Analysis Of Multimodal Data On Social Media Using Deep Learning Techniques

Abhishek Kumar[1], Himanshu Gupta[2], Rishabh Gupta[3], Vikas Maheshkar[4]

[1]Undergraduate Student (Information Technology),
Netaji Subhas University of Technology, New Delhi,
abhishekkr.kr28@gmail.com

[2]Undergraduate Student (Information Technology),
Netaji Subhas University of Technology, New Delhi,
himanshuguptaking12@gmail.com

[3]Undergraduate Student (Information Technology),
Netaji Subhas University of Technology, New Delhi,
rishabhgupta907@gmail.com
[4]Assistant Professor,
Netaji Subhas University of Technology, New Delhi,
vikas.maheshkar@nsut.ac.in

Abstract*: *Contextual text mining known as sentiment analysis identifies and extracts subjective information from the source content. It aids in the detection of sentiments that are good, negative, neutral, etc. It helps companies monitor internet debates in order to learn how the public feels about their brands, goods, and services. However, the only metrics generally utilized in social media stream analysis are straightforward sentiment analysis and count-based metrics [13]. This is analogous to simply scratching the surface and leaving out those priceless discoveries that are just waiting to be made. Sentiment analysis is quickly evolving into a crucial tool to track and comprehend the sentiment in all types of data because people express their thoughts and feelings more freely than ever before. This project's sole objective is to use various latest AI techniques to categorize various sentiments present in audio and text forms into categories like humorous, offensive, and sarcastic. Using datasets with audio files and image files, we trained the model, then we tested it using the test data.

***Keywords*: EfficientNet, Bert, CatBoost, Transfer Learning, Spectrogram**

I. INTRODUCTION

Sentiment analysis helps us to detect positive, negative, neutral, etc sentiments by contextual text mining which proves to be a very demanding technique nowadays, be it on social media or any other platform. However, the only metrics generally utilised in social media stream analysis are straightforward sentiment analysis and count-based metrics. This is analogous to simply scratching the surface and leaving out those truly priceless discoveries that are just waiting to be made. Recent developments in deep learning have greatly increased algorithms' capacity for text analysis[10]. The inventive application of cutting-edge artificial intelligence methods can be a useful instrument for doing in-depth study. To categorize various sentiments such as sad, hilarious, cheerful, sarcastic, and so on, we used cutting-edge AI approaches such as the Efficient Net Model. We also classified the information as offensive or non-offensive, and all of these classifications are more efficient than the prior model used to accomplish the same objective. Our main focus here is on image and audio files, where we analyze every element of the image file, which consists of visuals and texts in the image. Memes, pictures, etc are some examples of image files. In the case of audio, we first transformed the audio clip into a spectrogram, which allows us to better grasp nature and hence derive more accurate categorization findings. To improve accuracy, we employed large image and audio file datasets to train our model.

A. Motivation of the work

Sentiment analysis is transforming into a vital instrument to track and comprehend the sentiments in all data types, as people express their opinions and feelings more honestly and frankly than ever before. By automatically evaluating consumer feedback from places like social media posts and survey replies, brands may learn what makes customers pleased or upset so that their goods and services can be improved.

Our decisions are driven by emotional triggers. With the aid of sentiment analysis, you may identify which remarks and discussions serve as emotional catalysts for shifting client moods. Sentiment analysis can pinpoint pressing problems in real-time, such as whether a PR crisis on social media is getting worse? Is an angry customer poised to explode? With the help of sentiment

analysis tools, you can immediately identify these types of situations and take prompt action.

II. 1.4 RELATED RESEARCH PROJECTS

A. Sound analysis

- Audio Spectrogram representations for processing with Convolutional Neural Network (CNN)

  When constructing a neural network for any application, one of the decisions that must be made is how the data should be represented in order to be presented to and possibly generated by a neural network. For audio, the choice appears to be less obvious than it appears to be for visual images, and a variety of representations, including the raw digitized sample stream, hand-crafted features, machine-discovered features, MFCCs, and variants that include deltas, and a variety of spectral representations, have been used for different applications. This study examines several of these representations and the challenges that result, with a special emphasis on spectrograms for audio generation employing neural networks for style transfer.

  Mel Frequency Cepstral Coefficients (MFCCs), which characterize the form of a spectrum, have a long history in classification, particularly in speech. Despite being a lossy representation, they are employed for classification and identification effectiveness even at very low data rates as compared to sampled audio. MFCCs have also been employed with convolutional neural networks for environmental sound classification, albeit the claimed 65% classification performance could be improved with a less lossy representation. Raw audio samples have also been used in event classification, such as SoundNet.

- PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition

  In the field of machine learning, audio pattern recognition encompasses a variety of tasks, including audio tagging, music categorization, speech emotion classification, and acoustic scene classification. Neural networks have recently been used to solve audio pattern identification issues. However, earlier systems were developed using limited-time datasets that were peculiar to them. Recently, systems that were trained on massive datasets for computer vision and natural language processing have performed well across a variety of applications.

  There is, however, little research on pre training systems on massive audio pattern recognition datasets. Other audio-related responsibilities are transferred to these PANNs. Examination of PANNs modeled by several convolutional neural networks in terms of their performance and computational complexity. There can be a design known as Wavegram-Logmel-CNN, utilizing the waveform and log-mel spectrogram as input features. With the extensive ImageNet dataset, various picture classification methods have been developed in computer vision. With the use of massive text databases like Wikipedia, numerous language models for natural language processing have been developed. Systems trained on substantial audio datasets, however, have proven more constrained.

B. Image and Text Analysis

Several works have been proposed for the identification of humour using various modalities like text and image. A Multimodal strategy for detecting humour was proposed by (Yoshida et al., 2018) [3]. The modal was created by integrating a Long Short-term Memory Network (LSTM) (Hochreiter & Schmidhuber, 1997) [8] with a Convolutional Neural Network (CNN) (Fukushima & Miyake., 1982)[7]. Further, a ResNet-152 (He et al., 2016) model was also created by the authors. The authors also developed a loss function that calculated the loss by taking a funny score into account. Reviews from several humour websites were used, where stars were used to rate them, to calculate the funniness of a post.

Another procedure, apart from humour detection, that may be applied to text-image entities is sentiment analysis (Qian et al., 2019). The authors presented two approaches. First, a combination of AffectiveSpace 2 (Cambria et al., 2015)[5] and Support Vector Machines. Second, an architecture based on CNN inspired by AlexNet (Krizhevsky et al., 2012). In terms of visual characteristics, a single fully connected layer of size 4096 x 2 was used in place of many fully connected layers which distinguishes it from AlexNet. Moreover, the textual characteristics are retrieved using a 5-fold cross-validation based on AffectiveSpace 2. Their system outperforms the other available alternatives by 7.2 percent. In comparison to a single modality, the authors found that combining visual and textual features yields better outcomes.

For the meme classification task, the image segment has received a lot of attention. In his work (Kolawole, 2015)[9], the author concentrated mostly on meme images and attempted to extract visual elements such as contrast, edges, lines or interest points such as corners or blobs. There is a lot of ambiguity in the meme classification task because a meme consists of both text and images. Through observation and research, it was found that the text is the most significant aspect of a meme.

III. METHODOLOGY

A. Dataset

a) Images

We are provided with approximately 7000 images that have been categorized into various semantic classes [1]. For subtask-A, we need to classify the meme as positive, neutral or negative. While for subtask B we need to predict whether a meme belongs to the class(humour, sarcasm, motivational and offensive) or not. The dataset

also includes text/captions taken from meme images using the OCR service and then were manually corrected to fix errors. We have used 80 per cent of the dataset for training and the rest 20 per cent for testing purposes.

TABLE I. DATASET DISTRIBUTION

Dataset	Count
Training	5593
Test	1399
Total	**6992**

b) Sound files

The AudioMP3 directory houses the MP3 audio files presented to the raters. Approximately 7400 audio files from 91 actors are given to us. These audio clips are from male and female actors between a variety of ages (approximately from 20 to 74) coming from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic, and Unspecified). A selection of 12 sentences were read by the actors. Six different emotions (angry, disgust, fear, happy, neutral, and sad) and four different emotion levels were used to deliver the sentences (Low, Medium, High and Unspecified). 80 % of the data was used for training the model and the rest 20 5 was used for testing.

B. *Data Preprocessing*

The dataset contains both the images and the text and needs to be preprocessed separately before being used in any machine learning model or a neural network. For text, BeautifulSoup was used to remove any HTML tags and regex was used to allow only alphabets and whitespaces. Also, common stopwords were removed from the text. Images were reshaped to (224, 224, 3) resolution before passing it to the EfficientNet B0 model. The EfficientNet model accepts data ranging between [0, 255]. Normalization is included as part of the model.

IV. PROPOSED APPROACH

A. *For Images and Text*

1) Image-Only approach

In this approach, transfer learning using EfficientNet [2] was used. Transfer learning is a popular approach where the pre-trained model of a task is reused as a starting point for a similar task. EfficientNet is one of the most efficient models which is mostly used for image classification tasks. EfficientNet provides a family of models (B0 to B7) that offer a solid balance of accuracy and efficiency on a range of scales. To implement transfer learning, the last prediction layer of the pre-trained EfficientNet model needs to be removed. Furthermore, the weights of the pre-trained model are frozen and are not updated during training. The output vector of the EfficientNet model is passed to a simple Feed Forward Neural Networks and the final output is obtained.

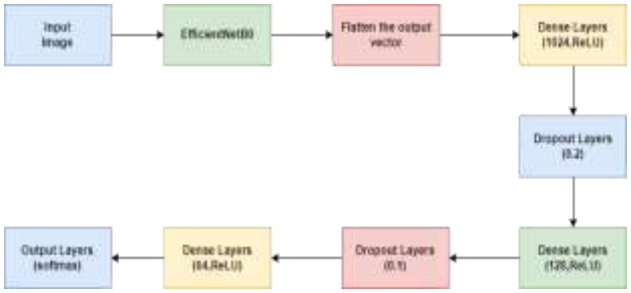

Fig. 1. Architecture model of image-based approach

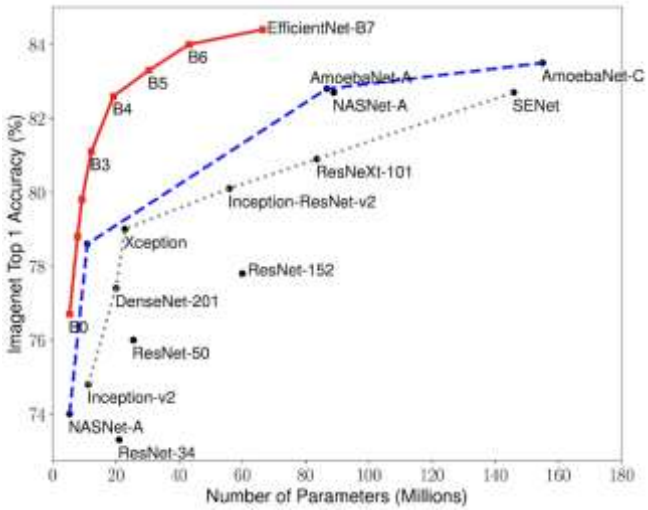

Fig. 2. Accuracy vs the number of parameters of various deep learning networks [4]

2) Text-Only approach

In this approach, we have used Bert [6] to extract the textual features. Bert stands for Bidirectional Encoder Representations from Transformers. It is a transformer-based machine learning technique, developed by Google, for natural language processing (NLP) pre-training. There are multiple Bert models available; we have used Bert-uncased for this task. The text, before being input to Bert, needs to be transformed into numeric tokens and arranged into several tokens, which is done by the preprocessing models provided by Tensorflow Hub. The pooled output obtained after passing each text through Bert is used as an input to Catboost Classifier with their labels for training. Catboost [11] is a distributed gradient boosting library used for regression and classification tasks.

Fig. 3. Architecture model of Text-based approach

3) Multi-Model approach
In multi-model approach we combined features extracted from both text and images [12]. The vector obtained after performing transfer learning on images and the pooled output vector obtained from the Bert model are concatenated. The final vector is then used as an input to Catboost Classifier.

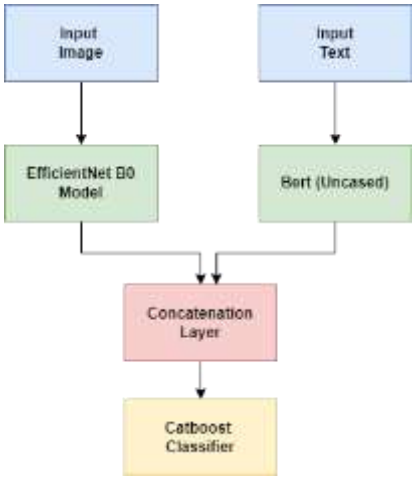

Fig. 4. Architecture model of multi-model approach

B. For Audio

The following algorithm is used for spectrogram generation: [14]
1. First, we need to import all the necessary libraries.
2. Then we need to load the audio files and visualise their waveform, which will simplify the conversion to a spectrogram.
3. Then a log transformation will be applied to the loaded audio signals.
4. Finally we pass the generated spectrogram through the image based approach discussed above and the results are obtained.

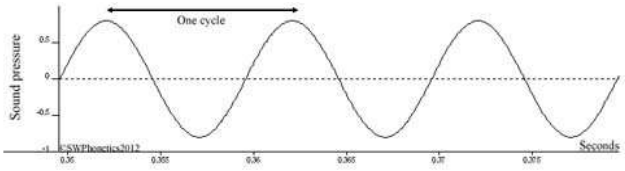

Fig. 5. Sound wave image

The y-axis represents sound pressure, the x-axis represents time.

V. Results and Discussion

A. Implementation environment

Hyper-parameters for the Neural Network model and the Catboost Classifier model are as follows.

TABLE II. NEURAL NETWORK MODEL HYPER-PARAMETERS

Hyper-parameter	Values
Optimizer	Adam
Loss	Sparse categorical crossentropy
Batch_size	20
Epochs	5 and 10

TABLE III. CATBOOST CLASSIFIER MODEL HYPER-PARAMETERS

Hyper-parameter	Values
Iterations	150
Learning Rate	1
Depth	3 and 5
Loss function	MultiClass

B. Performance result of the proposed model.

The results obtained are as follows:
1) For Images and Text
The results obtained are as follows:

TABLE IV. SUBTASK-A RESULTS

Approach Used	Overall Sentiment
Image only	0.3069
Text only	**0.3516**
Image+Text (Multimodal)	0.3329

TABLE V. SUBTASK-B RESULT

Approach Used	Humour	Sarcasm	Offense	Motivation
Image only	0.4962	0.4648	0.4288	0.4793
Text only	0.4906	**0.5076**	0.5110	0.5120
Image+Text (Multimodal)	**0.5222**	0.4979	**0.5144**	**0.5136**

The average score for subtask B is **0.5144**.

2) For Audio

TABLE VI. AUDIO RESULTS

Approach Used	Result
Spectrogram Images	0.1541

C. Comparison with existing works

The top 3 scores and the baseline score for subtask-A and subtask-B for image and text from SEMEVAL-2020 TASK 8 [15] are as follows:

TABLE VII. SUBTASK-A RESULTS AS PER THE COMPETITION

Participant / Team	Macro-F1
Vkeswani IITK	0.35466
Guoym	0.35197
Aihaihara	0.35017
Baseline	**0.21765**

TABLE VIII. SUBTASK-B RESULTS AS PER THE COMPETITION

Participant / Team	Humour	Sarcasm	Offense	Motivation	Average Score
George Vlad Eduardgzaharia UPB	0.51587	0.51590	0.52250	0.51909	0.51834
Guoym	0.51493	0.51099	0.51196	0.52065	0.51463
Sourvik Mishra Krakeu	0.51450	0.50415	0.51230	0.50708	0.50951
Baseline	**0.51185**	**0.50635**	**0.49114**	**0.49148**	**0.50021**

VI. CONCLUSION AND FUTURE WORK

The main aim of this research was to analyse the sentiments of multimodal data. We presented several techniques to find optimized solutions to the problem. Finally, we can conclude that the text-based approach outperformed the image-based approach for Classification of Images. Furthermore, the Multimodal approach surpassed the text-based approach by a very fine margin. For Audio classification we used spectrogram which is an image representation of an audio file.

In the future, we will look into the impact of adopting better transfer learning models for extracting features from the images, such as the EfficientNet(B1-B7) model, which provides better accuracy but requires a higher resolution of an image.

REFERENCES

[1] Sharma, C., & Deepesh. (2019). Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor. Memotion Dataset 7k. https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k

[2] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. https://arxiv.org/abs/1905.11946

[3] Yoshida, K., Minoguchi, M., Wani, K., Nakamura, A., & Kataoka, H. (2018). Neural Joking Machine : Humorous image captioning. https://arxiv.org/abs/1805.11850

[4] Tan, M., & V. Le, Q. (2019). EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling [Graph]. https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

[5] Cambria, E., Fu, J., Bisio, F., & Poria, S. (2015). Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In Twenty-ninth AAAI conference on artificial intelligence.

[6] Devlin, J., Chang, M.-W., Toutanova, K., & Lee, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

[7] Fukushima, K., & Miyake., S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets (pp. 267-285). Springer.

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

[9] Kolawole, O. T. (2015). Classification of internet memes. Ph.D. thesis.

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional

neural networks. In n Advances in neural information processing systems (pp. 1097-1105).

[11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. https://arxiv.org/abs/1706.09516

[12] Qian, C., Ragusa, E., Chaturvedi, I., Cambria, E., & Zunino, R. (2019). Text-image sentiment analysis.

[13] Gupta, S. (2018). Sentiment Analysis: Concept, Analysis and Applications. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

[14] MALLAM, S. (2018). Steps to convert audio clip to spectrogram. https://www.kaggle.com/code/msripooja/steps-to-convert-audio-clip-to-spectrogram

[15] Sharma, C., Bhageria, D., William Scott, & Chakraborty, T. (2020). Memotion Analysis -- The Visuo-Lingual Metaphor! https://arxiv.org/abs/2008.03781

Author Profiles

Abhishek Kumar
He is currently a B.Tech student in the Department of Information Technology at Netaji Subhas University of Technology, New Delhi. His research interests include image processing, Natural language processing and Computer Vision.

Himanshu Gupta
He is currently pursuing B.Tech in the department of Information Technology at Netaji Subhas University of Technology, New Delhi. His area of interest is Data Structures and Algorithms.

Rishabh Gupta
He is an undergraduate pursuing B.Tech in the field of Information Technology at Netaji Subhas University of Technology, New Delhi. His interests include Object Detection and Machine Learning.

Dr. Vikas Maheshkar
Dr. Vikas Maheshkar is an Assistant Professor at Faculty of Information Technology in NSIT Delhi. He earned his B.E. in Computer Technology from Nagpur University, M. Tech in Computer Science & Engineering from RGPV Bhopal and Ph.D from Motilal Nehru National Institute of Technology, Allahabad.
His current research interest includes Digital Image watermarking, Pattern Recognition, Computer Vision, Computer Forensics, Compression and Biometrics.