TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Pulchowk Campus

THESIS NUMBER: 075MSICE020

Categorization of Disaster Related Tweets using Multimodal Approach

by

Sumit Bidari

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

LALITPUR, NEPAL

August, 2021

Disaster Related Tweets Categorization using Multimodal Approach

By

Sumit Bidari

075MSICE020

Thesis Supervisor
Prof. Dr. Ram Krishna Maharjan

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information and Communication Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

Lalitpur, Nepal

August, 2021

COPYRIGHT©

DECLARATION

I declare that the work hereby submitted for Master of Science in Information and Communication Engineering (MSICE) at IOE, Pulchowk Campus entitled "**Disaster Related Tweets Categorization using Multimodal Approach**" is my own work and has not been previously submitted by me at any university for any academic award.

I authorize IOE, Pulchowk Campus to lend this thesis to other institution or individuals for the purpose of scholarly research.

Sumit Bidari

075MSICE020

August, 2021

RECOMMENDATION

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled "**Disaster Related Tweets Categorization using Multimodal Approach**" submitted by Sumit Bidari in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Information and Communication Engineering**".

……

Supervisor: Ram Krishna Maharjan, PhD

Professor,

Department of Electronics and Computer Engineering,

Institute of Engineering, Pulchowk Campus

…….

External Examiner: Kamal Chapagain, PhD

Assistant Professor,

Department of Electrical and Electronics Engineering,

School of Engineering, Kathmandu University

……..

Committee Chairperson: Basanta Joshi, PhD

Program Coordinator,

Information and Communication Engineering,

Department of Electronics and Computer Engineering,

Institute of Engineering, Pulchowk Campus

Date: August, 2021

DEPARTMENTAL ACCEPTANCE

The thesis entitled "**Disaster Related Tweets using Multimodal Approach**" submitted by **Sumit Bidari** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Information and Communication Engineering**" has been accepted as a bonafide record of work independently carried out by him in the department.

……
Prof. Dr. Ram Krishna Maharjan

Head of Department,

Department of Electronics and Computer Engineering,

Pulchowk Campus, Institute of Engineering,

Tribhuvan University,

Lalitpur, Nepal**.**

Acknowledgement

I would like to express my deep gratitude to **Prof Dr. Ram Krishna Maharjan** for valuable suggestions and feedback during the thesis progress. I would also like to pay my sincere thanks to **Dr. Sanjeeb Prasad Panday** and **Dr. Aman Shakya** for guiding me throughout the thesis progress. This report has been made in relation of UGC collaborative grand on ongoing disaster communication and management with **Dr. Sanjeeb Prasad Panday** and **Dr. Aman Shakya**.
I am thankful to **Dr. Basanta Joshi** and Department of Electronics and Computer Engineeging for guiding us in the process of thesis completion.

Sincerely,
Sumit Bidari
075MSICE020

Contents

COPYRIGHT© iii
Acknowledgement vii
List of Table v
Abstract 2
List of Abbreviations 3
CHAPTER 1 4
1. INTRODUCTION 4
1.1. Background 4
1.2. BERT model: 4
1.3. VGG16 (Visual Geometry Group): 7
1.4. Problem statement 8
1.5. Objectives 9
1.6. Scope of work 9
Chapter 2 10
2. LITERATURE REVIEW 10
2.1. Classification of disasters based on textual data 10
2.2. Classification of disaster based on image data 10
2.3. Classification of disaster based on both text data and image data 11
Chapter 3 12
3. METHODOLOGY 12
3.1. System 12
3.2. Softmax function: 12
3.3. Tools and Resources 12
3.4. Data set and Preprocessing 15
3.5. VGG16: Image Modality 17
3.6. BERT model: Text modality 18
3.7. Multimodal: text and images 18
Chapter 4 19
4. Evaluation Metrics 19
4.1. Precision: 19
4.2. Recall 19
4.3. F1 score 19
4.4. Accuracy 19

Chapter 5 20
5. IMPLEMENTATION AND ANALYSIS 20
5.1. BERT model 20
5.2. VGG16 model 22
5.3. Comparison table: 24
5.4. Multimodal Fusion: 24
Chapter 6 36
6. CONCLUSION 36
6.1. Conclusion 36
6.2. Future work 36
Chapter 7 37
7. Time schedule 37
References 38

List of Figures

Figure 1: Pretraining of BERT 5
Figure 2: BERT input representation [8] 6
Figure 3: VGG16 Architecture [13] 7
Figure 4: VGG16 [13] 8
Figure 5: Multimodal architecture for the classification task using both text and image as input to the system [1] 11
Figure 6: Methodology 14
Figure 7: Infrastructure and utility damages multimodal data 15
Figure 8: Not_humantarian multimodal data 16
Figure 9: Rescue and volunteering multimodal dataset 16
Figure 10: Affected individuals multimodal data 16
Figure 11: Other relevant multimodal data 16
Figure 12: BERT model 20
Figure 13: Confusion matrix of Bert model 21
Figure 14: VGG16 22
Figure 15: Accuracy vs Epochs of Vgg16 model using Adam optimizer 23
Figure 16: Loss vs Epochs of VGG16 using Adam optimizer 23
Figure 17: Confusion matrix of VGG16 model 24
Figure 18: Evaluation metrics for weights between 0.01 to 0.1 and 0.99 to 0.9 26
Figure 19: Evaluation metrics for weights between 0.11 to 0.2 and 0.89 to 0.8 27
Figure 20: Evaluation metrics for weights between 0.21 to 0.3 and 0.79 to 0.7 27
Figure 21: Evaluation metrics for weights between 0.41 to 0.5 and 0.59 to 0.5 28
Figure 22: Confusion Matrix of Multimodal System 29
Figure 23: Evaluation metrics for weights 0.11 to 0.2 and 0.89 to 0.8 31
Figure 24: Evaluation metrics for weights 0.21 to 0.2 and 0.79 to 0.7 32
Figure 25: Evaluation metrics for weights 0.31 to 0.4 and 0.69 to 0.6 33
Figure 26: Evaluation metrics for weights 0.31 to 0.4 and 0.69 to 0.6 34
Figure 27: Confusion Matrix of Multimodal System 35

List of Table

Table 1: Multimodal dataset used for the model 15
Table 2: Table depicting model performance of BERT..... 21
Table 3: Table depicting performance in individual class of VGG16 23
Table 4: Table depicting overall performance of theVGG16 model 24
Table 5: Comparison table of BERT and VGG16 24
Table 6: Evaluation metrics showing fusion model performance 25
Table 7: Evaluation metric showing individual class performance 25
Table 8: Results for the classification task 26
Table 9:: Evaluation metric showing individual class performance 28
Table 10: Evaluation metric showing individual class performance 29
Table 11: Results for the classification task 29
Table 12: Results for the classification task 35

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. Report of affected people as missing or injured, infrastructure and utility damages, rescue and volunteering needed, not humanitarian or other relevant information can also be found with this analysis. It has been found that only few researches focuses on text as well as image modality for such analysis. Also no works has been done for mixture of dissimilar and similar category text-image pairs. In this paper, we aim to use both text as well as 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

List of Abbreviations

FP: False Positive
TP: True Positive
TN: True Negative
FN: False Negative
FPR: False Positive Rate
TPR: True Positive Rate
VGG: Visual Geometry Group
CNN: Convolutional Neural Network
SGD: Stochastic Gradient Decent
BERT: Bidirectional Encoder Representation from Transformer

CHAPTER 1

1. INTRODUCTION

1.1.Background

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 which can be image, videos or text or all of them. 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 helps rescue teams and concerned authorities for their plans and actions. The concept of this 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 i] informative or not informative ii] 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 is 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.

1.2.BERT model:

A network architecture was created called Transformer that works on attention mechanism for solving problem of sequence modeling problem such as language modeling and machine translation [**4**]. Before transformer, LSTM networks were used to solve these problems but due to their computational complexity as training needs more time and for training words need to passed sequentially and also output was also generated as same way [**5**]. Even bidirectional LSTM are learning right to left and left to right separately and concatenating them and as a result true context is slightly lost [**6**]. Before LSTM, RNN networks were used for sequence to sequence modeling but unfortunately long gap dependencies problems was explored in RNN. The transformer networks are faster to train because words can be given simultaneously and contextual meaning of words are taken into consideration by learning from both directions simultaneously [**4**]. The transformer contains two key components encoder and decoder. When only decoders are stocked, GPT transformer architecture is obtained [**7**]. Conversely, when encoders are stocked BERT (Bidirectional Encoder Representation from Transformer) architecture is obtained.

BERT can be used for language translation, question answering, sentiment analysis, text summarization and many more tasks [**8**]. All of these required understanding of language. Training of BERT is done in two phases:

a. Pretrain BERT, where the model understands what language is and what the context of the words are.
b. Fine tuning BERT to do specific task, where it learns how to solve problems.

Pretraining:

The goal of pretraining is to know "what is language?" and "what is context?".
BERT can know about language by training on:

a. Masked Language Model (MLM)
b. Next Sentence Prediction (NSP)

a. Masked Language Model(MLM):
   In MLM, BERT takes some random sentence and replace some words with mask and job is to know the masked word which will help BERT understand a bidirectional context of a sentence.

b. Next Sentence Predictions(NSP):
   In NSP, BERT will be taking two sentences into account and determine if the second sentence has relation with first sentence.

With the help of these, BERT will know about the context of different sentences and with the help of sentences, it knows about language. In practice, both of them are trained simultaneously.

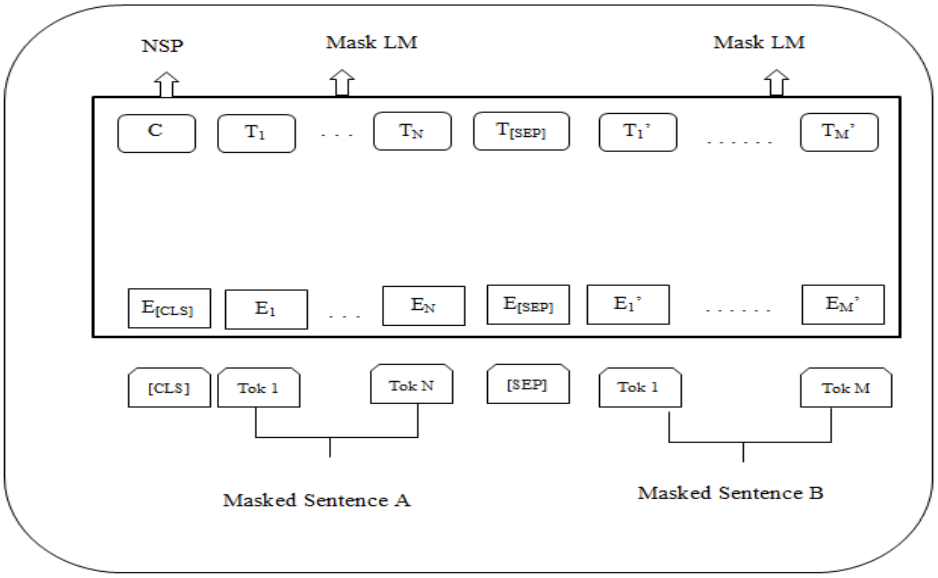


Figure 1: Pretraining of BERT

Here, input for the BERT can be two sentences and some words are masked. Each token

here is a word and each word is converted into embeddings using pretrained embeddings. In output, there is a "C" which represents either 0 or 1 for the NLP. So, if the "C" comes 1 then sentence A is followed by sentence B in context and output 0 indicates sentence A doesn't follows sentence B in context.
Each of T are word vector that correspond to outputs for MLM problem, so the no. of word vectors that we input is the same as the no. of word vectors that we output.

Here, initial embedding is constructed from 3 vectors.

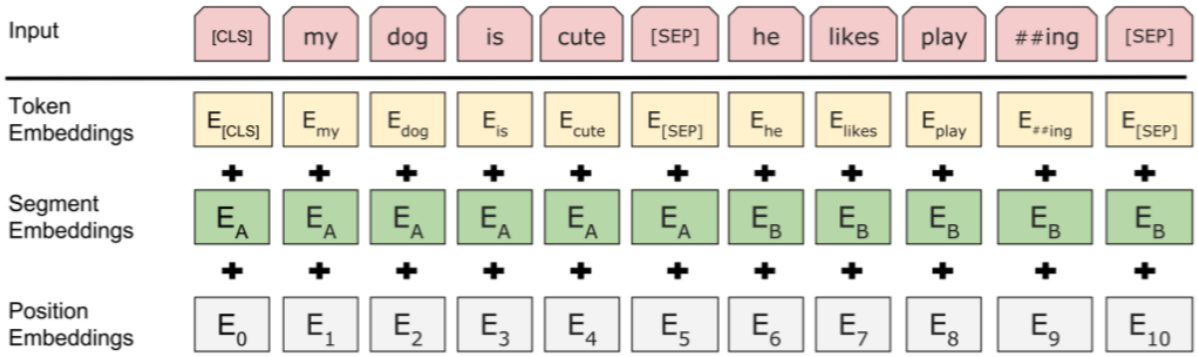


Figure 2: BERT input representation [8]

Token Embeddings are pretrained embeddings. "Wordpieces" embedding that have vocabulary of 30,000 vocabulary [**9**] is used. Segment Embedding is a sentence number and Position Embedding is position of words within the sentences that is encoded into a vector. For the input of a BERT embedding vector is used that is the addition of these three vectors. Segment embedding and position embedding add ordering for inputs. Since all these vectors are fed simultaneously into BERT and language model needs these ordering to be preserved.
A softmax activation function is applied and word vector is converted to a distribution and the actual label for this distribution will be one hot encoded vector for the actual word and these two distributions are compared and network is trained using cross entropy loss. The loss only considers the prediction of masked word and ignores all the other words that are output by network as focus is given on predicting the mask values and increases context awareness.

Fine tuning:
The goal of fine tuning is to determine "How to use language for specific task?"
Now, BERT can be trained further on very specific NLP task. For example, if we want to train BERT for text classification then input and output layer must be modified. The fully connected layers of the network must be replaced with fresh set of output layers that can basically predict the class of given sentences. Then supervised learning is performed depending on the task we want to solve, in our case with classification dataset. The time

duration must be less because it is the only parameter learned from scratch. The other model parameters needs to be fine-tuned and training time is faster.
Performance of BERT depends upon BERT size. BERT large model which have 340 million parameters can get high accuracy then BERT base model which has 110 million parameters.

1.3. VGG16 (Visual Geometry Group):

Convolutional networks have got a great success in large scale-image and video recognition [**10**] which became possible due to the large public image repositories such as ImageNet [**11**] and high performance computing system such as GPUs or large-scale distributed clusters [**12**]. In [**13**], depth of ConvNet architecture and used of very small (3 × 3) convolutional filters in all layers has been addressed. As a result, a significant ConvNet architecture (VGG16) has been achieved which got the state-of-the-art accuracy on ILSVRC classification and other image recognition datasets. The model achieved 92.7% top-5 best accuracy in Imagenet, which is a dataset of over 14 million images belonging to 1000 classes.

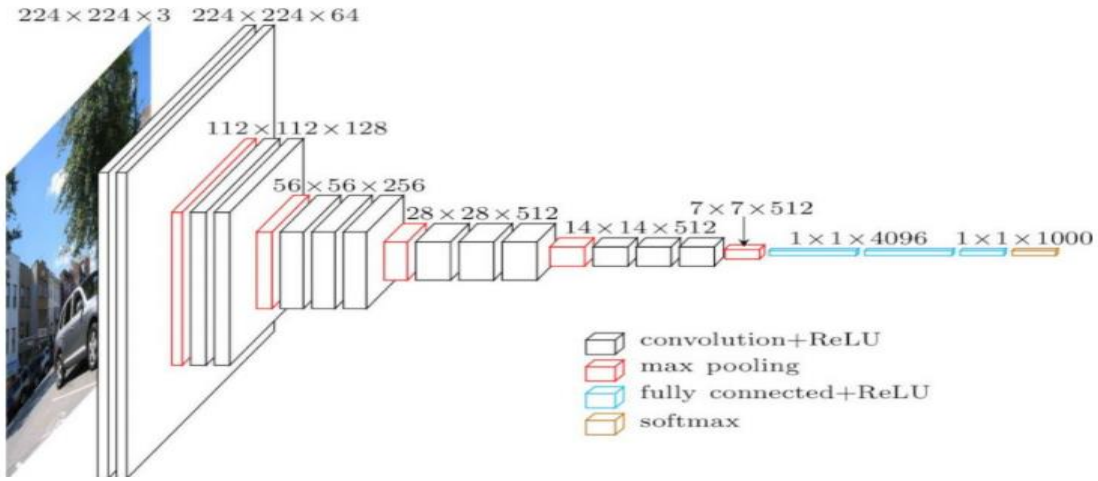


Figure 3: VGG16 Architecture [13**]**

The input to VGG16 224*224 RGB image is provided to VGG16. The provided image is passed to stack of convolutional layers where the image is convolved with fixed sized filter of 3*3. The convolutional stride performed on convolutional layer is 1 pixel and padding performed is such that spatial resolution is preserved after convolution. Spatial pooling is carried out by five max- pooling layers. Max-pooling is performed over a 2*2 pixel window, with stride 2 pixel.
Three Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer where classifications of images are obtained.

The model contains convolutional layers and maxpooling layers arranged as depicted bellow:

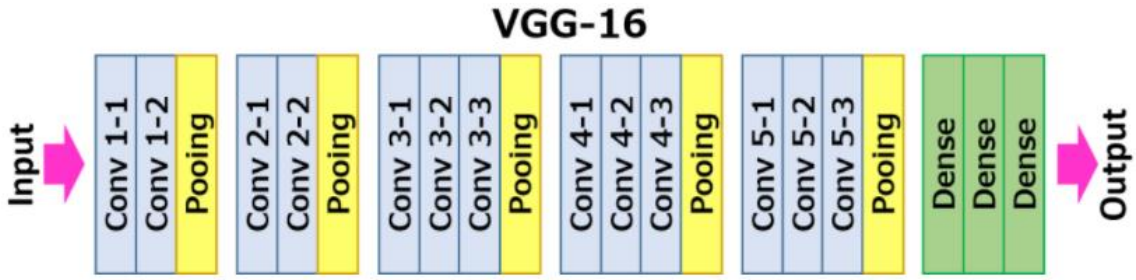


Figure 4: VGG16 [13**]**

Since, padding is same for convolution layers, when an image of 224*224*3 is passed to Conv1-1 layer, there will be the same output image as the previous layer and only the number of kernels will be changed. As the filters size is (3*3) and number of filters used is 64 so the output will be 224*224*64. But in max pooling filter size is (2*2) and stride is 3. So when image size of 224*224*3 is passed to max pooling layer the output can be calculated by eq1.

$$\frac{(I-F+2p)}{S}+1$$

(eq1)

where,

I= input image size
F = filter size
P= padding
S= stride

Similarly, for Conv1-2, being the padding same, the output image will be 112*112*128 where 128 indicates the number of filters. Similarly in every convolutional layer only the filter size is changing and padding is same but in maxpooling image dimensionality reduction is performed by keeping the important features of images preserved.
Vgg16 and VGG19 has got slight difference in accuracy. In comparison to ALEXNet, VGG16 has got more layers then ALEXNet and captures more features than ALEXNet. Since model is using RELU activation function there is no chance of vanishing gradient problem.

1.4.Problem statement

Convolutional Neural Network (CNN) used in [**14**] for text classification uses traditional embedding technique like word2vec, glove etc. Embedding performed there ignores a contextual meaning of the sentences where, polysemous words used in a sentence are not addressed. It also does not perform attention mechanism which looks relation between words with its neighbor. So the text classification based on ignoring context of words cannot give better accuracy. Similarly, the model [**15**] uses DenseNet for image

classification which decreases network computation efficiency and more prone to overfitting.

1.5. Objectives

The objectives of thesis are as follows:

a. Implementing BERT model for text analysis and VGG16 for image analysis and classifying tweet text-image pairs using post fusion technique.
b. Categorize tweet text-image pairs into five different class like affected individuals, infrastructure and utility damages, rescue and volunteering, not humanitarian and other relevant information.

1.6. Scope of work

The proposed system is focused on disaster related real image-text tweets for categorizing into humanitarian operation that will help for first responders. The proposed system does not focus on the perspective of sentiment analysis for the disaster related tweets. The sentiment analysis for the disaster related real tweets can be kept of future work.

Chapter 2

2. LITERATURE REVIEW

2.1. Classification of disasters based on textual data

Social media's content can be used to analyze to know people's conditions, their messages and their emotions according to the post shared by them. There are number of post shared on social media in the form of text, images, video and audio during disaster and after disaster. The information shared on tweeter by victims and by their relatives can be used to understand victim's conditions which will be really helpful for the first responders and rescue team to take immediate actions.

Focusing on 2012 Hurricane Sandy event, an approach has been made to take textual data posted on tweeter for identifying relevant tweets and categorizing only relevant tweets into reporting, sentiment, Information, actions, preparation and movement [**16**]. They have accessed three classification model named SVM, MaxEnt and Naïve Bayes. SVM (Support Vector Machine) performed well yielding the best F1 performance. Similarly focusing on only textual data, another study has been made to consolidate eight annotated data sources and provide 166.1k and 141.5k tweets for informativeness and humanitarian classification tasks [**17**] . The same data set is used in models to compare the performance of BERT and CNN for the classification of annotated disaster related data set. They classified textual data extracted from twitter into (i) informativeness which includes informative reports and not-informative reports), (ii) humanitarian information type classification which can include affected individual reports, infrastructure damage reports. The result of the study shows that BERT outperformed CNN. It has shown for the classification task among informativeness and not informativeness, both CNN and BERT model performed same accuracy but for the humanitarian task, the BERT model outperformed CNN model by 3.5 points in F1.During disasters, the information on social media not only comes on textual form but also on image form [**18**]. So images can also be used to detect damages and incidents.

2.2. Classification of disaster based on image data

Recently, there has been improvement in the categorization of the disaster related data by not only classifying based on text data but also based on images data. A research has been done to make a model to recognize and detect different incident based on image data [**19**]. The model is used for incident classification and incident detection by training the model with 1.1M images in dataset. It contains multitask architecture which jointly categorize images by detecting incident and place through a single architecture (by using Convolutional Neural Network (CNN) architecture with two task specific output layers. Similarly, for the analysis of the damages based on images was done on infrastructure like road, bridges, and buildings after disasters and for the categorization of the damages done after disaster, a model was proposed [**20**]. Where the study focused on the images extracted from twitter to pertinent humanitarian responses. The model classifies the disaster related images in 3 levels: (i) severe damage, (ii) mild damage, and (iii) no

damages. The disaster related images were collected from twitter using for 2 weeks using AIDR. They have used four different analysis modules to process images as image URL duplication, image deduplication, junk filtering and then unique images was finalized using severity assessment module. They have used automatic image processing system to understand the level of damages by processing 280k images and the information provided by the system was verified by domain expert and got accuracy of 76%.

Therefore disaster categorization can also be based on image data. Disaster images shared during disaster can also contain information on text format so to make an optimum analysis related to the damages and categorization of disaster can be done on relying both textual data and image data shared on social media.

2.3. Classification of disaster based on both text data and image data

Categorization of disaster related tweets can also be done based on image-data pair shared on social media which provides highly relevant information [**15**]. The multimodal framework works by fusing image and textual inputs. For the feature extraction of images, it has used DenseNet and for textual data it has used BERT model. A model [**14**] has been proposed for the categorization of disaster related tweets for humanitarian class by analyzing tweets based on both text data and image data using a joint representations. For tweet classification task, they have used deep learning based technique as CNN model. Similarly for image classification task, they have used VGG16 network. They have extracted features from both modality which is text and images and concatenated them for getting features from both modalities for classification of disasters as shown in figure 1.

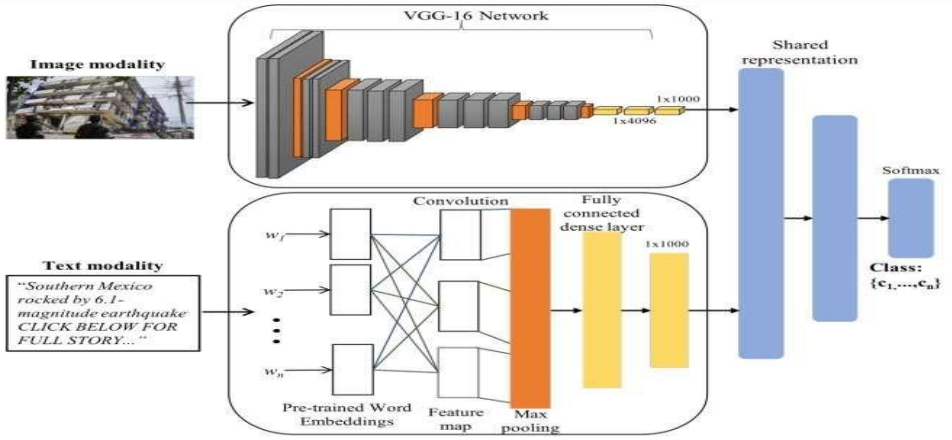


Figure 5: Multimodal architecture for the classification task using both text and image as input to the system [1]

However, in the model [**14**] the VGG16 used for extracting features of images outperforms the DenseNet used in model [**15**]. Similarly, the model [**14**] is using CNN for text classification and embedding performed here is traditional embedding. But in model [**15**] for text classification BERT algorithm is used. . A research has been proved that BERT model outperformed CNN for the disaster related text [**17**].

Chapter 3

3. METHODOLOGY

3.1. System

The 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 [**21**] and got good performance [**22**] 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.

3.2. Softmax function:

Softmax function of the model is used to convert output from the last layer of neural network into probabilities. It calculates probability for each class from where the input belongs. The number of softmax unit in output layer should always equals to the number of classes so that each unit can hold a probability of class.

$$softmax\ function = \frac{e^{yi}}{\sum_k e^{yk}}$$

Where, yi represents the raw output produced by neural network for particular class and Σyk represents the summation of all the raw output produced by neural network.

3.3. Tools and Resources

Hardware resources used are as follows:

a. Local Computer with following specifications:
   i. 4GB RAM
   ii. Intel i5 processor
   iii. GPU Raedon
   iv. 1TB harddisk
b. Cloud Platform with following specifications:
   i. GoogleColab
      - GPU NVIDIA K80
      - GPU memory 12 GB
      - Disk space 68 GB

Software used for the system

a. Python
b. GoogleColab
c. Numpy
d. Pandas
e. Tensor flow
f. Sckit-learn

Python
Python programing language is used for entire system coding.
Google Colab
Google colab is used as a platform for system development. GPU is used for training purpose.
Numpy
Numpy python package is used for numeric values.
Pandas
Pandas python package is used for accessing files and creating data frames for input and output data.
Tensor Flow
Tensorflow keras package is used for modeling machine learning model.
Sckit-learn
sckit-learn package is used for different metrics(accuracy, precision, recall, confusion matrix etc.)

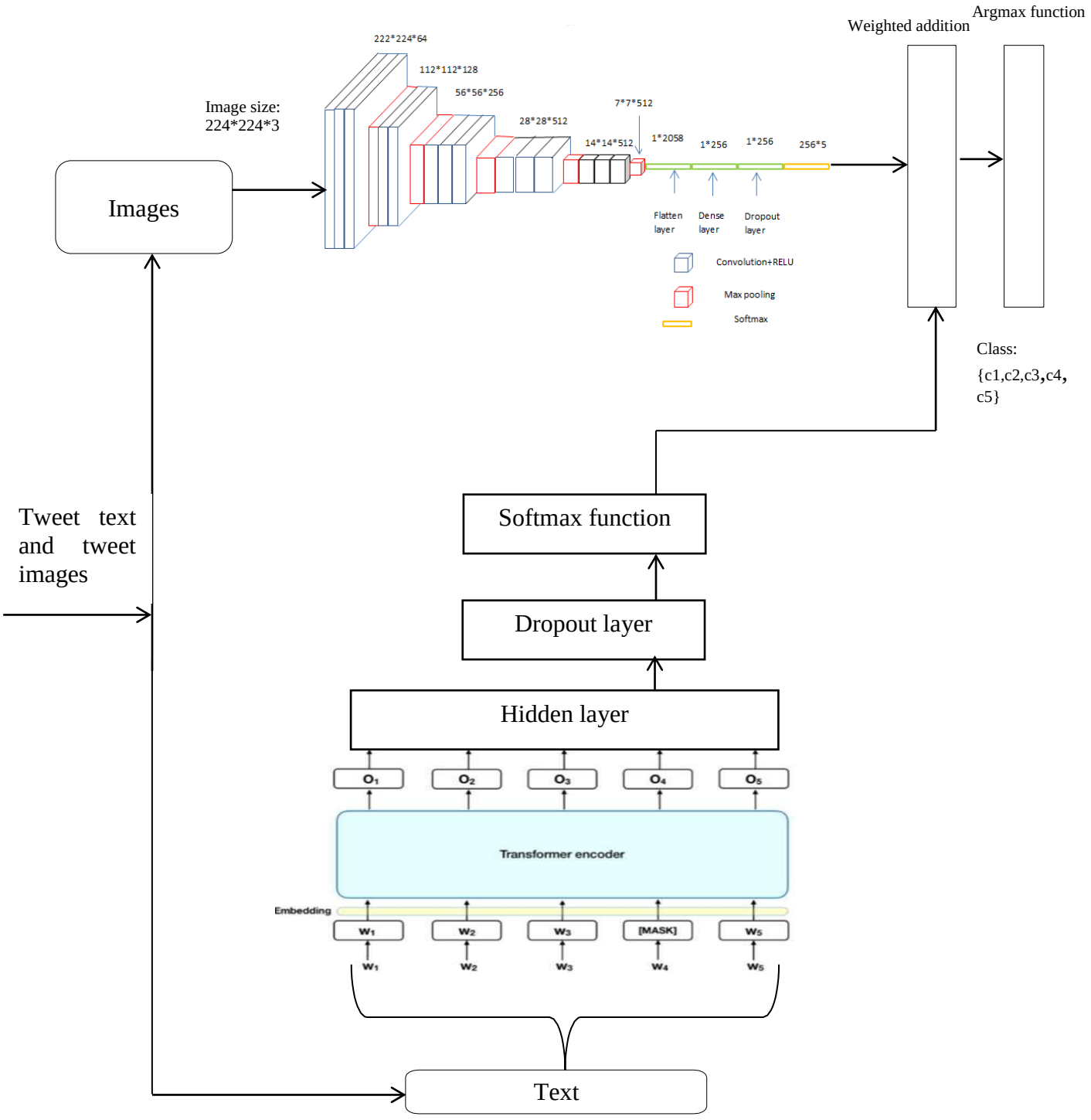


Figure 6: Methodology

3.4. Data set and Preprocessing

Multimodal dataset is used for our model, obtained from CRISISMMD which contains tweets and associated images [**21**]. 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

S.N	Category	texts	images
1.	not_humanitarian	4,394	17,57
2.	Other_relevant_information	5,686	929
3.	affected_individuals	947	775
4.	Infrastructure_and_utility_damages	1,213	2,387
5.	rescue_volunteering_or_donation_effort	3,197	1,235
	Total	15,437	7,083

(a) Hurricane Maria turns Dominica into 'giant debris field' https://t.co/rAISiAhMUy by #AJEnglish via @c0nvey https://t.co/I4zeuW4gkc

(b) A friend's text message saved Sarasota man from deadly California wildfire https://t.co/0TNMFgL885 https://t.co/CIzo44Npza

(c) Earthquake leaves hundreds dead, crews combing through rubble in #Mexico https://t.co/XPbAEIBcKw https://t.co/wGVxGD4xNd

Figure 7: Infrastructure and utility damages multimodal data

(d) @SueAikens hi su o back againe big hug FROM PUERTO RICO love you https://t.co/HCEyIHB0QZ

(e) https://t.co/jh0aQql3dR SEO ARTICLE GENERATOR https://t.co/2108RuhxgY #blogging #backlinks — Nurse fleeing California wildfires

(f) SEASON OVER???? WE COULD USE ABLE BODIES AT EARTHQUAKE IN MEXICO! DIG IN.... https://t.co/QlnYHtv9AI

Figure 8: Not_humantarian multimodal data

(g) Puerto Rico donation drive going on until 4 p.m. today and again on Oct. 28! https://t.co/zXZBrHeLCQ https://t.co/2T9k2mTCIs

(h) Raining Ash and No Rest: Firefighters Struggle to Contain California Wildfires https://t.co/G6pkvO53lJ #SocialMedia https://t.co/DRUCJ7t6G6

(i) Israeli aid team in #Mexico working day & night to find survivors #MexicoEarthquake https://t.co/UO2ZKkaisB

Figure 9: Rescue and volunteering multimodal dataset

(j) RT @ajplus: 85% of Puerto Rico remains without power. 40% of people still dont have access to drinking water. https://t.co/LKbGc7DI2R

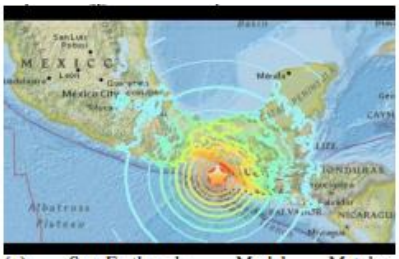
(k) RT @USRealityCheck: Homeowners cry as they return after fire https://t.co/kQIuhBCMQn #USNews #USRC https://t.co/A9ozlh2Mx1

(l) In Jojutla, Mexico, earthquake left hundreds homeless and hungry #TODAY https://t.co/jg6RFv8oHs https://t.co/iHUOYb0eEE

Figure 10: Affected individuals multimodal data

(m) #Maria remains a Category 1 Hurricane... Heavy rain by mid-week in the Outer banks https://t.co/Vm4qRPBMkY

(n) California is on fire! Please be safe out there everyone! https://t.co/dnuLv5FayS



(o) Sun-Earthquake Model Matches M8.1 in Mexico https://t.co/GEzzk9tECr https://t.co/48WWjCWw5p

Figure 11: Other relevant multimodal data

For image, data set it contains the image path labeled as different classes for the images stored in separate folders. From the image path, respective images are extracted for the individual classes and stored in drive naming individual classes. The images needs to be preprocessed before passing it to vgg16 model and for image processing, the image generator function is used to extract images from drive. The 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 contains 15,437 tweet text which are categorized into 5 different categories. The dataset is noisy and needs to be preprocessed before sending it to model for training. So, to remove noise from dataset following tasks are done:

a. converted to lowercase text

b. removed fully strip line breaks

c. replaced all URLs

d. removed all email addresses

e. removed all phone numbers

f. removed all numbers

g. removed all digits

h. removed all currency symbols

i. fully removed punctuation

j. removed all retweets.

3.5. VGG16: Image Modality

For images, VGG16 model is trained with 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 [**23**]. 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 5,256 numbers of images and validated using 1,827 numbers of images belonging to five different classes. The model is trained using Adam Optimizer. The model got 73 percent accuracy on validation dataset.

3.6. BERT model: Text modality

BERT model has achieved better performance on categorizing disaster related tweets compared to CNN [**17**]. 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 available dataset is used into train and validation set, where train data set containing 12,349 numbers of tweets is used to train the model and validation set containing 3,088 numbers 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.

3.7. 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.

Chapter 4

4. Evaluation Metrics

Performance of models can be measured using confusion matrix which contains False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN). The definition of each is given bellow:

FP: Values that are actually negative but predicted to positive.

FN: Values that are actually positive but predicted to negative.

TP: Values that are actually positive and predicted positive.

TN: Values that are actually negative and predicted to negative.

From the confusion matrix, accuracy, precision, recall and F1 score can be evaluated.

4.1. Precision: Precision is used to measure the number of right positive predictions among all positive predicted values. For example, if the model predicts 1000 numbers of tweets as not-humanitarian then precision gives the number of correct predicted tweets as not-humanitarian among predicted values.
It is calculated by using formula:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(eq2)}$$

4.2. Recall: Recall is used to measure the number of right positive predictions among all the true actual class. For example, if the model predicts 1000 numbers of tweets as not-humanitarian then recall gives us the correct predicted tweets as not-humanitarian tweets among the true values.
It is calculated by using formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(eq3)}$$

4.3. F1 score: F1 score is used to measure the test accuracy. It is the weighted mean of precision and recall. This score takes both false positives and false negatives into account.
It is calculated using formula:

$$\text{F1} = 2 * \frac{precision * recall}{(precision + recall)} \quad \text{(eq4)}$$

4.4. Accuracy: Accuracy is used to measure the performance of the model. It is the simple ratio of the correctly predicted observations to the total observations.
It can be calculated using formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(eq5)}$$

Chapter 5

5. IMPLEMENTATION AND ANALYSIS

The first two blocks of the proposed system, BERT for classifying the texts and VGG16 for images has been successfully completed. The fusion of two models outputs has been also completed. The performance of BERT model was evaluated on same multimodal dataset where it got a better accuracy compared to others model [**17**] so BERT model is selected. Models were trained using supervised learning and compared for classification performance as well. For supervised learning, 15,437 numbers of annotated data for five different categories were used to train and test the BERT model and 7,083 numbers of annotated images for five different categories were used to train and test the VGG16 model. We evaluated the performance of the models on the tweets dataset. The dataset consists of 12,349 numbers of tweet texts with five different classes with training and for validation 3,088 numbers of tweets are taken and for VGG16 5,256 numbers of images categorized into five different classes with training and validation consisting 1,827 numbers images are taken. The tweets are collected from the CrisisMMD [**21**] dataset.

5.1. BERT model

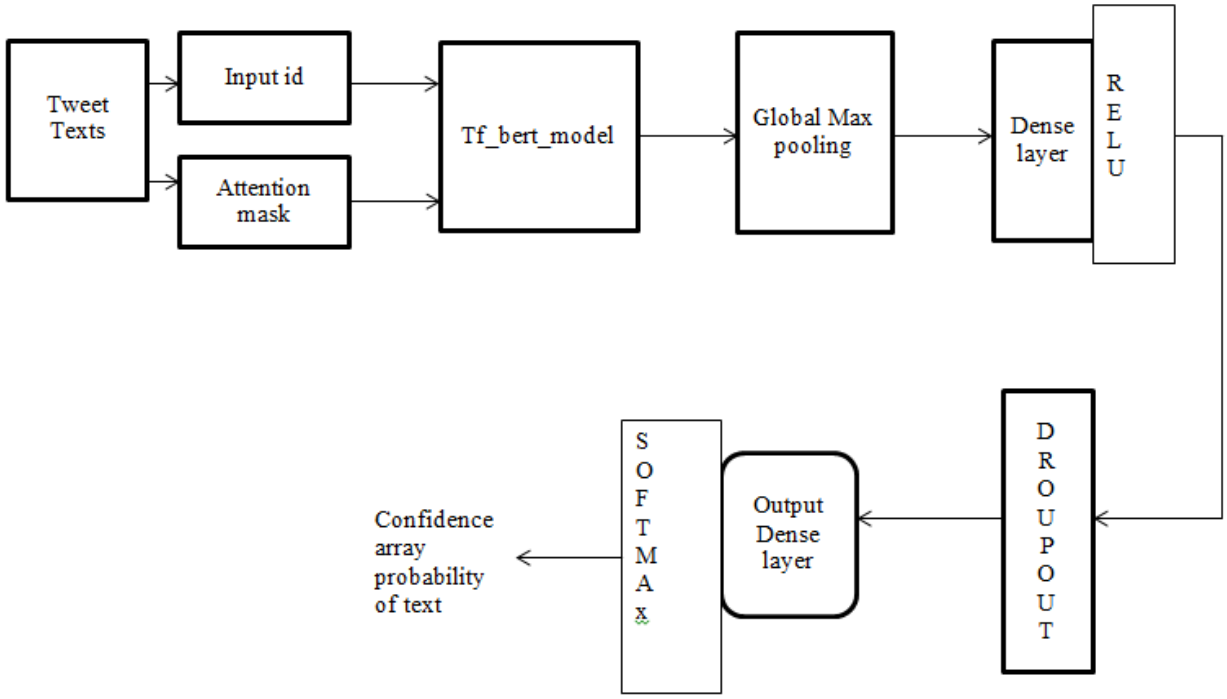


Figure 12: BERT model

To get the best performance classifier for tweet text classifications, we first downloaded a pretrained BERT model and fine-tuned it with our available categorized dataset. This is the simple supervised learning method implemented to let to know the classifier about the task. This model contains 2 input layer where 1 is attention_mask, 1 TFBert model, 1 global max, 1 dense layer of 768 units, 1 dropout layer of 0.9, and a dense layer with the

softmax activation function to categorize tweet texts in 5 categories.
To train the BERT model, I used a learning rate of 2e-05. A dropout value of 0.9 was applied in between the 2 dense layers to avoid overfitting. The batch size, the number of training samples at first of 16 texts and maximum sequence length of 128 is used. Evaluation metrics obtained by training the model and validating against the validation dataset shown in Table 2 and as confusion matrix in Figure 8.

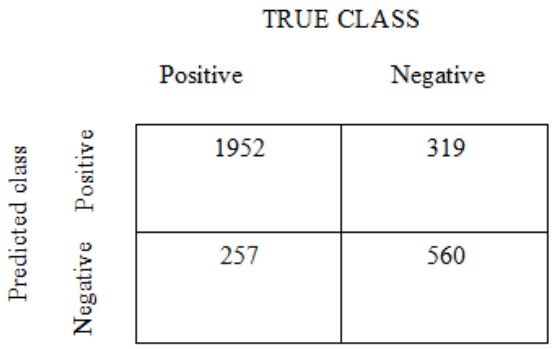


Figure 13: Confusion matrix of Bert model

Table 2: Table depicting model performance of BERT

Metrics	Performance
Accuracy	0.81
Precision	0.86
Recall	0.88
F1 score	0.86

5.2. VGG16 model

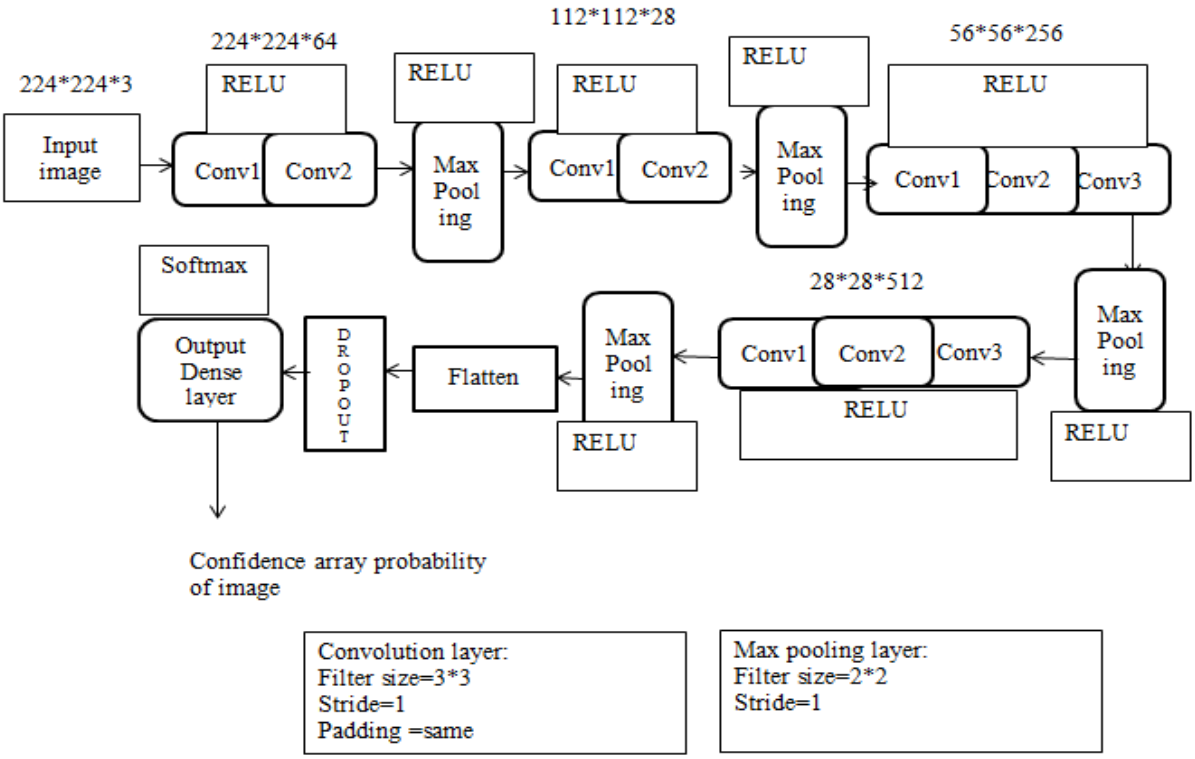


Figure 14: VGG16

To categorize the images into different classes, we first downloaded a pretrained deep convolutional network VGG16 where we can reuse the weights of VGG16 by providing the weights parameter as 'imagenet' and can be fine-tuned by removing the first and last layers of VGG16 and compute the performance of the classifier according to our task. Here, simple supervised learning method implemented to train the model with our available dataset to perform our required task. This CNN model consists of 13 convolutional layers, 5 max pooling layers with RELU activation functions and 3 fully connected layers. The last layer of VGG16 is of 1000 categories and we have the problem of classifying the images into 5 categories, so we will be putting the 5 dense layer in the output with softmax activation function. I downloaded VGG16 model and its pretrained weights. I passed pretrained weights to the VGG16 model. Then a flatten layer was added and a dense layer of 256 was added with RELU activation function. A dropout layer of 0.25 is added to avoid over fitting. According to our task, dense layer of 5 with softmax activation function is added as the final layer.

To train the model, I used a learning rate of 0.001 for same 5,256 numbers of images with Adam optimizer and tested in 1,827 numbers of images. For training the model, the epochs are set to 25 epochs. The model is trained with early stopping criteria with patience of 10 for validation accuracy got best validation accuracy as 55.30.Again, the model is trained SGD optimizer with learning rate of 0.0001 and momentum of 0.9 with 50 epochs. Figure 15 shows the plot of loss vs accuracy; which reveals the training set accuracy in 50 epochs and Figure 16 shows the plot of loss vs epochs which reveals the model loss in training set.

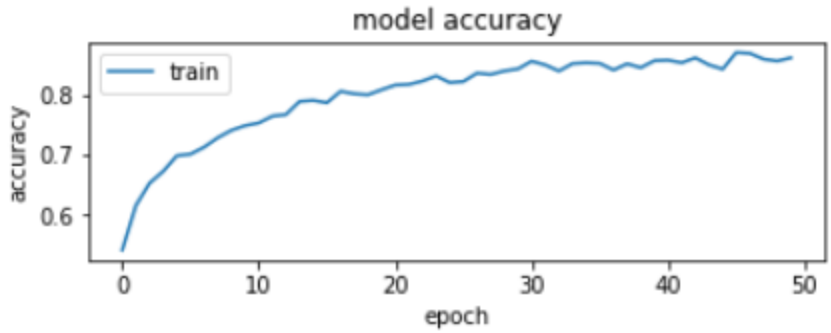


Figure 15: Accuracy vs Epochs of Vgg16 model using Adam optimizer

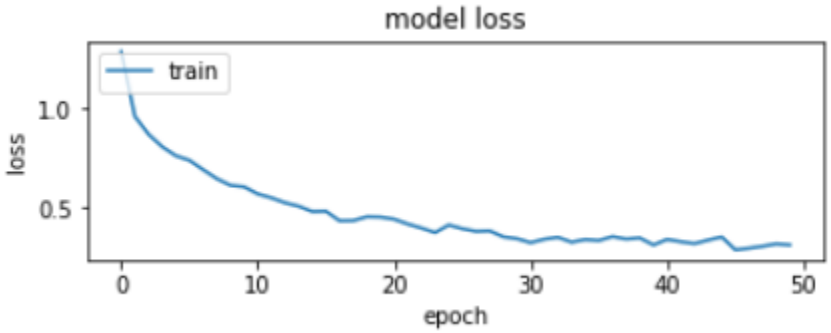


Figure 16: Loss vs Epochs of VGG16 using Adam optimizer

Evaluation metrics obtained for individual class and validating it against the validation dataset is shown in Table 3 and overall performance of model is shown in Table 4 and confusion matrix in Figure 17.

Table 3: Table depicting performance in individual class of VGG16

Class	Precision	Recall	F1-score
Affected_individuals	0.28	0.53	0.37
Infrastructure_and_utility_damage	0.69	0.8	0.74
Not_humanitarian	0.93	0.7	0.8
Other_Relevant	0.78	0.95	0.86
Rescue_voluntering_or_donation_effort	0.39	0.64	0.48

Table 4: Table depicting overall performance of theVGG16 model

Metrics	Performance
Precision	0.79%
Recall	0.73%
F1 score	0.75%
Accuracy	0.73%

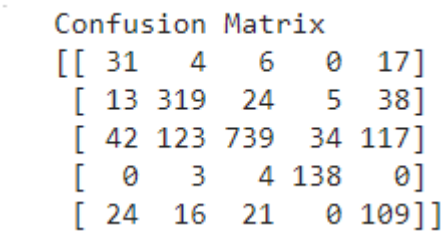


Figure 17: Confusion matrix of VGG16 model

5.3. Comparison table:

The comparison of the performance of VGG16 model and BERT model is depicted in table below:

Table 5: Comparison table of BERT and VGG16

Model	Accuracy	Precision	Recall	F1-score
VGG16	0.73%	0.79	0.73	0.75
BERT	0.81%	0.86	0.88	0.86

5.4. Multimodal Fusion:

For fusion of information from two models BERT and VGG16, the trained output layers are extracted from each model and tweet dataset containing text-images are passed as text goes to BERT layer and images goes to VGG16 for predictions. A csv file containing text-images linked with unique tweet id for each tweets is taken as a dataset. The prediction probabilities from each output layers are extracted in terms of arrays. As layers are extracted from trained model, the layers are expected for correct predictions and don't need to be further trained.

The prediction confidence of each model is compared with other and the model containing highest predictions are taken for the consideration and tweets are predicted

according to the highest confidence model. As I found that image results suppress the text results so weights are provided to each model prediction arrays. Texts result is provided with a weight of 0.63 and image result is provided with a weight of 0.37 and their confidence probabilities are compared. The model is tested on dataset of 1,150 tweets containing similar category text-image pairs but the multimodel accuracy was less even when compared with unimodal classification accuracy.

Again, the prediction probabilities of two BERT last layers and VGG16 last layers were added by using average ensembling approach where no weights were given. For testing initially our model on this approach, we take same dataset containing 1,150 tweets which contains similar category images and similar category texts and passed to the fused model. With this type of fusion, the obtained evaluation metrics are shown in table 6 and evaluation metrcis showing individual class performance is shown in table 7 and the multimodal fusion performance is compared with unimodal classification and results are shown in table 8.

Table 6: Evaluation metrics showing fusion model performance

Metrics	Performance
Accuracy	0.83
Recall	0.83
F1 score	0.84
Precision	0.86

Table 7: Evaluation metric showing individual class performance

Class	Precision	Recall	F1-score
Affected_individuals	0.28	0.85	0.42
Infrastructure_and_utility_damage	0.61	0.92	0.74
Not_humanitarian	0.91	0.85	0.87
Other_Relevant	0.92	0.80	0.86
Rescue_voluntering_or_donation_effort	0.73	0.71	0.72

Table 8: Results for the classification task

Model	Accuracy	Precision	Recall	F1-score
VGG16	0.73	0.81	0.73	0.76
BERT	0.82	0.86	0.82	0.82
Multimodal	0.83	0.86	0.83	0.84

Then, again for fusion, weighted ensebling approach was used and got following results. The weight was adjusted using hit and trial method. For the optimum weights for BERT and VGG16, weights taken was between 0.01 to 0.99 were taken as hit and trial method to get best accuracy, precision, recall and F1-score as shown in figure 18, figure 19, figure 20 and figure 21.

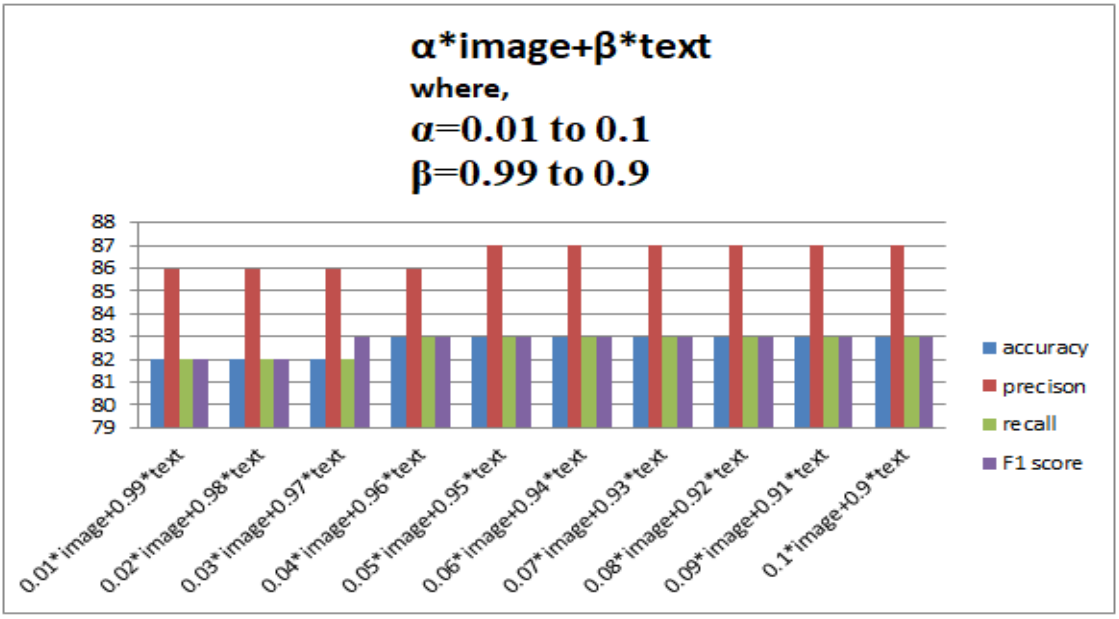


Figure 18: Evaluation metrics for weights between 0.01 to 0.1 and 0.99 to 0.9

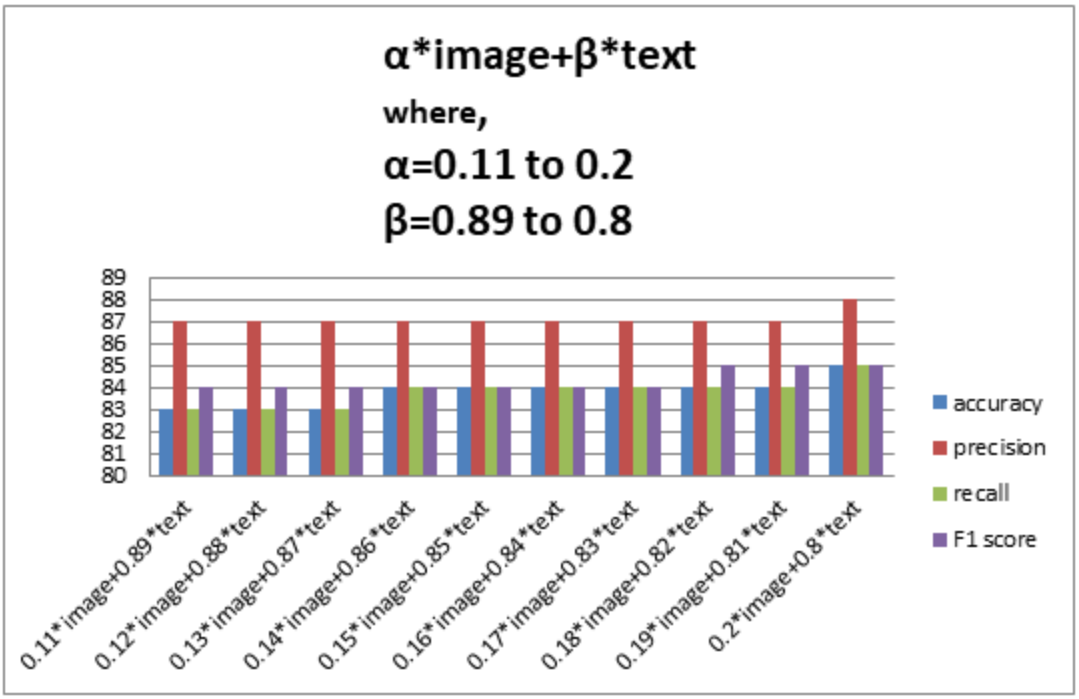


Figure 19: Evaluation metrics for weights between 0.11 to 0.2 and 0.89 to 0.8

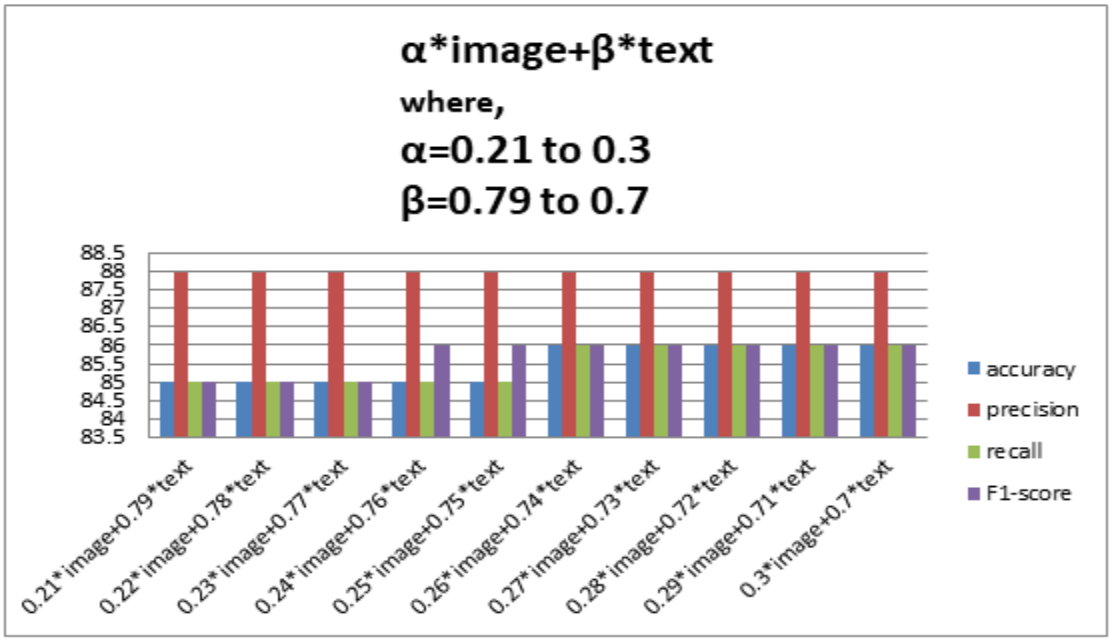


Figure 20: Evaluation metrics for weights between 0.21 to 0.3 and 0.79 to 0.7

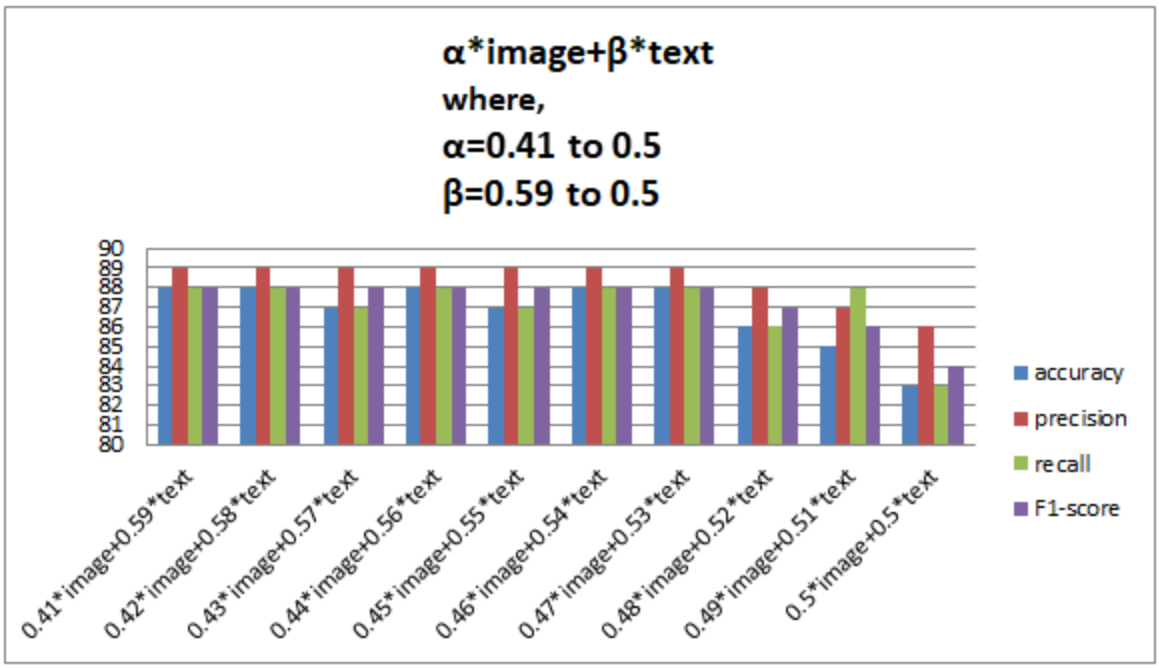


Figure 21: Evaluation metrics for weights between 0.41 to 0.5 and 0.59 to 0.5

The weights were found optimum at 0.6 to BERT and 0.4 to VGG16 with best accuracy, precision, recall and f1-score and following result were obtained as shown in table 11 and confusion matrix is shown in figure 19.

Table 9:: Evaluation metric showing individual class performance

Metrics	Performance
Accuracy	0.88
Recall	0.88
F1 score	0.88
Precision	0.89

Table 10: Evaluation metric showing individual class performance

Class	Precision	Recall	F1-score
Affected_individuals	0.55	0.92	0.69
Infrastructure_and_utility_damage	0.79	0.81	0.8
Not_humanitarian	0.97	0.85	0.91
Other_Relevant	0.81	0.95	0.88
Rescue_voluntering_or_donation_effort	0.79	0.87	0.83

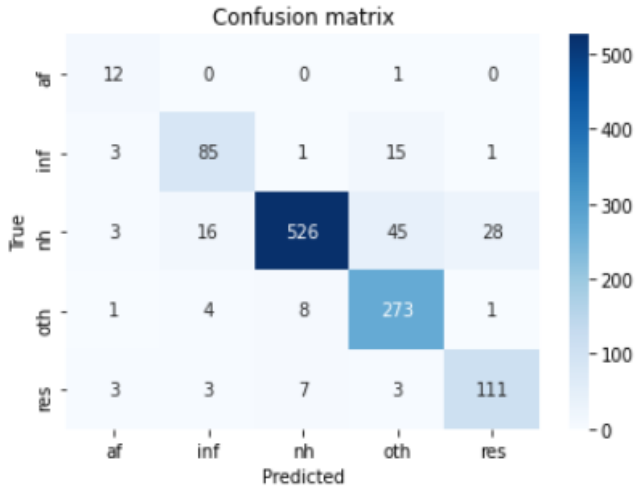


Figure 22: Confusion Matrix of Multimodal System

Table 11: Results for the classification task

Model	Accuracy	Precision	Recall	F1-score
VGG16(Image-only)	0.73	0.81	0.73	0.76
BERT(Text-only)	0.82	0.86	0.82	0.82
Multimodal(Image+text)	0.88	0.89	0.88	0.88

Our system outperforms the baseline model [**14**] in fusion of same CRISISMMD [**21**] dataset and obtained 87 percent accuracy which when compared with text modality, it is observed that multimodal system performs 5% better than BERT model and 14 % better than VGG16 model. This result confirms that multimodal classification works better than unimodal classifications.

Previously, the work has been done only on similar category text-image pairs of disaster related tweets, but our system works also on dissimilar category text-image pairs of disaster related tweets. Practically, in tweet we can find number of disaster related tweets

containing similar category text-image pairs and different category texts-image pairs too like image can be of not humanitarian and text can be of rescue and volunteering which is not addressed by previous model [**14**]. These types of tweets are also addressed by our system. Our system is tested on dissimilar and similar category text-image pairs of 1,351 numbers of tweets with weighted ensemble approach. The weight needed to be adjusted using hit and train method so weights were given between 0.01 to 0.99 to BERT and VGG16 to find best weight for greater accuracy, precision, recall and f1-score. The individual accuracy, precision, recall and f1-score on different weights on BERT output and VGG16 output is shown in figure19, figure 20, figure 21, figure 22, figure 23. The optimum weight for BERT was found with 0.79 and for VGG16 it was found 0.21 while performing hit and trial method. The weights were assigned to BERT and VGG16 for multimodal fusion and the information from BERT and VGG16 were fused using score fusion technique.

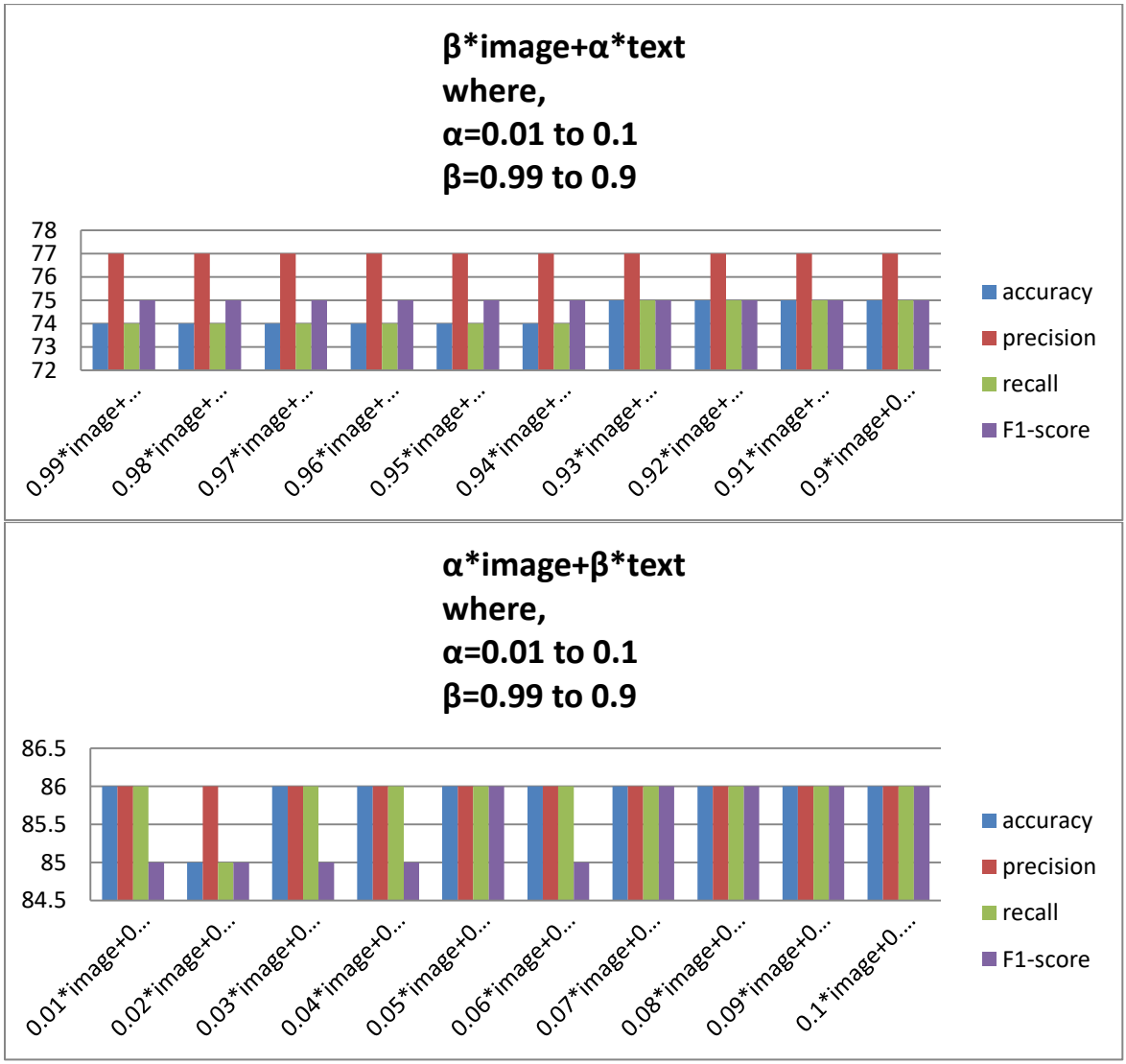


Figure 23: Evaluation metrics for weights 0.01 to 0.1 and 0.99 to 0.9

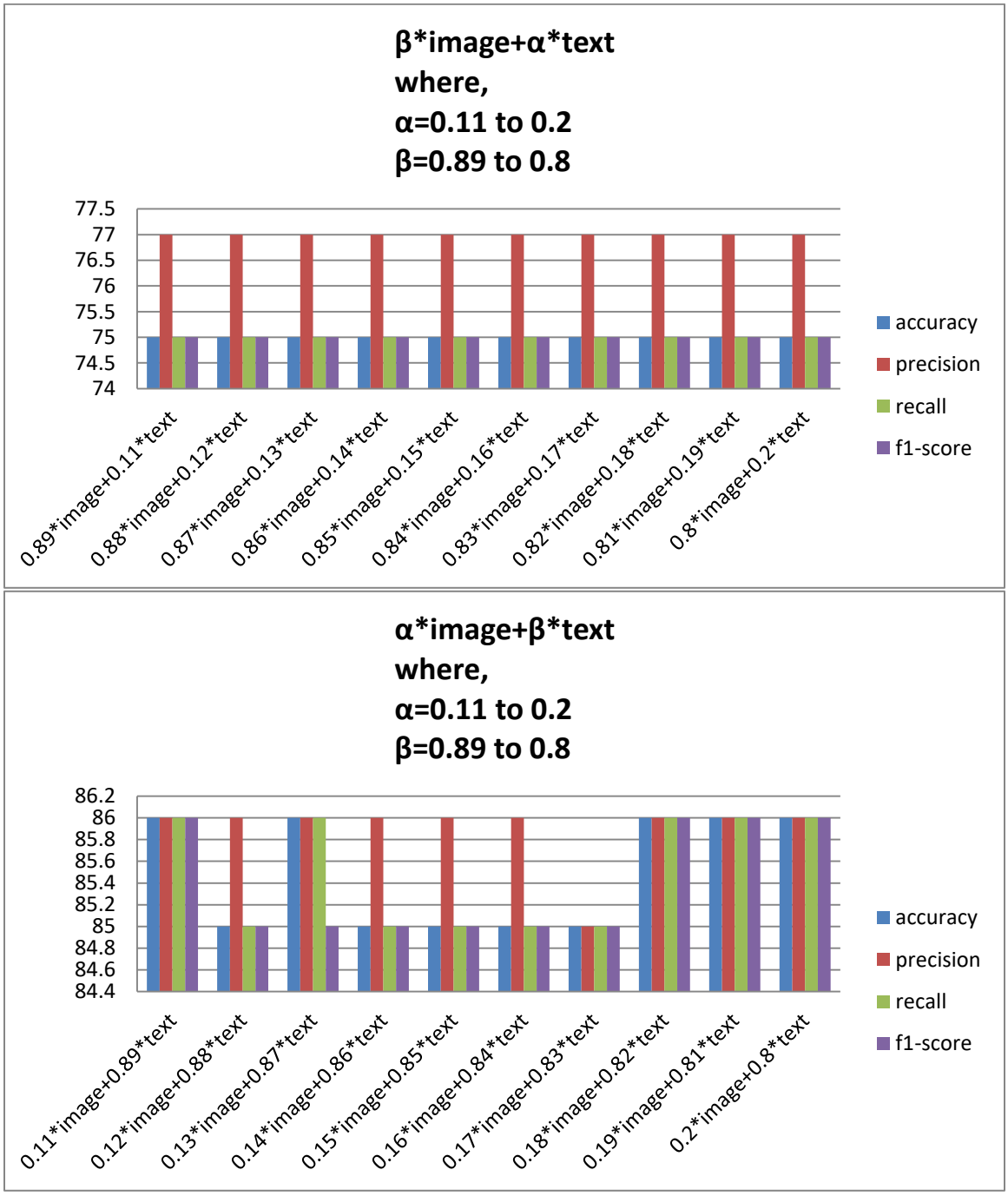


Figure 24: Evaluation metrics for weights 0.11 to 0.2 and 0.89 to 0.8

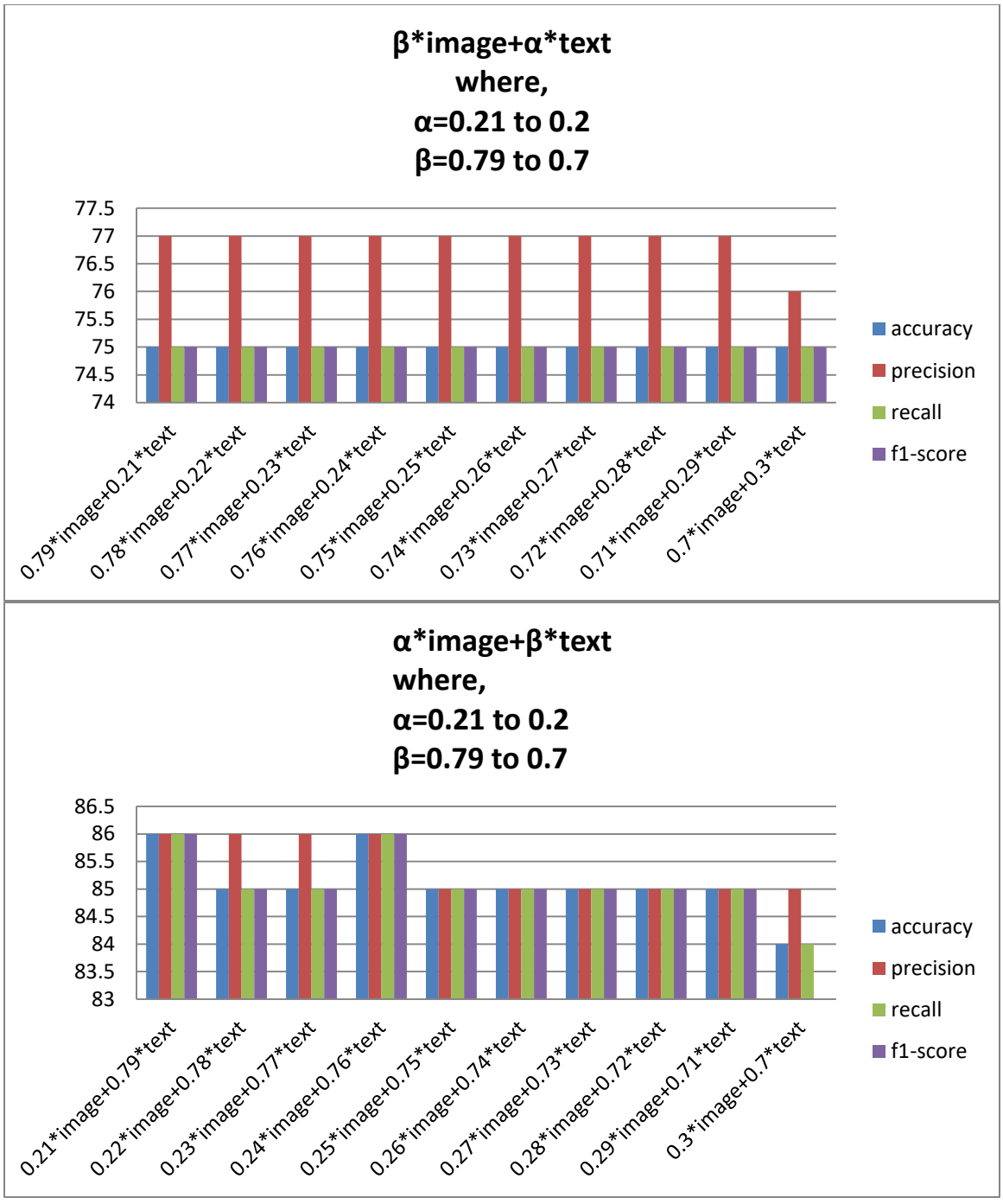


Figure 25: Evaluation metrics for weights 0.21 to 0.2 and 0.79 to 0.7

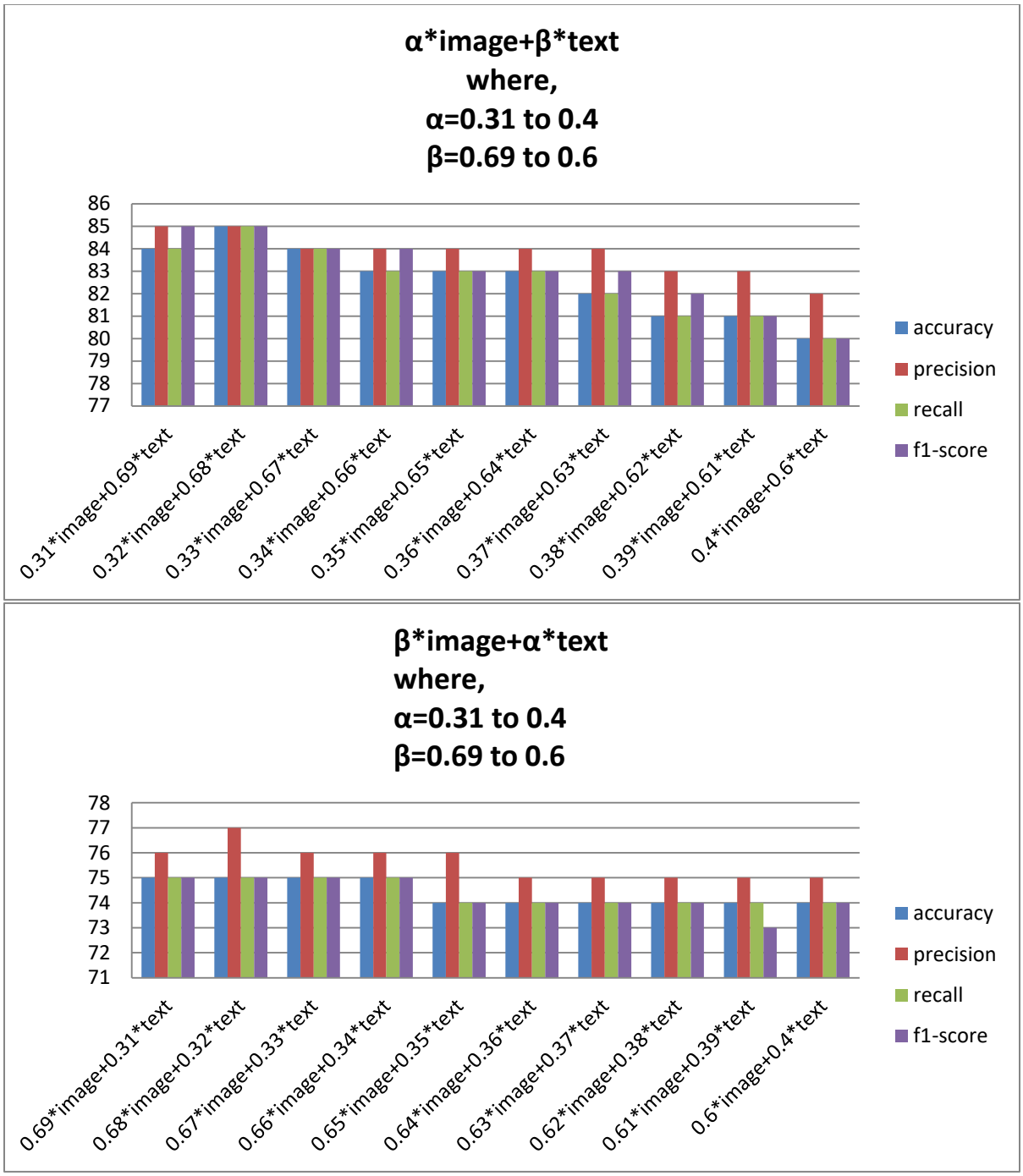


Figure 26: Evaluation metrics for weights 0.31 to 0.4 and 0.69 to 0.6

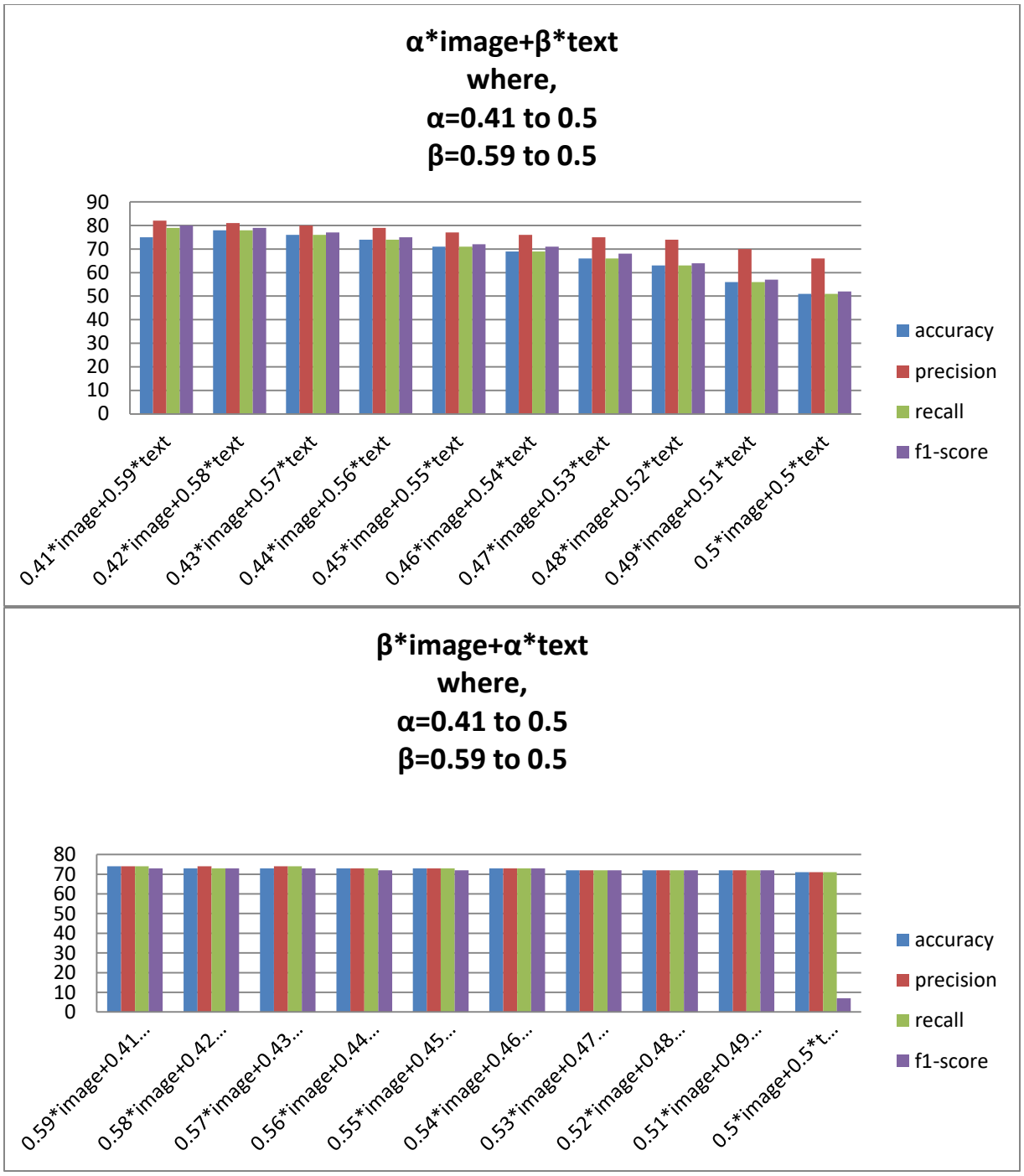


Figure 27: Evaluation metrics for weights 0.31 to 0.4 and 0.69 to 0.6

The weights for BERT is now adjusted to 0.79 and VGG16 is adjusted to 0.21 and again weights were swapped as BERT were given to 0.21 and VGG16 were given 0.79 for same similar and dissimilar category tweets and tested in multimodal system and results

obtained are shown in table 15 where accuracy for BERT has been increased from 0.8 to 0.82 percent and confusion matrix of multimodal system is shown in figure 24.

Table 12: Results for the classification task

Model	Accuracy	Precision	Recall	F1-score
VGG16(Image-only)	0.74	0.77	0.74	0.75
BERT(Text-only)	0.86	0.86	0.83	0.85
Multimodal(Image+text) (0.79*Text+0.21*Image)	0.86	0.86	0.86	0.86
Multimodal(Image+text) (0.79*Image+0.21*Text)	0.75	0.77	0.75	0.75

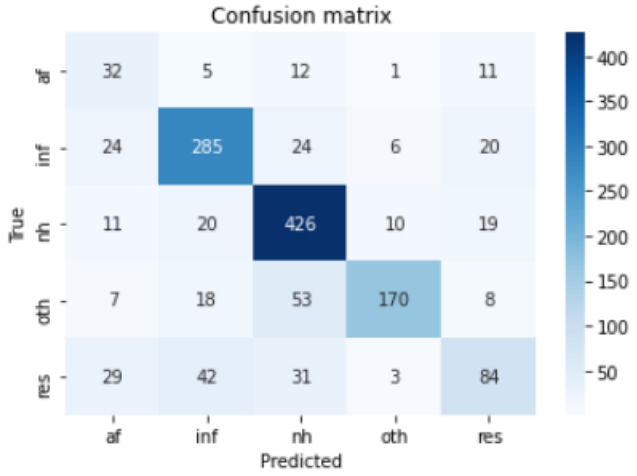


Figure 28: Confusion Matrix of Multimodal System

The confusion matrix shown in figure 19 is for the system which categorizes mixture of both similar and dissimilar tweet text-image pairs. One prominent and important column to observe here is “nh” which corresponds to not humanitarian category and shows all instances where the model prediction is not humanitarian. In particular, if we took at the instances where the actual label is “infrastructure and utility damage (denoted as inf) but the model prediction is not humanitarian (i.e the value of the cell at the intersection of row “inf” and column “nh”) we can see that the value is 24. A smiliar phenomenon can be observed for the case where actual label is rescue and volunteering (denoted as R), whereas the model predicted label is not humanitarian (i.e. the value of the cell at the intersection of res and column nh) has value 31 false negative instances. Similarly, for the case affected individual (denoted as af), whereas the model predicted label is not humanitarian (i.e. the value of the cell at the intersection of aff and column nh) has 12 false negative instances.

Chapter 6

6. CONCLUSION

6.1. Conclusion

Categorization of tweets based on both image and text is presented in this thesis. For categorization of text, BERT and for categorization of images, VGG16 has been implemented successfully and fusion of information is done with late fusion for categorization of text-image pairs. Classifying mixture of similar and dissimilar disaster related text-image pairs is a new topic of research.

For the experimentation with disaster related text-image pair tweets, 15,437 disasters related texts and 7,083 disaster related images which are connected by unique tweet id are collected and passed to BERT and VGG16 respectively. For BERT, 12,349 numbers of texts are used to train it by using hold out method and 3,088 numbers of texts are used for validating the model where the model got 81 percent accuracy and VGG16 is trained with 5,256 numbers of images and validated using 1,827 numbers of images where VGG16 got 73 percent accuracy. If there were tweets containing text-image pairs relating to disasters then how can we analyze the tweets based on both text and image? This was the main goal of the late fusion and this thesis.

Firstly, before selecting the BERT model and VGG16 for text and image classification respectively, BERT model was tested on same disaster related tweets and got greater accuracy compared to other models [**17**] and similarly VGG16 was also tested for better performance for categorizing disaster related images in same dataset [**22**]. Hence, BERT and VGG16 model were selected for categorization of disaster related text and disaster related images.

For the final experiments, after the fusion of two models for categorizing text-image pairs 1,150 numbers of tweets containing dissimilar and similar text-image pairs and linked by unique tweet id was taken. The tweet containing text-image pairs, where text were sent to BERT and image were sent to VGG16 and finally two model information were fused using weighted addition approach and got the accuracy of 82%

6.2. Future work

Categorization of disaster related tweets using multimodal approach are one of the most active research topics in disaster related works. Categorizing mixture of dissimilar and similar text-image pairs system is one of the most reliable systems for real tweets.

For further works, this research can be extended to:

- Propose a multimodal system for categorizing disaster related tweets using early fusion technique.
- Compare the multimodal system in real disaster related tweets.
- Implement the existing model to evaluate the level of damages left by disasters.

Chapter 7

7. Time schedule

TASKS	MARCH	APRIL					MAY						JUNE				
WEEKS	5	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4	5
Topic selection																	
Literature Review																	
Disaster classification using BERT																	
Disaster classification using VGG16																	
Disaster classification using Multimodal																	
Evaluation																	
Documentation																	

References

[1] Misra, L., Kumar, A., Misra, K., Aggarwal, S., and Shah, R. R. Gautam A. K., "Multimodal Analysis of Disaster Tweets," in *IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, pp. 94–103, 2019*, 2019.

[2] Cambria, E., Howard, N., Huang, G. B., and Hussain Poria S., "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," in *Neurocomputing 174, pp. 50–59*, 2016.

[3] Albanie, S., and Zisserman, A. Nagrani A., "Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Noam Shazeer,Niki Parmar,Jakob Uszkoreit,Llion Jones,Aidan N. Gomez,Illia Polosukhin Ashish Vaswani, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[5] Eric Rothstein Morris Ralf C. Staudemeyer, "Understanding LSTM –a tutorial into Long Short Term Memory RNN," September 23, 2019.

[6] Elisa Antolli, Giuseppe Serra,Carlo Tasso Marco Basaldella, "Bidirectional LSTM Recurrent Neural Network," *Research Gate*, 11 December 2018.

[7] Mu Li,Alexander J. Smola Chenguang Wang, "Language Models with Transformers," in *arXiv:1904.09408v2 [cs.CL]* , 17 Oct 2019.

[8] Ming-Wei Chang,Kenton Lee,Kristina Toutanova Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding," , 24 May 2019.

[9] Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi Yonghui Wu, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," in *arXiv:1609.08144v2 [cs.CL]*, 8 Oct 2016.

[10] A., Sutskever, I., and Hinton, G. E. Krizhevsky, "ImageNet classification with deep convolutional neural networks," in *NIPS, pp. 1106–1114*, 2012.

[11] J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L Deng, "Imagenet: A large-scale hierarchical image database," in *In Proc. CVPR*, 2009.

[12] J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang,K., Le, Q. V.,Ng, A. Y Dean, "Large scale distributed deep networks," in *In NIPS, pp. 1232–1240*, 2012.

[13] Andrew Zisserman Karen Simonyan, "VERY DEEP CONVOLUTIONAL NETOWRKS FOR LARGE-SCALE IMAGE RECOGNITION," in *ICLR 2015*, 10 April,2015.

[14] Firoj Alam, Muhammad Imran Ferda Ofli, "Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response," in *In Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Virginia, USA, 2020.

[15] Liwei Wu, Shengli Hu, Joel Tetreault, Alejandro Jaimes Mahid Abavisani, "Multimodal Categorization of Crisis Events in Social Media," in *IEEE Xplore*, 2020.

[16] Martha Palmer, Leysia Palen Kevin Stowe Michael Paul, "Identifying and Categorizing Disaster-Related Tweets," Ken Anderson University of Colorado, Boulder, CO 80309,.

[17] Hassan Sajjad, Muhammad Imran, Ferda Ofli Firoj Alam, "CrisisBench: Benchmarking Crisis-related Social Media Datasets," 17 April, 2021.

[18] Nuria Marzo, Dim P. Papadopoulos, Aritro Biswas, Agata Lapedriza ,Ferda Ofli Ethan Weber, "Detecting natural disasters, damage, and incidents in the wild," in *In Proceedings of the 16th European Conference on Computer Vision (ECCV),* , 2020.

[19] Alam, F., Ofli, F., and Imran, M. Nguyen D. T., "Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises," in *International Conference on Information Systems for Crisis Response and Management* , 2017.

[20] Firoj Alam,Umair Qazi,Steve Peterson,Ferda Ofli Muhammad Imran, "Rapid Damage Assessment UsingSocial Media Images by Combining Human and Machine Intelligence," in *In Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM)* , Virginia, USA, 2020.

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

[22] Ferda Ofli,Muhammad Imran,Tanvirul Alam,Umair Qazi, Firoj Alam, "Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Dhaka, Bangladesh, 2020.

[23] J., Clune, J., Bengio, Y., and Lipson, H Yosinski, "How Transferable Are Features in Deep Neural Networks?," in *Advances in Neural Information Processing Systems, pp. 3320–3328*, 2014).

[24] Brian Ramsay,Anca Ralescu,Esther van der Knaap Sofia Visa, "Confusion Matrix-based Feature Selection," in *Research Gate*, January 2011.

[25] Huging Face. [Online]. https://huggingface.co/