Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Disaster Related Tweets Categorization using Multimodal Approach

Sumit Bidari [a], Ram Krishna Maharjan [b], Sanjeeb Prasad Panday [c], Aman Shakya [d]

[a, b, c, d] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] 075MSICE020.sumit@pcampus.edu.np, [b] rkmahajn@ioe.edu.np,
[c] sanjeeb@ioe.edu.np, [d] aman.shakya@edu.np

Abstract
Contents shared in form of text and images in multimedia during and after disasters can be used to analyze the information about the event. The analysis can be done to know the report of affected people as missing or injured, infrastructure and utility damages, rescue needed for victims and many others. Many researches in the past have been done on focusing on either text modality or on image modality for disaster response. Only few work has been done till now for the use of both text and image modality for disaster response and they only focus on same category of text and images and practically it is found less reliable. In this paper, we propose to use both text and image of different category and fuse them using score fusion for joint representation of text and images. For text modality, we have used BERT model and for image modality we have used VGG16 modality and fused them using late fusion for multimodal analysis of disaster related tweet categorization.

Keywords
Multimodal content, Multimodal fusion, Disasters and analysis

1. Introduction

Social media is used as a handy platform for sharing people's emotions and messages. Messages in social media can be shared in terms of different multimedia content such as text, images, audio, video etc. From all over the world, at every second, billions of information is shared in social media in the form of images and texts. Information shared as text and images can be used to identify a concept, an event and many more. By not relying only on text data but by also analyzing information coming from different modes, the concept of multimodal learning came in. The concept of multimodal learning was applied to social media data analysis of the multimedia content [1] posted during disaster which help humanitarian organizations in preparedness, mitigation, response and recovery efforts.

This concept of multimodal learning has been also applied to different fields such as audio-visual analysis [2] cross-modal study [3] and speech processing (e.g. audio and transcriptions). This model aims to classify disaster related tweets by using information of both real text and image modalities extracted from twitter data. This system aim to categorize the tweet as whether it contains useful humanitarian information as infrastructure damages, vehicle damages, rescue, volunteering or donation efforts and affected individuals (injury, dead, missing, found etc.).Information related to ongoing disasters are often shared in image-text pairs by victim or by general public. By categorizing only text or only images can be sometimes less informative for first responder, rescue and awareness members. So, multimodal learning approach can also be used to categorize disasters related tweets extracted in real time using image-text pairs.

The major objectives of thesis are to implement BERT model with VGG16 for disaster tweets classifications using text and images where texts are analyzed using BERT and images are analyzed using VGG16 and text-images are analyzed using post fusion technique and use disaster related real tweet extracted from tweeter in image-text pairs to categorize for humanitarian operations which includes Infrastructure and utility damages, Vehicle damages, recue, volunteering or donation efforts and affected individuals using multimodal approach and comparing our results with other multimodal approaches for disaster realated tweet categorization.

2. Methodology

The proposed system is as shown in figure1 is for the classification task which takes image- text pairs as an input posted during disaster or after disaster in social media. Real tweet can be extracted from tweeter in the form of images and texts using tweeter API. The system contains vgg16 network to extract feature maps from images and VGG16 has been implemented to test its performance in this dataset CRISISMMD [4] and got good performance [5] and BERT model to extract text features from texts which has also been implemented in the same dataset and got good accuracy. The hidden layer of equal size is kept at both models to get equal number of output. The feature extracted by each model is passed to shared representation layer where the post fusion of the both text and image features is done and the result is given to softmax layer which classify the tweets into five classes according to the probability calculated by the softmax function.

2.1 Dataset

Multimodal dataset is used for our model, obtained from [4]which contains tweets and associated images which was collected from seven different natural disasters [4]. The collected data set contains tweet text and tweet images about humanitarian categories and not humanitarian categories .The humanitarian categories contained 8 classes. The category 'injured or dead people' and 'missing or found people' and 'vehicle damage' contains less number of tweets. So, 'injured or dead people' and missing or found people category are merged into 'affected individual' category. Similarly, 'vehicle damage' category is merged into 'infrastructure and utility damages' category. After merging these categories only 5 classes are obtained. The data set contains retweets and redundant tweets and those tweets were removed. Similarly tweet containing 70 percent similarity was also removed from dataset. As a result we got following number of dataset as shown in table 1.

Table 1: Multimodal dataset used for the model

Category	Texts	Images
not humanitarian	4,394	17,57
Other relevant information	5,686	929
affected individuals	947	17.83
Infrastructure and utility damages	1,213	2,387
rescue volunteering or donation effort	3,197	1,235

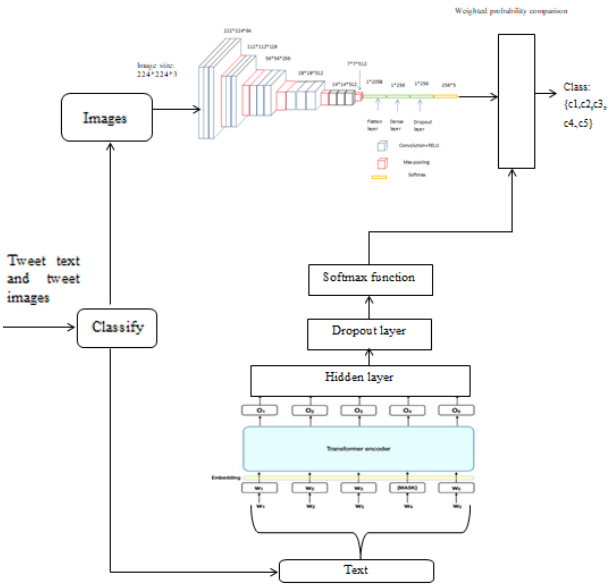

Figure 1: Multimodal achitecture for classification task using both text and images as input to the system.

2.2 Data Processing

For image, unbalanced image dataset is observed for different classes. So, for balancing unbalanced dataset image augmentation techniques is used before passing images to model. For image augmentation zoom, shift, rescaling, rotation, increase in width, height, and horizontal flip has been done. The image dimension is set to 224*224*3 and passed to model.

For text dataset, the dataset is noisy and needs to be preprocessed before sending it to model for training. So, to remove noise from dataset texts are converted to lowercase text, removed fully strip line breaks, replaced all URLs , removed all email addresses, removed all phone numbers, removed all numbers, removed all digits, removed all currency symbols , fully removed punctuation, removed all retweets.

2.3 VGG16: Image Modality

For images, VGG16 model is used to train the model for available dataset. The image size given to the system is 224*224*3 and idea of transfer learning approach is used for using existing weights of a pre-trained model on ImageNet [6]. Weights of a VGG16 model is used which is pre-trained on ImageNet to initialize the model. The last three layers are modified of VGG16 where last two layers are using Relu as activation function and units given are 224 for both layers. The last layer is taken as softmax layer where classification is done into 5 different classes. The model is trained with 12,352 numbers of images and

validated using 3,085 numbers of images belonging to five different classes. The model is trained using Adam Optimizer. The model got 73 percent accuracy on validation dataset.

2.4 BERT: Text Modality

BERT model has achieved better performance on categorizing disaster related tweets compared to CNN [7]. Bert model has achieved state of art result on Natural Language Processing tasks predicting next sentence and understanding context of words from sentences. For our system, the preprocessed dataset is passed into BERT model for training the model. The model is trained using hold out method with batch size of 16 and initial learning rate is taken as 2e-5. The dataset is divided into train and validation set, where train data set containing 12349 numbers of tweets is used to train the model and validation set containing 3088 number of tweets is used to validate the model. The model is evaluated using confusion matrix where we got FP, TP, FN, TN. From confusion matrix, the model got classification accuracy of 81 percent.

2.5 Multimodal: text and images

Multimodal deep neural network used is as shown in figure for our experiment. As mentioned earlier, for the image modality VGG16 model is used for classifications and for text modality BERT model is used for classifications. Late fusion, as comparing the confidence probabilities between two modes is used and the model getting better confidence probabilities will contribute more in final classifications. The last layer is the softmax layer for categorizing the class of given inputs.

3. Results and Discussion

In table 2, we present the performance of unimodal and multimodal for similar category text-image pairs, where text modality got 82 percent accuracy and for image modality we got 73 percent and our proposed multimodal system 87 percent accuracy on similar category text-image pairs.

Similarly, table 3 shows the accuracy of different category text-image pairs where the text when analyzed got 86 percent accuracy and image when analyzed got 73 percent accuracy and our proposed model got accuracy of 81 percent. Though the model got less accuracy compared to text and image modality but our system takes information of both text and images but is more reliable and applicable as it can fuse both similar category and different category text-image pairs.

Table 2: Results for the humanitarian classification task for similar category image-text pairs

Modality	Accuracy	Precision	Recall	F1 score
Image	73	81	73	76.79
Text	82	86	82	83.9
Multimodal	87	89	87	87.9

Table 3: Results for the humanitarian classification task for different category image-text pairs

Modality	Accuracy	Precision	Recall	F1 score
Image	74	77	74	75.4
Text	86	86	86	86
Multimodal	82	84	82	82.98

4. Conclusion

During and after disasters many disasters related tweets are tweeted by victims and their well-wishers in the image-text pairs. When images and texts are analyzed separately then much information can be missed regarding victims and affected areas. So our proposed system takes both images and text information using late fusion technique and analyzes both text- image pairs in different humanitarian category. Previous model only categorizes tweets containing same label image and same label texts but our model has no such limitations. In summary, our study has two main contributions which are that It has got better accuracy compared to previous baseline model and has no limitation that texts and images pairs must be of same category.

Acknowledgments

This work was supported by Tribhuvan University, Institute of Engineering in 2020.

References

[1] L. Misra, A. Kumar, K. Misra, S. Aggarwal, R. R. Shah, and A. K. Gautam. Multimodal analysis of disaster tweets. *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 1(3):10, 2019.

[2] E. Cambria, N. Howard, G. B. Huang, and Hussain Poria S. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 25:50—-59, 2016.

[3] S. Albanie, A. Zisserman, and Nagrani A. Seeing voices and hearing faces: Cross-modal biometric matching. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2321–9653, 2018.

[4] F. Ofli, M. Imran, and F. Alam. Crisismmd: Multimodal twitter datasets from natural disasters. In *in AAAI press, pp. 465–473, In:Proc. of the 12th ICWSM*, pages 41–48, 2018.

[5] Ferda Ofli, Muhammad Imran, Tanvirul Alam, Umair Qazi, and Firoj Alam. Deep learning benchmarks and datasets for social media image classification for disaster response. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Dhaka, Bangladesh*, 52(1):1–4, 2020.

[6] J. Clune, J. Bengio, Y. Lipson, and H. Yosinski. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320—3328, 2014.

[7] Hassan Sajjad, Muhammad Imran, Ferda Ofli, and Firoj Alam. Crisisbench: Benchmarking crisis-related social media datasets. (1):22–36, 2021.