated tweets collected during seven major natural disaster events: Hurricane Irma 2017, Hurricane Harvey 2017, Hurricane Maria 2017, the California Wildfires 2017, the Mexico Earthquake 2017, the Iraq–Iran Border Earthquake 2017, and the Sri Lanka Floods 2017.

The data are annotated with three types of labels that are based on three different classification tasks: (1) Task 1 evaluates whether the data is related to the disaster or humanitarian aid. If the given tweet/image is related, it is considered to be an "Informative" tweet/image or a "Not informative" tweet/image otherwise; (2) Task 2 aims to further divide the "informative" tweet/image into "Affected individuals", "Infrastructure and utility damage", "Injured or dead people", "Missing or found people", "Rescue, volunteering, or donation effort", "Vehicle damage", "Other relevant information", and "Not relevant or can't judge"; and (3) Task 3 assesses the severity of the damage reported/shown in the "Infrastructure and utility damage" images. Damage severity categories include "Severe damage", "Mild damage", "Little or no damage", and "Don't know or can't judge".

The number of images and text messages in the dataset is described in Table 1. Since one tweet may contain one text message and more than one image, the images outnumber the text messages. In the dataset, a pair of a text message and an image could be annotated with different labels because the operations of the annotations are independent. Therefore, we only use the images and text messages with the same label as the experimental data [23]. The filtered dataset for Task 1 is shown in Table 2. For Task 2, we combined similar categories because there were few pairs of text messages and images in these categories. As such, "Affected individuals", "Injured or dead people", and "Missing or found people" were combined (using P to represent it), and "Infrastructure and utility damage" and "Vehicle damage" were combined (with D representing them). Data labeled "Rescue, volunteering or donation effort" were denoted by R. Data labeled "Other relevant information" were represented by O. The combined dataset is as shown in Table 3. The image data, which included the crisis and its severity in Table 4, were used to perform Task 3. "Severe damage", "Mild damage", and "Little or no damage" were represented by S, M, and L respectively. Sample images along with their text and annotations are shown in Table 5.

Table 1. The number of images and text messages in the dataset.

Crisis Name	Images	Text Messages
Hurricane Irma	4504	4021
Hurricane Harvey	4434	3992
Hurricane Maria	4556	3995
California wildfires	1589	1486
Mexico earthquake	1380	1238
Iraq–Iran earthquake	597	496
Sri Lanka floods	1022	830
Total	18,082	16,058

Table 2. The dataset division according to Task 1.

Crisis Name	**Images**		**Text Messages**	
	Informative	**Not Informative**	**Informative**	**Not Informative**
Hurricane Irma	2018	766	1836	678
Hurricane Harvey	2258	906	2082	800
Hurricane Maria	1813	1295	1594	1139
California wildfires	923	282	873	261
Mexico earthquake	806	315	732	285
Iraq–Iran earthquake	398	102	330	83
Sri Lanka floods	229	632	184	527
Total	8445	4298	7631	3773
	12,743		11,404	

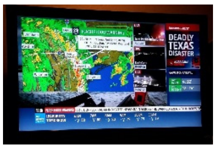
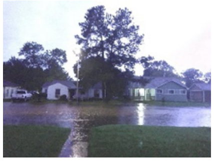
Table 3. The dataset division according to Task 2.

Crisis Name	**Images**				**Text Messages**			
	P	**D**	**R**	**O**	**P**	**D**	**R**	**O**
Hurricane Irma	6	207	214	657	6	174	187	623
Hurricane Harvey	22	233	402	397	21	194	367	385
Hurricane Maria	11	173	276	478	11	141	230	446
California wildfires	8	83	52	96	8	80	47	89
Mexico earthquake	9	37	166	64	9	32	154	62
Iraq–Iran earthquake	28	22	26	51	27	20	17	47
Sri Lanka floods	6	18	56	16	6	15	36	16
Total	90	773	1192	1759	88	656	1038	1668
	3814				3450			

Table 4. The dataset of crisis and its severity for Task 3.

Crisis Name	**Images**		
	S	**M**	**L**
Hurricane Irma	316	229	250
Hurricane Harvey	556	220	116
Hurricane Maria	509	273	80
California wildfires	465	51	15
Mexico earthquake	148	25	5
Iraq–Iran earthquake	158	11	4
Sri Lanka floods	60	30	5
Total	2212	839	475

Table 5. Sample images along with their text and annotations.

			
Astros pummel Harvey in his return, top Mets 12-8	Not always good when your city shows up on a severe weather map. #HurricaneHarvey #ItAintOverYet	Three people, two dogs ride out Hurricane Harvey in 'pod' at Holiday Beach	#HurricaneHarvey Victim Relief-Ways you can help those effected by the storm. Click HERE: https://t.co/m13Lj10an2, accessed on 19 November 2017 https://t.co/lEf3HDxCyQ, accessed on 19 November 2017
Not informative	Informative Other relevant information	Informative Affected individuals	Informative Rescue volunteering or donation effort
			
RT @Nairametrics: Reports suggest Hurricane Harvey cars could be on its way to Nigeria	RT @stephentpaulsen: My street in SE #Houston is now a river. That light is from lightning; it's 10pm #Harvey	RT @worldonalert: #Texas: Photos show destruction in #Bayside after hurricane #Harvey.	The hurricane "Harvey" in the USA: first victims and destructions-RIA Novosti, 8/27/20...
Informative Vehicle damage	Informative Infrastructure and utility damage Little or no damage	Informative Infrastructure and utility damage Mild damage	Informative Infrastructure and utility damage Severe damage

3.2. Model

The framework of the multimodal classification model consists of three modules: an image feature extractor, a textual feature extractor, and multimodal fusion (Figure 1).

3.2.1. Image Feature Extractor

The image feature extractor adopted the VGG16 model [24], which is a CNN. A CNN is mainly composed of convolution layers, a rectified linear unit (ReLU), and max pooling layers. Convolution is a function that maps a tuple of sequences into a sequence of tuples. ReLU is an activation function, and max pooling is a function to reduce input dimensionality. In the image feature extractor module, the parameter size in the second to last layer of its fully connected layer was adjusted to 500 from the original size of 1000. This is because 500 could obtain the same or even better results under the same training epochs in our experiment. For example, in Task 1, which had 30 training epochs, the accuracy of the 500 size was 83.3%, and the accuracy of the 1000 size was 83.1%. Meanwhile, it also reduced the amount of calculations needed in multimodal fusion. The parameter size in the last layer of the fully connected layer, i.e., FC-Num_Classes, was set with respect to the different tasks. For Task 1, the disaster image binary classification task, FC-Num_Classes was set to 2. Transfer learning is an effective method to speed up model convergence [39]. Accordingly, our model directly used the weights from the VGG16 model that had been pretrained on ImageNet by means of transfer learning. In addition, the size of the images in the dataset is not uniform. Therefore, these images all need to be cropped and resized to 224 × 224 before training can begin. The adjusted VGG16 is architecture shown in Table 6.

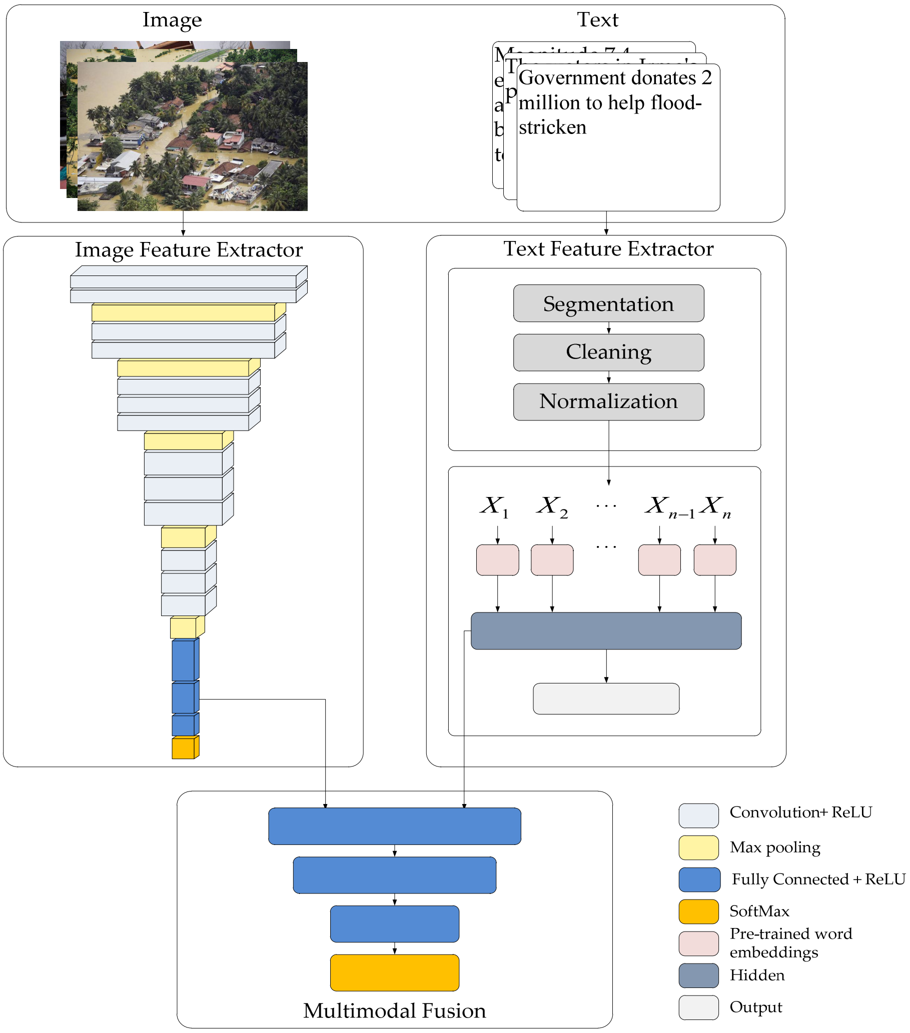


Figure 1. The proposed model framework.

Table 6. The hierarchical architecture of adjusted VGG16 (Num_Classes indicates the number of classes, and FC means fully connected layer).

Layer	Output Size
conv3-64	224 × 224 × 64
conv3-64	224 × 224 × 64
max-pooling	112 × 112 × 64
conv3-128	112 × 112 × 128
conv3-128	112 × 112 × 128
max-pooling	56 × 56 × 128
conv3-256	56 × 56 × 256
conv3-256	56 × 56 × 256
conv3-256	56 × 56 × 256
max-pooling	28 × 28 × 256

Table 6. *Cont.*

Layer	Output Size
conv3-512	28 × 28 × 512
conv3-512	28 × 28 × 512
conv3-512	28 × 28 × 512
max-pooling	14 × 14 × 512
conv3-512	14 × 14 × 512
conv3-512	14 × 14 × 512
conv3-512	14 × 14 × 512
max-pooling	7 × 7 × 512
FC-4096	1 × 1 × 4096
FC-500	1 × 1 × 500
FC-Num_Classes	1 × 1 × Num_Classes
SoftMax	1 × 1 × Num_Classes

3.2.2. Text Feature Extractor

This module extracts text features, including location, temporal, disaster name, and other information from text messages, by using the FastText model [24]. FastText is a model for the efficient learning of word representations and sentence classification. We chose FastText instead of a model based on CNN for two reasons: (1) FastText increases the training and testing speed with similar accuracy and (2) FastText does not need pretrained word embeddings because it can generate word embeddings automatically. Once the text was collected from the Internet, several key steps of the text preprocessing process were necessary, such as segmentation, cleaning, and normalization. The cleaning step includes decapitalization and the removal of stop words and special characters. The normalization step consists of stemming and lemmatization. A sample of text preprocessing is shown in Table 7. Next, the preprocessed text data are used to train the FastText model. The final output of the FastText model, textual features ($TextF[500]$), is the average of the word embeddings ($WordF[500]$) according to Formula (1).

$$TextF[500] = \sum_{i=1}^{num} [WordF_i[0], WordF_i[1], \ldots, WordF_i[499]] / num \tag{1}$$

where $TextF$ represents textual features, 500 is the feature size, $WordF_i[m]$ denotes the i-th word embedding's m-th value, and num is the number of words in the sentence.

Table 7. A sample of text preprocessing.

Text Sample	Segmentation	Cleaning	Normalization
RT @worldonalert: #Texas: Photos show destruction in #Bayside after hurricane #Harvey.	['RT', '@worldonalert', ':', 'Texas', ':', 'Photos', 'show', 'destruction', 'in', 'Bayside', 'after', 'hurricane', 'Harvey']	['texas', 'photos', 'show', 'destruction', 'bayside', 'hurricane', 'harvey']	['texa', 'photo', 'show', 'destruct', 'baysid', 'hurrican', 'harvey']

3.2.3. Multimodal Fusion

Multimodal fusion consists of three fully connected layers and a SoftMax layer. Instead of using complex eigenvector alignment methods [4], we adopted a simple model architecture, i.e., a simple concatenation in series, to combine two 500-dimensional eigenvectors into one 1000-dimensional eigenvector. Then, through the above four network layers, the final prediction results were able to be obtained. The overall model architecture is shown in Table 8.

Table 8. Multimodal fusion model architecture (Num_Classes means the number of classes, and FC means fully connected layer).

Layer	Output Size
FC	1 × 1 × 1000
FC	1 × 1 × 500
FC	1 × 1 × Num_Classes
SoftMax	1 × 1 × Num_Classes

4. Experiments and Results

The experiments include three parts and are according to the data labels categories that correspond to three tasks described in Section 3.1. Each task was completed based on the previous one, which means that the input data of each task was filtered by the previous task. Note that the procedure of Task 2 was also redesigned and improved in this work. Based on our manual examination, most of the images labeled as "Other relevant information" were satellite images, weather maps and news reports, etc., which were quite different from the others. Therefore, Task 2 was further divided into two steps: (1) step one was a binary classification task that selected images and text not labeled as "Other Relevant Information". In this step, information from other categories was combined as one category; (2) step two was a tri-categorization task that divided the data into three categories that were merged in step one.

For each of these tasks, three classification experiments were performed, where the models were trained by (i) only text, (ii) only images, and (iii) both text and images. To evaluate the performance of the trained models, the well-known metrics that were used in Ofli's work [23], such as accuracy, precision, recall, and F1-score, were adopted. All of these metrics used the weighted-average method and are calculated as follows:

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \\ Precision = \sum_{i=1}^{N} \frac{TP_i}{TP_i+FP_i} \cdot \frac{Num_i}{ALL} \\ Recall = \sum_{i=1}^{N} \frac{TP_i}{TP_i+FP_i} \cdot \frac{Num_i}{ALL} \\ F1-score = \frac{2 \times precision \times recall}{precision + recall} \end{cases} \tag{2}$$

where *Accuracy* measures the proportion of the correctly labeled samples among all of the data, *Precision* measures the proportion of the truly positive samples among the predicted positive samples, and *Recall* is the proportion of the correct positive samples among the positive samples belong to this category in the real world. TP, TN, FP, and FN means true positive, true negative, false positive, and false negative, respectively. If a metric includes a subscript (e.g., TP_i), it measures the performance of a certain data category. N is the number of data categories. ALL is the number of all of the samples. Num_i is the number of samples in the i-th category.

In this work, the training set, the verification set, and the test set were divided with a ratio of 70:15:15. In order to avoid a similar situation to that of the training set, which mostly consisted of earthquake data, even though the majority of the verification and test set is composed of other different disasters, we divided the datasets into seven major natural disaster events. To ensure that the experimental data were balanced, we divided the data starting from the bottom task. For example, the damage severity assessment dataset in Task 3 (S, M, L) is the data subset of Task 2 "Infrastructure and Utility Damage" (D), as shown in Tables 3 and 4. The experimental results of the three tasks are shown in Table 9. The classification results of a sample with images, text, and annotations are shown in Table 10.

Table 9. The results of experiments.

Task		Models	Accuracy	Precision	Recall	F1-Score
Task 1		Only Text	0.852	0.863	0.852	0.858
		Only Images	0.833	0.831	0.833	0.832
		Text and Images	0.876	0.875	0.876	0.875
Task 2	Step 1	Only Text	0.907	0.908	0.906	0.907
		Only Images	0.922	0.922	0.922	0.922
		Text and Images	0.926	0.927	0.926	0.926
	Step 2	Only Text	0.922	0.922	0.920	0.918
		Only Images	0.885	0.847	0.885	0.864
		Text and Images	0.9125	0.872	0.911	0.891
Task 3		Images	0.689	0.663	0.669	0.670

Table 10. The classification results of a sample with images, text, and annotations (OT: Only Text, OI: Only Images, TI: Text and Images).

Sample	Task		Model	Classification	Annotation
	Task1		OT	Informative	Informative
			OI	Informative	
			TI	Informative	
RT @worldonalert: #Texas: Photos show destruction in #Bayside after hurricane #Harvey.	Task2	Step 1	OT	P+D+R	D
			OI	P+D+R	
			TI	P+D+R	
		Step 2	OT	D	
			OI	D	
			TI	D	
	Task 3		TI	Mild damage	Mild damage

In Task 1, when only using image data or text data, the accuracy was 83.3% and 85.2%, respectively. Using the multimodal method, the accuracy was 87.6%. Clearly, in Task 1, the multimodal approach works better than the unimodel one.

Task 2 divided the experiment into two steps. In the first step, the instances labeled “Other relevant information” were eliminated according to the previous analysis. Using only text and images, the accuracy was 90.7% and 92.2%, respectively. However, the accuracy reached 92.6% by applying the proposed multimodal method, which demonstrates that the multimodal method shows improvement over the unimodal method. In the second step, we further classified the information into the categories P, D, and R. Using text or images only, the accuracies were 92.2% and 88.5%, respectively. The accuracy of the multimodal method reached 91.2%, which was only better than the model using images exclusively. Although the accuracy of the method when using text only outperformed our method by 0.95%, using text on its own lacks reliability and intuitiveness. Most work on disaster assessment focuses on using images since images are reliable and intuitive.

In Task 3, our model obtained a relatively low accuracy (68.9%) since that the test samples were very unbalanced due to the lack of data labeled “L”. In fact, the accuracy of Task 3 could be further improved by increasing the number of samples. We acquired more data from AIDR (https://crisisnlp.qcri.org/, accessed on 1 October 2020, RESOURCE # 9), and the Task 3 experiment was repeated. The accuracy of this experiment was improved from 68.9% to 79.6% when the number of data labeled as L was increased from 475 to 1553. However, our target was to explore a method that is able to complete damage assessment tasks with high accuracy with minimal data. For example, generative adversarial networks (GAN) can be explored to assess the severity of disaster damage [40]. In future work, we will focus on addressing this open problem.

5. Discussion

This work only used a portion of the CrisisMMD dataset, where the image labels were the same as the text labels. As shown in Table 9, the method using multimodality had better results than the method using unimodality. Compared to the model designed by Gautam [37] and Ofli [23], the architecture of the model designed in this paper was simpler and easier to train. Specifically, in the image feature extractor module, we adjusted the parameter size in the second to last layer of its fully connected layer to 500 from the original size of 1000, which means that the image feature is simpler and that the multimodal fusion input is simpler as well. In the text feature extractor module, we utilized the FastText model instead of a common model based on CNN, and we analyzed the advantages of the FastText model in Section 3.2.2. Furthermore, the results here were also better than those from the aforementioned methods. Specifically, this multimodal method achieves about 3% higher accuracy than Ofli's method does [23] in Task 1 due to the following advantages: (1) the accuracy of our text processing module is better than that of Ofli's method [23], and (2) a simpler series concatenation is explored to fuse the above two modules.

In Task 1, the experimental results and procedures were in line with our expectations. In other words, the proposed model had better accuracy than the methods using a unimodal model, and the model accuracy trend steadily increased with the increasing number of epochs used in the model training process. The training process of the multimodal model using under 100 epochs is shown in Figure 2. At the beginning of the training process, the accuracy of the model was approximately equal to the only images model. Additionally, the process steadily improved and reached a peak at about 80 epochs. We further analyzed the performance of the three models (i.e., only text, only images, and text + images) by examining their confusion matrices, which contain some specific cases, such as 1925 pairs of text messages and images. The confusion matrix of these three models on Task 1 is shown in Table 11. From these three confusion matrices, it can be seen that the only text model missed 138 useful instances and that the only images model missed 137 useful instances. Meanwhile, the multimodal approach only missed 86 useful instances. Obviously, many less instances were missed when using our multimodal approach.

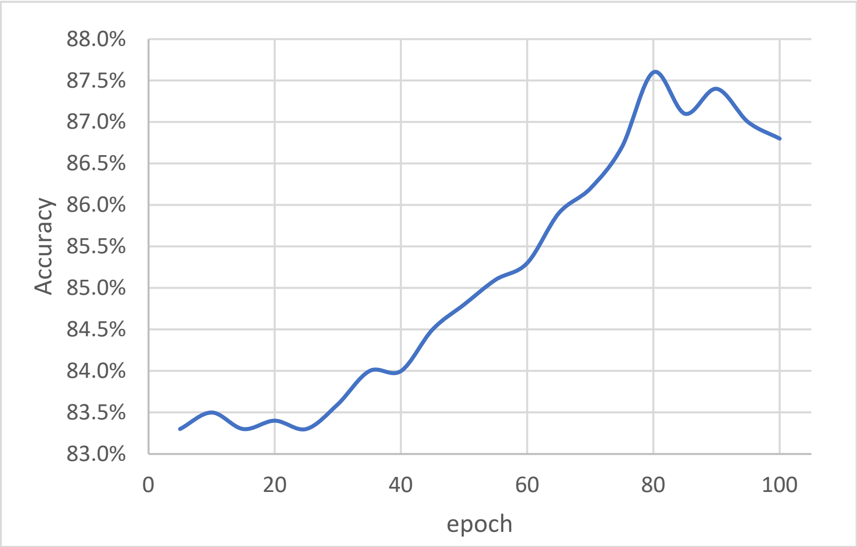


Figure 2. Accuracy of the proposed multimodal model for Task 1.

Table 11. Confusion matrix of the three models on Task 1. Inf represents informative. Not-Inf represents not informative.

Data	Label	Predicted	
		Inf	**Not-Inf**
Only Text	Inf Not-Inf	737 58	138 396
Only Images	Inf Not-Inf	1135 183	137 470
Text + Images	Inf Not-Inf	1186 151	86 502

The first step of Task 2, the training process, is shown in Figure 3. Overall, the accuracy fluctuated between 91.9% and 92.6%. This is because that some training samples were mistakenly classified, as shown in the confusion matrices (see Table 12). We further analyzed the confusion matrices and found that there were 26 instances missed when using only text, and there were 22 instances missed with the images only model. In contrast, there were only 20 instances missed when using our multimodal model.

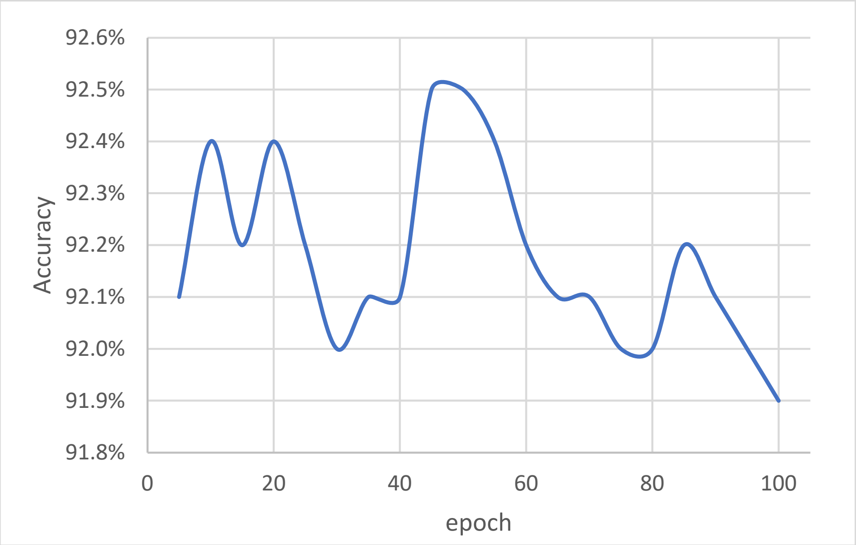


Figure 3. The accuracy of the proposed multimodal model in the first step of Task 2.

Table 12. Confusion matrix of the three models on Task 2.

Data	Label	Predicted	
		P + D + R	**O**
Only Text	P + D + R O	248 22	26 222
Only Image	P + D + R O	252 18	22 226
Text + Image	P + D + R O	254 18	20 226

Figure 4 shows the second part step of Task 2, which was the training process of our multimodal model. We found that before epoch 60, the accuracy increased steadily and that it had some fluctuation after 60 epochs. Therefore, we set the number of training process epochs to 60. In addition, although the classification accuracy for P set was relatively low, the overall accuracy of our model was reasonably good, achieving 91.2% accuracy.

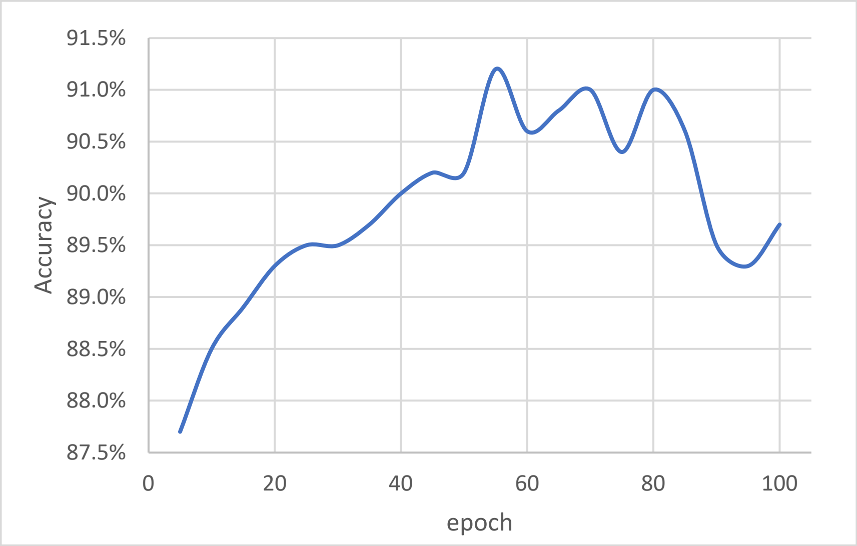


Figure 4. The accuracy of the multimodal model in the second step of Task 2.

Figure 4 depicts the second step of Task 2, which was the training process of our multimodal model. We found that before epoch 60, the accuracy increased steadily and that had some fluctuation after 60 epochs. Therefore, we set the number of training process epochs to 60. In addition, although the classification accuracy for P set was relatively low, the overall accuracy of our model was reasonably good, achieving 91.2% accuracy.

It can be seen from the above discussion that a multimodal machine learning method can improve the classification accuracy of disaster images, and the classified data can be applied to solve the problems of disaster assessment and management, such as SA [26]. Specifically, in Task 1, we extracted disaster-related information in order to generate a general awareness of a disaster situation. In Task 2, we divided the relevant information into P, D, and R. The data labeled as P contribute to providing medical services for injured people and relief supplies for victims while the data labeled as D can be used for disaster assessment, and the data labeled as R can help communities to better prepare for the dispatch of relief supplies. In Task 3, the data labeled as D can be explored to assess damage severity, helping managers respond to and recover from crises.

6. Conclusions

When disasters occur, victims post a large volume of messages about what they have experienced and have witnessed on social media. For relief workers, sorting out relevant information from these massive social media data can help them in assessing disaster severity. The data on social media are often massive and messy, which means that a method is needed to retrieve disaster-relevant information. However, most of the previous studies surrounding this topic have focused on either text analysis or image analysis, and few studies have used multimodal methods. Even when using multimodal approaches, their classification accuracies are not very satisfactory. In this research, we proposed a multimodal method for disaster image classification. Specifically, the deep learning method was used to extract the image features, and the text features were integrated simultaneously for the classification tasks. The experimental results on real disaster datasets demonstrate the effectiveness of the proposed method. In general, the proposed multimodal approach shows better performance than the unimodal one and achieves higher accuracy than the existing multimodal approach [23]. In addition, the architecture of the model is simpler to train.

Although geospatial data have been used in the text feature extractor module, our task can be further optimized by mining spatio-temporal information. Specifically, the data subset contains some geographic information, such as longitude and latitude in images as well as street information in text. In Task 2, the places where medical staff and relief supplies are needed can be marked on the map. In Task 3, the damage severity of different regions can be better assessed with the help of spatio-temporal information.

In the future, there are three possible directions for further improvements. The first direction is to annotate classification results with spatio-temporal information. The second possible direction is to address the problem of unbalanced data. We noticed that unbalanced samples resulted in low accuracy in Task 3. Subsequent work will concentrate on finding an appropriate approach to handle this problem. The third direction could be data fusion. Besides the method used in this work, many methods can be applied to fuse multimodal data, and we will further examine novel deep learning methods to integrate social media multimodal data in the future.

Author Contributions: Conceptualization, Zhiqiang Zou and Qunying Huang; methodology, Zhiqiang Zou and Hongyu Gan; software, Zhiqiang Zou, Hongyu Gan, and Tianhui Cai; validation, Zhiqiang Zou, Hongyu Gan, Qunying Huang, and Tianhui Cai; formal analysis, Zhiqiang Zou and Hongyu Gan; investigation, Zhiqiang Zou, Hongyu Gan, and Tianhui Cai; resources, Zhiqiang Zou; data curation, Hongyu Gan; writing—original draft preparation, Zhiqiang Zou, Hongyu Gan, and Tianhui Cai; writing—review and editing, Zhiqiang Zou, Qunying Huang and Kai Cao; visualization, Hongyu Gan and Tianhui Cai; supervision, Qunying Huang; project administration, Zhiqiang Zou and Qunying Huang; funding acquisition, Zhiqiang Zou and Qunying Huang. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDA23100100), the Chinese Scholarship Council (grant number 202008320044), the National Natural Science Foundation of China (grant number 42050101), the Vilas Associates Competition Award from the University of Wisconsin–Madison (UW-Madison), and Microsoft AI for Earth. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Microsoft or UW–Madison.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [https://crisisnlp.qcri.org/crisismmd (accessed on 1 October 2020)].

Acknowledgments: The authors would like to thank Nanjing University of Posts and Telecommunications and University of Wisconsin–Madison. Meanwhile, we thank the editors and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, Q.; Xiao, Y. Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1549–1568. [CrossRef]
2. Kumar, A.; Sangwan, S.R.; Nayyar, A. *Multimedia Social Big Data: Mining*; Springer: Singapore, 2020; ISBN 9789811387593.
3. Liu, X.; Kar, B.; Zhang, C.; Cochran, D.M. Assessing relevance of tweets for risk communication. *Int. J. Digit. Earth* **2019**, *12*, 781–801. [CrossRef]
4. Zhao, Z.; Resnick, P.; Mei, Q. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts Categories and Subject Descriptors Detection Problems in Social Media. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 1395–1405.
5. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [CrossRef] [PubMed]
6. Zubiaga, A.; Liakata, M.; Procter, R. Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. *arXiv* **2016**, arXiv:1610.07363.
7. Mendoza, M.; Poblete, B.; Castillo, C. Twitter under crisis: Can we trust what we RT? In Proceedings of the First Workshop on Social Media Analytics—SOMA'10, Washington, DC, USA, 25–28 July 2010; pp. 71–79.
8. Burnap, P.; Williams, M.L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **2015**, *7*, 223–242. [CrossRef]
9. Singh, A.; Shukla, N.; Mishra, N. Social media data analytics to improve supply chain management in food industries. *Transp. Res. Part E Logist. Transp. Rev.* **2018**, *114*, 398–415. [CrossRef]

10. Oliveira, N.; Cortez, P.; Areal, N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Syst. Appl.* **2017**, *73*, 125–144. [CrossRef]
11. Ganesan, T.; Anuradha, S.; Harika, A.; Nikitha, N.; Nalajala, S. Analyzing Social Media Data for Better Understanding Students' Learning Experiences. *Lect. Notes Data Eng. Commun. Technol.* **2021**, *57*, 523–533. [CrossRef]
12. Sailunaz, K.; Alhajj, R. Emotion and sentiment analysis from Twitter text. *J. Comput. Sci.* **2019**, *36*, 101003. [CrossRef]
13. Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access* **2019**, *7*, 70701–70718. [CrossRef]
14. Chen, Y.; Zhou, Y.; Zhu, S.; Xu, H. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 71–80.
15. Zhao, S.; Ding, G.; Huang, Q.; Chua, T.S.; Schuller, B.W.; Keutzer, K. Affective image content analysis: A comprehensive survey. *IJCAI Int. Jt. Conf. Artif. Intell.* **2018**, *2018*, 5534–5541. [CrossRef]
16. Rao, T.; Li, X.; Zhang, H.; Xu, M. Multi-level region-based Convolutional Neural Network for image emotion classification. *Neurocomputing* **2019**, *333*, 429–439. [CrossRef]
17. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 423–443. [CrossRef] [PubMed]
18. Yuhas, B.P.; Goldstein, M.H.; Sejnowski, T.J. Integration of Acoustic and Visual Speech Signals Using Neural Networks. *IEEE Commun. Mag.* **1989**, *27*, 65–71. [CrossRef]
19. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *IJCAI Int. Jt. Conf. Artif. Intell.* **2015**, *2015*, 4188–4192. [CrossRef]
20. Lienhart, R. Comparison of Automatic Shot Boundary Detection Algorithms. *Event Electron. Imaging* **1999**, *3656*, 290–301.
21. Evangelopoulos, G.; Zlatintsi, A.; Potamianos, A.; Maragos, P.; Rapantzikos, K.; Skoumas, G.; Avrithis, Y. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimed.* **2013**, *15*, 1553–1568. [CrossRef]
22. Zou, Z.; He, X.; Zhu, A. An Automatic Annotation Method for Discovering Semantic Information of Geographical Locations from Location-Based Social Networks. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 487. [CrossRef]
23. Ofli, F.; Alam, F.; Imran, M. Analysis of Social Media Data Using Multimodal Deep Learning for Disaster Response. *arXiv* **2020**, arXiv:2004.11838.
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
25. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 2, pp. 427–431.
26. Vongkusolkit, J.; Huang, Q. Situational awareness extraction: A comprehensive review of social media data classification during natural hazards. *Ann. GIS* **2021**, *27*, 5–28. [CrossRef]
27. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing Social Media Messages in Mass Emergency: Survey Summary. In Proceedings of the Companion of the The Web Conference 2018 on The Web Conference 2018—WWW'18, Lyon, France, 23–27 April 2018; Volume 2, pp. 507–511.
28. Liu, X.; Kar, B.; Montiel Ishino, F.A.; Zhang, C.; Williams, F. Assessing the Reliability of Relevant Tweets and Validation Using Manual and Automatic Approaches for Flood Risk Communication. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 532. [CrossRef] [PubMed]
29. Yang, T.; Xie, J.; Li, G.; Mou, N.; Li, Z.; Tian, C.; Zhao, J. Social Media Big Data Mining and Spatio-Temporal Analysis on Public Emotions for Disaster Mitigation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 29. [CrossRef]
30. Bai, H.; Yu, G. A Weibo-based approach to disaster informatics: Incidents monitor in post-disaster situation via Weibo text negative sentiment analysis. *Nat. Hazards* **2016**, *83*, 1177–1196. [CrossRef]
31. Ragini, J.R.; Anand, P.M.R.; Bhaskar, V. Big data analytics for disaster response and recovery through sentiment analysis. *Int. J. Inf. Manag.* **2018**, *42*, 13–24. [CrossRef]
32. Wu, Z.; Shen, Y.; Wang, H. Assessing Urban Areas' Vulnerability to Flood Disaster Based on Text Data: A Case Study in Zhengzhou City. *Sustainability* **2019**, *11*, 4548. [CrossRef]
33. Alam, F.; Imran, M.; Ofli, F. Image4Act: Online social media image processing for disaster response. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, 31 July–3 August 2017; pp. 601–604.
34. Said, N.; Ahmad, K.; Riegler, M.; Pogorelov, K.; Hassan, L.; Ahmad, N.; Conci, N. Natural disasters detection in social media and satellite imagery: A survey. *Multimed. Tools Appl.* **2019**, *78*, 31267–31302. [CrossRef]
35. Hassan, S.Z.; Ahmad, K.; Hicks, S.; Halvorsen, P.; Al-Fuqaha, A.; Conci, N.; Riegler, M. Visual Sentiment Analysis from Disaster Images in Social Media. *arXiv* **2020**, arXiv:2009.03051.

36. Dao, M.-S.; Quang Nhat Minh, P.; Kasem, A.; Haja Nazmudeen, M.S. A Context-Aware Late-Fusion Approach for Disaster Image Retrieval from Social Media. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 266–273.
37. Gautam, A.K.; Misra, L.; Kumar, A.; Misra, K.; Aggarwal, S.; Shah, R.R. Multimodal Analysis of Disaster Tweets. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 94–103.
38. Alam, F.; Ofli, F.; Imran, M. CrisisMMD: Multimodal twitter datasets from natural disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media, ICWSM 2018, Palo Alto, CA, USA, 25–28 June 2018; pp. 465–473.
39. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
40. Tilon, S.; Nex, F.; Kerle, N.; Vosselman, G. Post-Disaster Building Damage Detection from Earth Observation Imagery Using Unsupervised and Transferable Anomaly Detecting Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 4193. [CrossRef]