

Hierarchical Deep Multi-Modal Neural Network for Classification of Small Scale Incidents in Microblogs

*Sako, DJS, Bennett, EO and Igiri, CG

Department of Computer Science, Rivers State University, Port Harcourt, Nigeria.

*Correspondence: dum.sako@ust.edu.ng

DOI: 10.56201/ijcsmt.v9.no1.2023.pg36.51

Abstract

Data imbalance is defined by great differences in the distribution of the classes in the dataset and is predominant and inherent in the real world. Addressing imbalanced data distribution is a difficult task for many classification algorithms as algorithms do not learn properly when a massive difference in size between data classes exist. This classification problem exists in many real world application domains, one of which is emergency situation management where we identify incident-related informative content from the Twitter streams especially for smaller scale incidents where there are only small bits of information. In this paper, we present a two-stage hierarchical multimodal deep learning neural network model (2HDLnet) to identify incident-related informative content from social media. This top-down level-based classification method entails the temporary regrouping and re-labelling of classes of the training data toward a more balanced distribution and performing hierarchical data classification using stacks of deep learning architectures to allow both overall and specialized learning at each level of the data hierarchy. We train separate CNN models for text, image and multimodal approaches at each level and combine the models at both levels to perform hierarchical classification. We consider the use of message contents from Twitter microblogging site consisting of texts and images in which they exist together or separately. The experimental analysis on a home-grown incident dataset demonstrates that the proposed approach can effectively classify crisis images and/or text tweets at each logical layer.

Keywords: Deep learning, Data imbalance, Hierarchical classification; Emergency situation, Multimodal Learning

1 Introduction

Addressing imbalanced data distribution is a difficult task for many classification algorithms as algorithms do not learn properly when a massive difference in size between data classes exist (Bader-El-Den et al, 2016). This classification problem is predominant and inherent in many real world application domains, one of which is emergency situation management and reporting where we identify incident-related informative content from the Twitter streams especially for small scale events where there are only small bits of information. Dataset that suffers from a high class imbalance can be made more useful by organizing the document into the correct categories and sub-categories to reflect their similarities.

Research in recent years has uncovered the increasingly importance of social media during emergency situations and shown that information broadcast via social media can enhance situational awareness during a crisis situation (Vieweg et al. 2010; Alsaedi et al. 2015; Vieweg et al. 2014; Alsaedi et al. 2017). During emergency, people start posting tweets containing texts, images, and videos as soon as the incident occurs in an area. The

analysis of these incident-related tweet texts, images, and videos can help emergency response organizations in better decision-making and prioritizing their tasks. Current analyses for the use of social media in emergency management mostly focus on detecting large scale incidents in microblogs (Kumar et al. 2020; Vieweg et al. 2010, Sakaki et al. 2013). The number of information shared on social media platforms is usually high, because many people might be affected. In contrast, the absolute amount of information on smaller scale incidents, like car crashes or fires, is comparably low (Schulz and Ristoski, 2013). Finding the informative contents out of the massive volume of Twitter content for personal or small scale incidents is much more difficult (Kumar et al. 2020) with only a dozen of postings available.

Also, many real-world applications do not involve only one data modality. Social network post and message contents could consist of either text only, image only or images along with corresponding textual descriptions. Limiting oneself to using only one modality would involve losing all the information contained in the others. As humans, we perceive our environment in a multimodal way and would often need those multiple complementary information sources, where they exist, to fully understand the social media posts and messages. However, successfully integrating heterogeneous information streams (two or more views of one data point that belong to very different spaces) has become a challenging task.

Therefore, this paper aims to propose a two-stage hierarchical deep learning based model (2HDLnet) that combines the CNN–CNN networks to automatically classify images and/or text into target categories and sub-categories. This two-phased data level classification approach entails the temporary regrouping and re-labelling of classes of the training data toward a more balanced distribution and performing hierarchical data classification using stacks of deep learning architectures to allow both overall and specialized learning by level of the data hierarchy. The incident types categories are temporarily regrouped and relabeled as incident-related in the first stage of the classification to balance it up with the more populated non-incident class. In the second stage, we then have a specialized learning of the different target incident types/categories considered in this study. The hierarchical classification scheme breaks down incident/disaster-related categories into tree structure which organizes general to specific categories using a pre-defined hierarchy.

The contributions of this study are summarized as follows:

- i) Developing approach that entails the temporary regrouping and re-labelling of classes of the training data toward a more balanced distribution.
- ii) Hierarchical classification of multimodal document
- iii) A detailed experimental analysis is provided in terms of accuracy to measure the performance of the proposed system.

The paper is structured as follows: Section 2 reviews recent scholarly works related to this study. A description of the proposed approach is presented in section 3. Section 4 describes the datasets used in carrying out the experiments including the experimental setup and configuration. Section 5 describes the results and discussion of the undertaken experiments while section 6 concludes the paper.

2 Related Literature

In research closely related to this study, Kowsari et al (2017) proposed an approach to hierarchical document classification, called Hierarchical Deep Learning for Text classification (HDLTex), that combines multiple deep learning approaches to produce hierarchical classifications of textual documents obtained from the Web of Science.

Salakhutdinov et al. (2013) used deep learning to hierarchically categorize images. At the top level the images are labeled as animals or vehicles. The next level then classifies the kind of animal or vehicle. Huang et al (2012) also proposed a method for learning hierarchical representations for face verification with convolutional deep belief networks.

In solving data imbalance problem, Mohamed Bader-El-Den et al (2016) proposed a two-phased data-level approach named TempC, which introduces the temporary re-labelling of classes aimed at reducing the level of class imbalance and also to identify and treat difficult areas of a dataset separately. In the original training set, the minority examples and their k nearest majority neighbours are given a new class label which serves as the new minority class. A classification model is built which is used to classify unlabelled instances in the test set. The derived minority class which captures the difficult areas is used as the training set in the 2nd phase. Another classification model is built and the test instances that were predicted as belonging to the minority class in the 1st phase are used as test set. An approach which also re-labels instances is the SPIDER (Selective Pre-processing for Imbalanced Data) method proposed by Stefanowski and Wilk (2008). It systematically removes or re-labels majority class examples while difficult instances from the minority class are amplified.

This paper considers newer methods of machine learning for multimodal (text and image) classification taken from deep learning. Deep learning, a sub-field of machine learning, is based on a set of algorithms that attempt to model high level abstractions in data (Shridhar, 2017), with successive layers of the increasingly representations. It is based on artificial neural networks with representation learning. Deep learning is an efficient version of neural networks that, like conventional machine learning, can be supervised, semi-supervised or unsupervised (Bengio et al., 2013). The machine learning algorithms are based on learning multiple levels of representation/abstraction. It is the implementation of neural networks with more than a single hidden layer of neurons (Mayo, 2016). Over the last few years, Artificial Neural Networks (ANNs) and Deep Learning have achieved state-of-the-art performances that have seen their use become widespread in many domains (Zhou et al., 2017; Krizhevsky et al, 2012; Leung et al, 2014; Haj-Yahia, 2018). They have shown great improvement in recognizing and classifying objects and images and have changed the way we handle data in computational data analytics (Nooka, 2016).

Deep Learning Neural Networks, like Convolutional Neural Networks (CNNs), as generic feature extractors, have been continuously improving the image classification accuracy, avoiding the traditional hand-crafted feature extraction techniques in image classification problems; hand-crafted features are usually low-level features without enough mid-level and high-level information, which hinders the performance of the system (Yim et al. 2015). Deep architectures with hierarchical frameworks enable the representation of complex concepts with fewer nodes than shallow architectures (Nooka, 2016). Unlike traditional statistical approaches where humans were required to study the data carefully and design useful features, deep learning places the machine in charge of learning useful features (or representations) directly from the data without human intervention (LeCun et al. 2015). Deep learning typically uses high-capacity neural network models, which are trained by a large number of training samples (Choi, 2018).

The convolutional neural network (CNN) is a class of deep learning neural networks. Convolutional Neural Networks (CNNs or ConvNets) are a specialized kind of neural network for processing data that has a known grid-like structure (Goodfellow et al., 2016; Eskesen, 2017). They are modeled after the architecture of the visual cortex where neurons are not fully connected but are spatially distinct (LeCun et al 1998; and provide excellent results in generalizing the classification of objects in images (Oquab et al. 2014; Peters and Albuquerque, 2015; Daly and Thom, 2016; Lagerstrom et al. 2016, Alam et al. 2017). More

recent work has used CNNs for text mining and classification (Lee and Dernoncourt, 2016; Verma et al. 2011, Cameron et al. 2012, Abel et al.2012; Chowdhury et al. 2013, Makino et al. 2018).

In research closely related to the work in this paper, (Mouzannar et al. 2018; Zahavy et al. 2016, Wang et al. 2017, Kiela et al. 2018, Rizk et al. 2019, Offi et al. 2020) proposed multimodal deep learning classification frameworks where one modality is discrete, e.g. text, and the other is continuous, e.g. visual representations using convolutional neural networks.

3 The Proposed Model

In this work, we propose a two-stage Hierarchical multi-modal Deep Learning neural network for Data Classification (2HDLnet). 2HDLnet employs a stack of deep learning architectures to provide specialized understanding at each level of the data hierarchy. The hierarchical classification scheme breaks down incident/disaster-related categories into tree structure which organizes general to specific categories. classifiers are built for each category. The hierarchy tree structure is shown in figure 1. The top-down category tree (Level I) classifies tweets into two classes/categories: I. INCIDENT and II. non-INCIDENT. These are the parent-level models. The tweets are further classified into a sub-category of specific incident types (Level II) (child-level models) if the tweets are classified as INCIDENT-related in Level I. They include *accident, crime and civil disorder, damaged infrastructure, fire and flood*. Figure 2 shows the overall structure of the 2HDLnet. At level I, classification task is done in order to determine whether or not the tweet is incident-related. If a tweet is incident-related, classification of the tweet is carried out at level II to determine which category of incidents it belongs to.

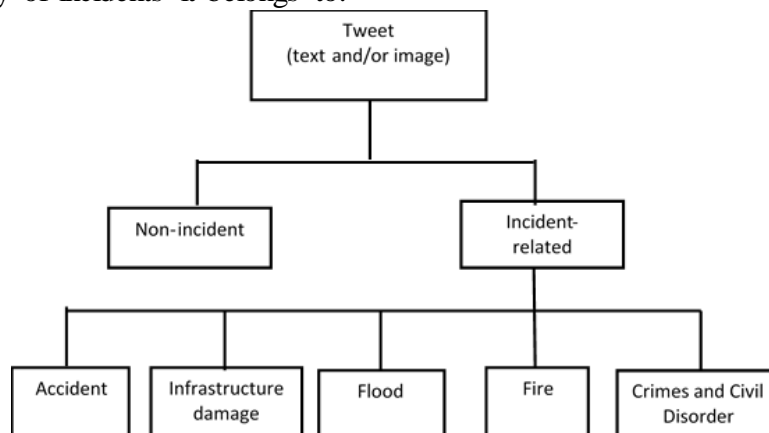


Figure 1. Hierarchical Structure of the data classification groups

Once a tweet, containing textual and/or visual components, is retrieved, the text and image are extracted and processed independently and in a different pipe. Textual and visual features extracted are used to train the classifiers (text, image and multimodal classifiers) that run in parallel. Multi-modal learning is partitioned between feature-level fusion techniques and decision-level fusion techniques. For a feature-level multimodal classifier, after a certain depth, the pipes are concatenated as a shared representation of image and text modalities that preserves their cross-modality correlation (Wang et al. 2017) followed by multi-modal layers. Each modality, including the feature-level multimodal, is processed in a different pipe and gives a prediction from the classifier trained on them. At the decision-level multimodal learning, a policy network is learned to decide which of the three classifiers to use. This is shown in Figure 3, Left. Basically, the architecture is composed of a combination of dual path

CNNs to simultaneously and/or independently learn visual and textual representations in an end-to-end fashion, consisting of a deep image CNN for image input, deep text CNN for text input and joint model to learn joint/shared representation of the two modalities. They are forged together by a policy network in each level of the classification hierarchy, as can be seen in Figure 3, Right. Classifications are performed at each level of the classification process of the model.

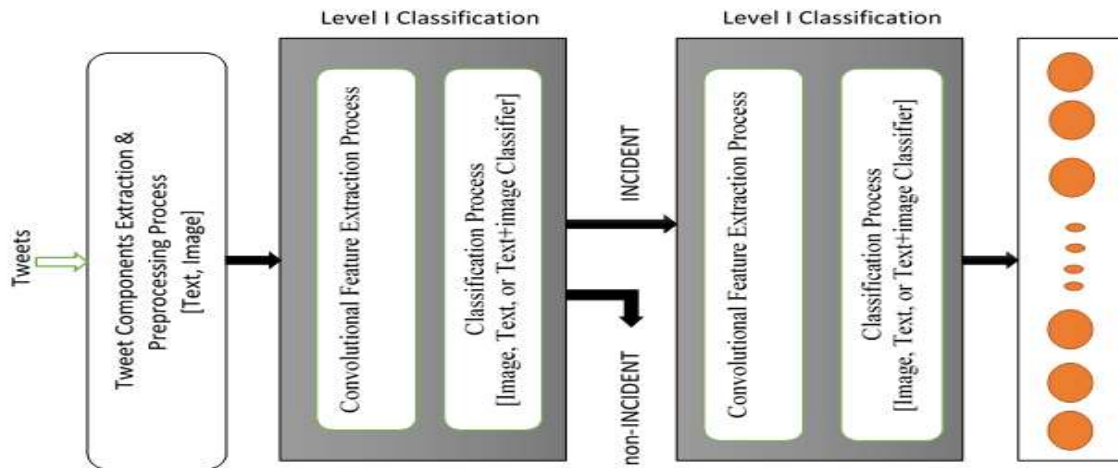


Figure 2. Overall structure of our system for classifying incident types of reported events.

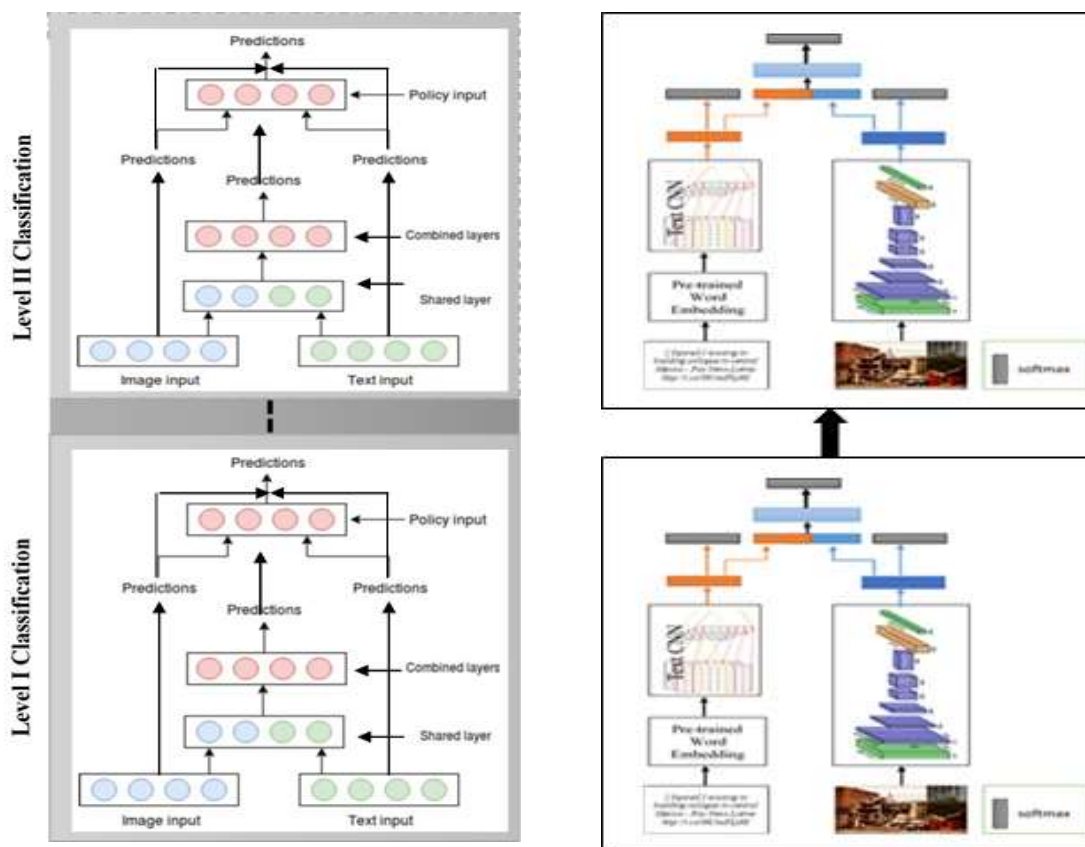


Figure 3. 2HDLnet Multi-modal fusion architectures.

Left: Feature-level and Decision-level fusion. **Right:** The proposed 2HDLTnet multi-modal architecture.

Text CNN. The text modality relies on GloVe (Pennington et al., 2014) word embeddings, which utilizes dense and real-valued numerical vectors to represent semantics of words, and

CNN nodes to model word vectors sequentially. Each word in the vocabulary V is represented by a D dimensional vector in a shared look-up table $L \in \mathbb{R}^{|V| \times D}$ where L is considered a model parameter to be learned. We initialize L randomly or using the GloVe pretrained word embedding vectors.

Given an input tweet text $s = (w_1, \dots, w_T)$, we first transform it into a feature sequence by mapping each word token $w_t \in s$ to an index in L . The look-up layer then creates an input vector $x_t \in \mathbb{R}^D$ for each token w_t which are passed through a sequence of convolution and pooling operations to learn high-level feature representations of the textual data and finally to a fully connected layer to perform prediction.

Image CNN: The image side of the network relies on convolutions and flattening layers to extract features and reshape image data. Typically, input image to a CNN is broken down into pixels for processing. Every pixel has a value between 0 and 255. The image is further transformed into a shape with values between 0 and 1.

Multimodal. The proposed hierarchical multi-modal system uses both tweet texts and images for the classification of incident tweets. For each level, the model combines the early fusion (feature level) with late fusion (decision level) strategies. The early fusion (feature level fusion) strategy in which the text and image features obtained from CNNs are directly combined in shared representation learning is implemented to better explore the correlation information between the image modality and tweet text modality. The decision-level strategy ensures a policy network is learned to decide the classifier to use for final prediction.

Hierarchical Deep Learning. The primary contribution of this study is the hierarchical classification of multimodal data. The structure of our hierarchical classification for the 2HDLnet architecture consists of combination of the classifiers at level I and level II of the classification tasks for text only, image only, multimodal (image + text) at feature level and multimodal (image + text) at decision level.

In this study, we experiment with a comprehensive set of models. In all cases, the models are trained with the standard backpropagation algorithm using rectified linear unit (ReLU) as activation function. ReLU, which sets negative value to zero, is defined in Eq. (1) and its gradient has the form of Eq. (2).

$$f(x) = \max(0, x) \quad (1)$$

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (2)$$

The output layer of the models uses softmax activation function with categorical cross-entropy as a loss function, which can be defined by Eq. (3).

$$\text{softmax function} = f(x_i) = \frac{e^{x_i}}{\sum_{k=1}^M e^{x_k}}, \text{ where } k = 1, 2, \dots, M \text{ and } x_i \in R \quad (3)$$

where x_i is the numerical value coming at the output layer from its previous layer which is the input value to the present layer and M represents the number of classes. The softmax function is calculating the probabilities of each target class over all possible target classes.

In all cases, given a set of N data samples, the objective is to minimize the negative log likelihood over the classes:

$$total\ loss = -\frac{1}{N} \sum_{i=1}^N \log(\text{softmax}(o(x_i), y_i)) \quad (4)$$

where o is the network's output, x_i is the unimodal or multi-modal input and y_i is the label. The softmax function represents the predicted class probability of the model for the i th training sample in a batch of N training samples.

Algorithm 1 Two-level Hierarchical Classification

Input: Dataset, DS , comprising of text and/or image

- 1: Split DS into majority DS_a and minority DS_b // DS_b consisting of 5 sub-categories
 - 2: **for** each modality in DS **do** // including bimodal [text+image] as a dataset group
 - 3: **for** $i = 1$ to DS_b **do**
 - 4: Relabel the class of DB_{bi} to C
 - 5: Copy DS_{bi} to DS_c
 - 6: **end for**
 - 7: **end for**
 - 8: Build level 1 classifiers [text, image & multimodal] using $DS_a \cup DS_c$
 - 9: Build level 2 classifiers [text, image & multimodal] using DS_b
-

As shown in Algorithm 1, given an imbalanced dataset, divided into majority class DS_a and minority classes DS_b where DS_b consists of five sub-categories, a new label C is given to the minority classes in DS_b . The newly labelled DS_c is used to replace the original minority class DS_b . DS_a and DS_c are combined to form a new dataset DS_d . DS_d is split into training, TR_d , validation, VA_d , and testing, TS_d , sets.

DS_b is also split into *training*, TR_b , *validation*, VA_b , and *testing*, TS_b , sets. In the first stage, classifiers for text, image and multimodal (image+text) are learned on the new dataset TR_d and validated on VA_d . TS_d is tested on the built classification models. In the second stage, classifiers for level 2 classification are built on TR_b , validated on VA_b and tested on TS_b .

4 Experiments

4.1 Dataset Description

Most of the publicly available real-world big data classification datasets, an essential building block of deep learning systems, typically contain only one data type (Zahavy et al 2017). We hardly find large-scale multimodal classification datasets available especially for personal and small scale incidents. Hence we tested our models using target incident-related real world home-grown dataset consisting of tweets and associated images collected using Twitter Streaming API over different time periods to insure a representative sample of crisis related images and description/captions using the appropriate keywords and hashtags in our target incident categories. Keywords and hashtags such as damaged building, heavy rainfall, flood, fire, shootings, terrorist attack, explosion, bomb, accident, etc., were queried to retrieve corresponding tweets (both text and images when available). We also collected some samples that suit our study from human and infrastructural dataset (Mouzannar, et al., 2018). The collected data consisting of text and/or images were filtered and manually annotated for

supervised learning into six categories for tasks (the one extra category is for non-incident) (i) Incident vs. non-Incident (ii) Incident categories (five classes).

Incident vs. non-Incident. The level I task is used to determine whether or not the tweet text and/or image is incident-related. The incident types (5 categories) were relabeled as a group with the general name ‘incident-related’ to address the problem of data imbalance in the data distribution using part of the procedure outlined in algorithm 1.

Incident Categories. The purpose of this level II classification task is to determine the category of incidents the tweet belongs if it is incident-related. Incident samples were annotated into one of the five different categories: accidents, crimes and civil disorder, damaged infrastructure, fire, and flood

The original dataset which contains a total of 408,670 samples consists of images, text and images with textual information. We divided the data set into two parts as shown in Table 1: Dataset 1 (DS-1) containing 39,279 samples is a subset of this full data set and Dataset 2 (DS-2), the full dataset with all the 408,670 data samples. Table 2 contains the list of categories and a breakdown of the distribution of samples in the datasets. Some images with associated text samples in the dataset are presented in Table 3.

Table 1 Details of datasets used

Dataset	Text	Images	Image with associated text	Total	Level I (Labels)	Level II (Labels)
DS-1	29684	6483	2625	39279	2	5
DS-2	383562	15483	9625	408670	2	5

Table 2 List of categories and distribution of samples in datasets

Label	Text		Images		Image with associated text		Total	
	DS-1	DS-2	DS-1	DS-2	DS-1	DS-2	DS-1	DS-2
Level I :Classification Task								
<i>Incident</i>	14843	190986	3141	7680	1789	4620	19773	203286
<i>Non-incident</i>	14841	190576	3342	7803	1323	5005	19506	205384
Total	29684	383562	6483	15,483	2625	9625	39,279	408670
Level II: Incident/Emergency Classification Task								
<i>Accident</i>	2830	38266	620	1506	345	922	3795	40694
<i>Crime and Civil Disorder</i>	3036	38804	621	1510	230	932	3887	41246
<i>Damaged Infrastructure</i>	2935	37815	634	1512	349	895	3918	40222
<i>Fire</i>	3004	37900	632	1550	388	921	4024	40371
<i>Flood</i>	3038	38201	634	1602	387	950	4059	40753
Total	14843	190986	3141	7680	1789	4620	19773	203286

4.2 Data Preprocessing







Twitter text were preprocessed for training, validation and testing. We remove all special characters, URLs, stop words and convert the text to lowercase characters. We preprocess the

images by reshaping them into the shape the network expects and scaling them so that all values are in the [0,1] interval.

4.3 Experimental Setup

We use 80% of the datasets as the training set and 20% as the test set. The validation set is 20% of the training set. To build the proposed hierarchical multimodal system some of the unimodal text and image models that were built in our previous works (Sako et al 2021a, 2021b) were reused in this study. We evaluate our unimodal and multimodal deep learning models and the hierarchical multimodal models/classifiers for the emergency situation identification tasks.

Table 3 Sample Dataset Images with associated text in tweets

Class	Image	Textual Description
accident		<i>I-77 Mile Marker 31 South Mooresville Iredell Vehicle Accident Ramp Closed at 8/6 1:18 PM</i>
damaged infrastructure		<i>2 Injured 1 missing in building collapse in central Mexico - Fox News Latino http://t.co/10UnaFLy0Y</i>
fire		<i>Raging #fire #apartment #apartmentfire #buildingfire #surreybc #surreyfire #nofilterReally hope everyone made it out</i>
flood		<i>#odorna #odornamarket #accraffloods #photography #lives #work</i>
crime and civil disorder		<i>Children are the primary target of Syrian regime bombardment#assadcrimes #syrie #isiscrimes #syria #syrians #syrianorphans #children #childrenofsyria#orphans #suffering #sufferingofsyrians</i>
non-incident		<i>We have a special guest tomorrow in our Library! Come checkout @littlesparklebaby jewelry collection! It's amazing! Show Your Love And Shop Small. #SmallBusinessSaturday</i>

Baseline Model. Support Vector Machine (SVM) was trained and tested to validate the accuracy advantage of our models.

CNN for Text Modality. For the text models, we experiment with two different CNN architectures consisting of 4 layers as follows:

- i) CNN_{t_2} : CNN Model with pre-trained 300-dimensional vectors from GloVe for word embedding. All word-including the unknown ones that are randomly initialized – are kept static and only the other parameters of the model are learned.
- ii) CNN_{t_3} : Same as in CNN_{t_2} but with parameters of the model being learned.

CNN for Image Modality. For the image model, we experiment with two different CNN models in increasing order of complexity and with varying configurations. The models are summarized as follows:

- (i) CNN_{v_2} : 5 layers CNN with image augmentation
- (ii) $VGG16_v$: Pre-trained VGG16 model (Simonyan and Zisserman, 2014) trained on ImageNet dataset (Deng et al., 2009), and the weights of the networks’ last layers were fine-tuned on our dataset for the classification tasks instead of the original 1000-way classification.

Multimodal. We perform experiments for both early fusion (feature-level fusion, FF) and late or decision-level fusion (DF). The models from each modality (text and images) were combined in the multimodal fusion. In the early fusion (FF) multimodal approach, once the text and image features have been extracted (i.e. from the last layer generated from each model), both feature vectors were concatenated and used to train the meta classifier. In the late or decision-level fusion (DF) multimodal approach, the decision scores from the unimodal text and image models and the decision scores of the early fusion models are fused.

Hierarchical Deep Learning. The primary contribution of this study hierarchical classification of multimodal data. The structure of our Hierarchical Deep Learning for the 2HDLnet architecture consists of combination of the models at level I and level II of the classification tasks for text only, image only, multimodal (image + text) at feature level and multimodal (image + text) at decision level.

For all the models, the training process is done using an early stopping where the model will stop training before it overfits the training data. The maximum number of epochs is set to 100. We experiment with {0.25, 0.5} dropout rates (Srivastava et al. 2014) to avoid overfitting and {16, 32} mini-batch sizes. We use the Adam optimizer (Kingma and Ba 2014) to optimize the cross entropy). The hyper-parameters used in this study are listed in Table 4. In all experiments, the models (and hyper parameters) are tuned on the validation (development) set and the final performance evaluated against the test set.

Table 4 Hyper-parameter settings for the proposed model

	Text Model	Image Model	Multi-modal Model (Text + Image)
Optimizer	Adam	Adam	Adam
Loss function	Categorical cross entropy	Categorical cross entropy	Categorical cross entropy
Learning rate	0.0001	0.0001	0.0001
Batch size	32	16	16
Activation	ReLu, Softmax	ReLu, Softmax	ReLu, Softmax
Epochs	100	100	100

5 Results and Discussion

In this section, we present the experimental results and discussions of our proposed 2HDLnet model for image and/or text classification on the DS-1 and DS-2 datasets. We consider different possibilities for the bi-modal input data consisting of text and image, i.e., missing

text (i.e. image only), missing image (i.e. text only), and bi-modal input (i.e. text and image). Hence, all the models; namely Tweet text models, Image models and multi-modal models (Tweet text + Image), are trained separately at each level of the classification task with the datasets. For our proposed 2HDLnet, we experiment with combinations of the different models in the Level I and Level II of the classification tasks respectively. We calculate the accuracy to evaluate the performance of the models.

Accuracy is the fraction of predictions our model got right. It is calculated by dividing the correctly classified tweets by the total number of tweets as in Eq. (5). The higher the accuracy, the better the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

where True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are four indicators that were measured.

Tables 5 and 6 show the results from our experiments with the overall accuracy for DS-1 and DS-2 datasets respectively. We report the results in four settings: text only, image only, text+image (early fusion), and text+image (late fusion).

Table 5 2HDLnet Accuracy of DS-1 dataset

	Text only		Image only		Text+Image (FF)		Text+Image (DF)					
	Methods	Accuracy	Methods	Accuracy	Methods	Accuracy	Methods	Accuracy				
Baseline (SVM)	83.25	82.02	82.64	81.02	80.88	80.95	79.76	81.90	80.83	82.34	82.11	82.23
2HDLnet	$CNN_{t2}^{(1)}$	$CNN_{t2}^{(2)}$	86.26	$CNN_{v2}^{(1)}$	$CNN_{v2}^{(2)}$	81.41	$CNN_{t2}^{(1)}$	$CNN_{t2}^{(2)}$	82.12	$CNN_{t2}^{(1)}$	$CNN_{t2}^{(2)}$	81.68
	86.65	85.87		81.37	81.45		$CNN_{v2}^{(1)}$	$CNN_{v2}^{(2)}$		$CNN_{t2}^{(1)}$	$CNN_{t2}^{(2)}$	
	$CNN_{t2}^{(1)}$	$CNN_{t3}^{(2)}$	85.54	$CNN_{v2}^{(1)}$	$VGG16_v^{(2)}$	83.36	$CNN_{t2}^{(1)}$	$CNN_{t3}^{(2)}$	85.28	$CNN_{t2}^{(1)}$	$CNN_{t3}^{(2)}$	84.52
	86.65	84.43		81.37	85.34		$VGG16_v^{(1)}$	$VGG16_v^{(2)}$		$CNN_{t2}^{(1)}$	$CNN_{t3}^{(2)}$	
	$CNN_{t3}^{(1)}$	$CNN_{t2}^{(2)}$	86.88	$VGG16_v^{(1)}$	$CNN_{v2}^{(2)}$	82.74	$CNN_{t3}^{(1)}$	$CNN_{t2}^{(2)}$	86.05	$CNN_{t3}^{(1)}$	$CNN_{t2}^{(2)}$	83.68
	87.89	85.87		84.03	81.45		$CNN_{v3}^{(1)}$	$CNN_{v3}^{(2)}$		$CNN_{t3}^{(1)}$	$CNN_{t2}^{(2)}$	
	$CNN_{t3}^{(1)}$	$CNN_{t3}^{(2)}$	86.16	$VGG16_v^{(1)}$	$VGG16_v^{(2)}$	84.69	$CNN_{t3}^{(1)}$	$CNN_{t3}^{(2)}$	86.60	$CNN_{t3}^{(1)}$	$CNN_{t3}^{(2)}$	88.95
	87.89	84.43		84.03	85.34		$VGG16_v^{(1)}$	$VGG16_v^{(2)}$		$CNN_{t3}^{(1)}$	$CNN_{t3}^{(2)}$	

Table 6 2HDLnet Accuracy of DS-2 dataset

	Text only			Image only			Image + Text (FF)			Image+Text (DF)		
	Methods		Accu- racy	Methods		Accu- racy	Methods		Accu- racy	Methods		Accu- racy
Baseline (SVM)	82.19	82.33	82.26	80.07	79.90	79.99	80.12	81.26	80.69	81.23	80.78	81.01
2HDLnet	$CNN_{t2}^{(1)}$	$CNN_{t2}^{(2)}$	88.30	$CNN_{v2}^{(1)}$	$CNN_{v2}^{(2)}$	84.22	$CNN_{t2}^{(1)}$ +	$CNN_{t2}^{(2)}$ +	82.83	$CNN_{t2}^{(1)}$ +	$CNN_{t2}^{(2)}$ +	82.57
	88.98	89.42		85.22	83.21		$CNN_{v2}^{(1)}$ +	$CNN_{v2}^{(2)}$		$CNN_{v2}^{(1)}$ +	$CNN_{v2}^{(2)}$	
	$CNN_{t2}^{(1)}$	$CNN_{t3}^{(2)}$	89.20	$CNN_{v2}^{(1)}$	$VGG16_v^{(2)}$	85.34	$CNN_{t2}^{(1)}$ +	$CNN_{t3}^{(2)}$ +	86.34	$CNN_{t2}^{(1)}$ +	$CNN_{t3}^{(2)}$ +	87.26
	88.98	89.65		85.22	85.45		$VGG16_v^{(1)}$ +	$VGG16_v^{(2)}$		$VGG16_v^{(1)}$ +	$VGG16_v^{(2)}$	
	$CNN_{t3}^{(1)}$	$CNN_{t2}^{(2)}$	89.69	$VGG16_v^{(1)}$	$CNN_{v2}^{(2)}$	84.58	$CNN_{t3}^{(1)}$ +	$CNN_{t2}^{(2)}$ +	85.62	$CNN_{t3}^{(1)}$ +	$CNN_{t2}^{(2)}$ +	85.23
	89.95	89.42		85.95	83.21		CNN_{v3} +	CNN_{v3}		CNN_{v3} +	CNN_{v3}	
$CNN_{t3}^{(1)}$	$CNN_{t3}^{(2)}$	89.80	$VGG16_v^{(1)}$	$VGG16_v^{(2)}$	85.70	$CNN_{t3}^{(1)}$ +	$CNN_{t3}^{(2)}$ +	88.12	$CNN_{t3}^{(1)}$ +	$CNN_{t3}^{(2)}$ +	89.63	
89.95	89.65		85.95	85.45		$VGG16_v^{(1)}$ +	$VGG16_v^{(2)}$		$VGG16_v^{(1)}$ +	$VGG16_v^{(2)}$		

These results show that overall performance improvement for general data classification is obtainable with the proposed deep learning approaches compared to traditional methods. The 2HDLnet approaches with stacked, deep learning architectures clearly provide superior performance in all the different modalities across the two datasets. For text unimodal hierarchical classifiers, the best accuracy is obtained by the combination CNN_{t3} for the level I and CNN_{t3} for level II of the classification. This gives accuracies of 87.89% for level I, 86.56% for the level II and 87.23% overall on DS-1 and 88.90% for the level I, 89.65% for level II and 89.28% overall on DS-2 datasets respectively. This is slightly better than all of the other combination of CNN models. For the image models, the best scores are again achieved by VGG16 for level I and VGG16 for level II in both datasets. The closest scores to these are obtained by $VGG16_{v1}$ and CNN_{v2} in levels I and II respectively. For the multimodal hierarchical models (text and image) using shared representative learning at feature level (FF), the accuracy is obtained by the combination of CNN_{t3} and $VGG16_v$ for the level I and $CNN_{t3}+VGG16_v$ for the level II of classification. This gives accuracies of 87.01% for the first level, 86.19% for the second level and 86.60% overall. The Decision-level Fusion (DF) multimodal model has a winner in CNN_{t3} and $VGG16_v$ at level I and CNN_{t3} and $VGG16_v$ at level II. Across the models and modalities and datasets, the best performance is seen in the DF hierarchical models.

The pretrained input word embeddings enjoy great success in the experiments with text modality classification, even without fine-tuning. This proves that the pre-trained word vectors have good generalization and adaptability in different classification tasks. Above all, the proposed hierarchical unimodal and multimodal fusion strategies that selectively combine early and late fusions further improve these results. Testing on datasets of data samples obtained from the social media shows that combinations of CNN_{t3} and $VGG16_v$ at the higher level and CNN_{t3} and $VGG16_v$ at the lower level produced accuracies consistently higher than

those obtainable by conventional approaches using SVM. These results showed that deep learning methods can provide improvements for data classification and that they provide flexibility to classify data within a hierarchy. Figure 4 shows the real-time classification output of our proposed system.

```
Performing Level I classification...
-----
[[0.85820894 0.14179106]]
incident: 0.86
non_incident: 0.14
-----
Prediction: incident 85.82089403053365

Homes affected by the blast. #accrafloods photo by: @teresameka we'll be distributing care packages soon,
if you are interested in assembling them with us email ghanafloodrelief@gmail.com. We will also let you
know how to donate soon #ghana #life => incident
Tweet Prediction [level I] => incident: Confidence(%): 85.82

Performing Level II Classification
-----
[[0.00529322 0.00132822 0.95572176 0.00452288 0.03313391]]
accident: 0.0053
crime_n_civil_disorder: 0.0013
damaged_infrastructure: 0.96
fire: 0.0045
flood: 0.033
-----
Prediction: damaged_infrastructure 95.57217636326716

Homes affected by the blast. #accrafloods photo by: @teresameka we'll be distributing care packages soon,
if you are interested in assembling them with us email ghanafloodrelief@gmail.com. We will also let you
know how to donate soon #ghana #life => damaged_infrastructure
Tweet Prediction [level II] => damaged_infrastructure: Confidence(%): 95.57

Performing Level I classification...
-----
[[0.02262969 0.97737031]]
incident: 0.023
non_incident: 0.98
-----

Tweet=> What a wonderful day! =>non_incident
-----
Tweet Prediction [level I] => non_incident: Confidence(%): 97.74
```

Figure 4. Real-time classification output of our system

6. Conclusion

In this study, we proposed a two-stage Hierarchical Multimodal Deep Learning Neural Network Model (2HDLnet) using Convolutional Neural Networks (CNNs) for the classification of text and/or image contents of Twitter posts. This classification approach entails the temporary regrouping and re-labelling of classes of the training data toward a more balanced distribution and performing hierarchical data classification using stacks of deep learning architectures to allow both overall and specialized learning by level of the data hierarchy and modality for modeling incidents that contribute to emergency situation awareness. The results of the experimentation – training, validation and testing - of the proposed model on the home-grown datasets produced accuracies consistently higher than those obtainable by conventional approaches like the use of Support Vector Machines. The experimental analysis demonstrates that the proposed 2HDLnet approach can provide improvements for image and/or text data classification at each logical layer. The results showed best performance of the 2HDLnet model at decision level of the hierarchical model. The main contribution of this study is hierarchical classification of multimodal documents.

References

- Abel, F., Hauff, C., Houben, G.-J., Tao, K. and Stronkman, R. (2012) Twitcident: Fighting fire with information from social web streams. *Proceedings of the 21st International Conference on World Wide Web*. ACM, Lyon, France, 305–308.
- Alam, F., Imran, M. and Ofli, F. (2017). Image4Act: Online Social Media Image Processing for Disaster Response. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 601-604. <http://dx.doi.org/10.1145/3110025.3110164>
- Bader-El-Den, M., Teitei, E. and Adda, M. (2016). Hierarchical classification for dealing with the Class imbalance problem, *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 3584-3591, doi: 10.1109/IJCNN.2016.7727660.
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828
- Cameron, M.A., Power, R., Robinson, B. and Yin, J. (2012). Emergency situation awareness from twitter for crisis management. *Proceedings of the 21st International Conference on World Wide Web*. 695-698.
- Choi, E. (2018). Doctor AI: Interpretable deep learning for modeling electronic health records. A PhD dissertation. Georgia Institute of Technology.
- Chowdhury, S., Imran, M., Asghar, M.R., Amer-Yahia, S., and Castillo, C. (2013). Tweet4act: using incident-specific profiles for classifying crisis-related messages. *Proceedings of 10th International ISCRAM Conference*, Baden, Germany.
- Daly, S. and Thom, J.A. (2016). Mining and Classifying Image Posts on Social Media to Analyse Fires. *Long Paper – Social Media Studies Proceedings of the ISCRAM 2016 Conference – Rio de Janeiro, Brazil*.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 248-255.
- Eskesen, S. (2017). Improving product categorization by combining image and title. Master Thesis.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning. MIT Press online. <http://www.deeplearningbook.org>.
- Haj-Yahia, Z. (2018). Introduction to Deep Learning. <https://www.kdnuggets.com/2018/09/introduction-deep-learning.html>
- Huang, G. B. Lee, H. and Learned-Miller, E. (2012). Learning hierarchical representations for face verification with convolutional deep belief networks, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2518–2525.
- Kiela, D., Grave, E., Joulin, A. and Mikolov, T (2018). Efficient large-scale multi-modal classification, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 5198–5204.
- Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 26th Annual Conference on Neural Information Processing Systems.*, Lake Tahoe, Nev., 1097–1105
- Kumar A, Singh JP, Dwivedi YK and Rana NP (2020) A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-020-03514-x>

- Lagerstrom, R., Arzhaeva, Y., Szul, P., Obst, O., Power, R., Robinson, B., and Bednarz, T. (2016). Image classification to Support Emergency Situation Awareness. *Frontiers in Robotics and AI*. <https://doi.org/10.3389/frobt.2016.00054>
- LeCun, Y, Bengio, Y and Hinton, G. (2015). Deep learning, *Nature*, 521(7553), 436–444
- LeCun, Y., Bottou, L., Bengio, Y, and Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lee, J. Y. and Deroncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks, arXiv preprint arXiv:1603.03827
- Leung, M. K. K., Xiong, H. Y., Lee, L. J. and Frey, B. J.(2014). Deep learning of the tissue-regulated splicing code, *Bioinformatics*, 30(12), 1121–1129.
- Makino, K., Takei, Y., Miyazaki, T., & Goto, J. (2018). Classification of Tweets about Reported Events using Neural Networks. *NUT@EMNLP*.
- Mayo, M. (2016). 7 Steps to Understanding Deep Learning. <https://www.kdnuggets.com/2016/01/seven-steps-deep-learning.html>
- Mouzannar, H., Rizk, Y. and Awad, M. (2018). Damage Identification in Social Media Posts using Multimodal Deep Learning. *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management*, Rochester, 529-543.
- Nooka, S.P. (2016). Fusion of Mini-Deep Nets. Thesis. Rochester Institute of Technology. <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=10351&context=theses>
- Ofi, F., Alam, F., and Imran, M. (2020). Analysis of social media data using multimodal Deep Learning for Disaster Response. arXiv:2004.11838v1 [cs.CV]
- Oquab, M., Bottou, L. Laptev, I. and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.
- Peters, R., and Albuquerque, J. P. D. (2015). Investigating images as indicators for relevant social media messages in disaster management, in *The 12th International Conference on Information Systems for Crisis Response and Management*, Kristiansand, Norway.
- Rizk, Y., Jomaa, HS., Awad, M. and Castillo. C. (2019). A computationally efficient multi-modal classification approach of disaster-related Twitter images. In *the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19)*, Limassol, Cyprus. <https://doi.org/10.1145/3297280.3297481>
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.* 25, 919–931. doi:10.1109/TKDE.2012.29
- Sako, DJS., Onuodu, FE. and Eke, BO. (2021a). A Method for Classifying Tweets about Emergency Events using Deep Neural Networks. *Journal of Software Engineering and Simulation*, 7(9), 38-45.
- Sako, DJS., Onuodu, FE. and Eke, BO. (2021b). A Model for Improving Image Classification Using Convolutional Neural Network for Emergency Situation Reporting. *International Journal of Advanced Research in Computer and Communication Engineering*. 10(9), 53-60.
- Salakhutdinov, R., Tenenbaum, J. B. and Torralba, A. (2013). Learning with hierarchical deep models, *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1958–1971.
- Schulz, A. and Ristoski, P. (2013). The car that hit the burning house: Understanding small scale incident related information in microblogs. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*
- Shridhar, K. (2017). A Beginners Guide to Deep Learning. <https://medium.com/@shridhar743/a-beginners-guide-to-deep-learning-5ee814cf7706>

- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of MLR*, 15(1),1929–1958.
- Stefanowski, J. and Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *Data Warehousing and Knowledge Discovery*, 283–292. Springer
- Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A. and Anderson, K.M. (2011). Natural Language Processing to the Rescue? Extracting “Situational Awareness” Tweets During Mass Emergency. *Proceedings of the Fifth International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media*. 385-392.
- Vieweg, S., Hughes, A. L, Starbird, K. and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *Proceedings of the SIGCHI conference on human factors in computing systems*. 1079-1088. <https://doi.org/10.1145/1753326.1753486>
- Wang, D., Mao, K., and Ng, G-W (2017) Convolutional Neural Networks and Multimodal Fusion for Text Aided Image Classification. *Proceedings of the 20th International Conference on Information Fusion*, Xi'an, China.
- Yim, J, Ju, J., Jung, H and Kim, J. (2015). Image classification using convolutional neural networks with multi-stage feature. *Advances in Intelligent Systems and Computing*, 345, 587–594
- Zahavy, T., Mannor, S., Magnani, A. and Krishnan, A. (2016). Is a Picture worth a Thousand Words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-Commerce. arXiv:1611.09534v1 [cs.CV]
- Zhou, L., Li, Q., Huo, G., and Zhou, Y. (2017). Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features. *J. of Computational Intelligence and Neuroscience*. <https://doi.org/10.1155/2017/3792805>