

Multimodal Combination of Text and Image Tweets for Disaster Response Assessment

Saideshwar Kotha, Smitha Haridasan, Ajita Rattani*, Aaron Bowen, Glyn Rimmington and Atri Dutta

¹Wichita State University, Wichita, KS, USA

Abstract

Social media platforms are a vital source of information in times of natural and man-made disasters. People use social media to report updates about injured or dead people, infrastructure damage, missing or found people among other information. Studies show that social media data, if processed timely and effectively, could provide important insight to humanitarian organizations to plan relief activities. However, real-time analysis of social media data using machine learning algorithms poses multiple challenges and requires processing large amounts of labeled data. Multi-modal Twitter Datasets from Natural Disasters (CrisisMMD) is one of the dataset that provide annotations as well as textual and image data to help researchers develop a crisis response system. In this paper, we analyzed multi-modal data from CrisisMMD, related to seven major natural calamities like earthquakes, floods, hurricanes, wildfires, etc., and proposed an effective fusion-based decision making technique to classify social media data into Informative and Non-informative categories. The Informative tweets are then classified into various humanitarian categories such as rescue volunteering or donation efforts, not-humanitarian, infrastructure and utility damage, affected individuals, and other relevant information. The proposed multi-modal fusion methodology outperforms the text tweets-based baseline by 6.98% in the Informative category and 11.2% in the Humanitarian category, while it outperforms image tweets-based baselines by 4.5% in the Informative category and 6.39% in the humanitarian category.

Keywords

Social Media, Disaster Management, Multi-modal Deep Learning

1. Introduction

Event analysis using social media is a widely researched topic and it helps in identifying trending topics, gives a sense of public sentiments about events that happen at a particular location [1]. With ever increasing access to mobile devices and social media platforms, information related to events like natural disasters is posted on social media like Facebook, Instagram, and Twitter. Although sharing of such information is useful for humanitarian support, rampant sharing of crisis related posts has led to the need of categorizing data into Informative vs Non-Informative.

International Workshop on Data-driven Resilience Research 2022, July 6, 2022, Leipzig, Germany


*Corresponding author.

✉ sxkotha3@shockers.wichita.edu (S. Kotha); sxharidasan@shockers.wichita.edu (S. Haridasan);
ajita.rattani@wichita.edu (A. Rattani); aaron.bowen@wichita.edu (A. Bowen); glyn.rimmington@wichita.edu
(G. Rimmington); atri.dutta@wichita.edu (A. Dutta)

🆔 0000-0001-8605-9908 (S. Haridasan); 0000-0002-1541-8202 (A. Rattani); 0000-0001-5511-0694 (A. Bowen);
0000-0003-2018-0099 (G. Rimmington); 0000-0003-2191-0305 (A. Dutta)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The Information is then categorized into various humanitarian aid categories. Humanitarian aid workers [2] help people during crisis events to save lives, reduce suffering and rebuild affected communities. It ensures that basic necessities like food, water, shelter, and medical assistance are provided to all affected individuals. Words in tweets such as ‘caution’, ‘help needed’, ‘warnings’, ‘rescue’, and ‘donation request’ fall into the category of Informative class. Tweets that do not shed any light related to disasters are considered Non-Informative. A social media post brings in attention and aids in getting help from Humanitarian aid workers.

Social media posts with different types (text, images, videos) of information are found to be able to provide the best results. Different modalities of information provide complementary signals about a concept, an item, or an incident. It is even more accurate to draw conclusions using a variety of methods rather than one. This well-researched approach to machine learning has been used in a variety of domains, including audio-visual analysis [3], cross-modal study [4],[5], and speech processing [6]. In order to help humanitarian organizations [2] plan, mitigate, respond to, and facilitate recovery from disasters, it is necessary to conduct time-critical analyses of the multimedia information uploaded on social media during the crisis. The majority of prior research that used social media data to analyze catastrophic occurrences focuses on text data [7]. Concentrating on a particular mode of data might miss information in some circumstances. To overcome this limitation, this research focuses on the combination of text and visual data. A powerful model has been built using these two types of data and deep learning techniques.

Social media data contains a wealth of information about environmental conditions, human activities, and geographic data that data scientists or domain scientists can analyze. Social media contains wide variety of data like text, images, and videos. For instance, timestamps, user tweets, geo-location and retweets, are also included in Twitter posts [8]. Wu et al. [9] proposed a multi-label multimodal which captures correlation and independence between modalities and can very well adapt to label inconsistency. A single tweet image can render information about the situation before and after the disaster. For example, by posting a picture of a flooded road on social media, helps others to divert to a different route. Disaster management plays a vital role in alleviating and reducing loss of life and infrastructure damage. Social media is an additional source of information beyond what was traditionally available using sensors, video streaming and satellite information can facilitate effective disaster management. The ubiquity of social media has enabled humans to post information about disaster, which can assist in disaster management.

We aim to develop a robust classification system that combines text and image modality to predict whether the tweet is informative or not and whether it is suitable for humanitarian aid workers. For this task, fusion of text and image tweets are used to implement multi-modal classification system.

In summary, the main **contributions** of this work are as follows:

- Evaluating the accuracy of text features using the model DistilBERT for classification of tweet text in CrisisMMD dataset [10].
- Analyzing the accuracy of several pre-trained CNN architectures like VGG16, VGG19 [11], ResNet50 [12], DenseNet121 [13] and RegNetY320 [14] on CrisisMMD for classification

of Image tweets.

- Comparative evaluation of various deep learning architectures on the combination of image and text tweets from CrisisMMD dataset.

The rest of this paper is structured as follows: In Section 2 prior work is discussed. Section 3 presents the dataset used in this study along with information about experiment protocols. Section 4 share insights from the experimental results for experiments on text, image and combined text and image modes. Section 5 provides the conclusions arising from this work.

2. Prior work

Several previous studies have shown that pictures posted on the internet following a crisis event may benefit humanitarian groups in a variety of ways. For example, images from Twitter may allow determination of the extent of the damage to the infrastructure. Daly et al. [15] studied the occurrence of fire in geotagged Flickr photos by considering only the images. The Fast Library for Approximate Nearest Neighbors (FLANN) was used to build a visual vocabulary of K-means, clustered key points of the image. Alam et al. [4] developed a mechanism to purify noisy social media imagery data by removing duplicate, near-duplicate and irrelevant image content. It then uses VGG-16 to classify the social media photos to extract information during an on-going crisis to create core situational awareness and to assess the severity of damage. Chen et al. [16] studies the correlation between tweet texts and tweet images in relevant and irrelevant tweets and uses SIFT descriptors, clustering them to form descriptors. Mouzannar et al. [17] conducted a study to detect damage that considered both human and environmental consequences. Six categories, including: infrastructure damage; environmental damage; injuries; and fatalities, were identified in the collection of multi-modal social media postings, which were used to examine alternative multi-modal modeling settings according to different categories. Using a decision fusion methodology for crisis-related social media data, Gautam et al. [18] examined uni-modal and multi-modal methodologies to classify tweet text and image combinations into informative and non-informative classes. Semi-supervised auto-encoding with sequential variation was developed to solve the shortcomings of a semi-supervised encoder for text classification using RNNs. LSTM structures and unlabeled data were utilized to assess the encoder's performance.

3. Dataset and Protocol

3.1. Dataset

During natural and man-made disasters, people use social media platforms such as Twitter, Facebook and Instagram to post textual and multimedia content to report updates about injuries or death, infrastructure damage and missing or found people among other information types. Studies have revealed that this information, if processed contemporaneously and effectively, is extremely useful for humanitarian organizations to gain situational awareness and plan relief operations.

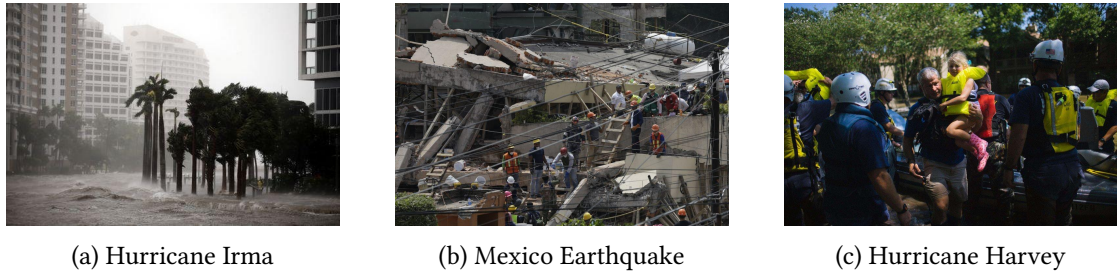


Figure 1: Image tweets from different disaster events

We used the Multimodal Crisis (CrisisMMD) [10] which consists of thousands of tweet texts and images, annotated by hand, collected during multiple major natural calamities including earthquakes, floods, hurricanes and wildfires in 2017 in various parts of the world. Fig 1 shows sample image tweets pertaining to particular disasters. The data was annotated for three tasks: (i) Informative vs Non-informative; (ii) Humanitarian Categories (8 classes); and (iii) Damage Severity Assessment (3 classes). Train, Validation and Test set is used exactly as provided by the CrisisMMD dataset rather than using 20:80 ratio which could result in overfitting.

- **Informative vs Non-Informative:** The purpose of this task is to determine whether the tweet and the image posted with the tweet during the disaster was useful for humanitarian aid purposes. If the tweet was useful for humanitarian aid, it was considered as an informative tweet. The number of data points used in the dataset for Informative vs Non-Informative tweets with text tweets, image tweets (Table - 1) and (text + image tweets) is shown in (Table-2).
- **Humanitarian Categories:** The goal of this task is to find out what information was communicated in a tweet text/image during a crisis. This information was used to classify a tweet text/image into one of the following categories: (i) Rescue volunteering or donation effect; (ii) Not-humanitarian; (iii) Infrastructure and utility damage; (iv) Other relevant information; and (v) Affected Individuals. Table-3 provides the total number of tweets in each Humanitarian category.

An important aspect of the CrisisMMD dataset is that co-occurring tweet text and image pairs have different labels for the same task because text and image modality were annotated separately and independently. Therefore, in this experiment, we consider only a subset of the original dataset where text and image pairs has the same label for a given task.

3.2. Experiments Conducted

Classification of tweets into Informative vs Non-Informative and Humanitarian categories fell into three separate classification experiments where we trained the models using: (i) text tweets; (ii) image tweets; and (iii) a fusion of text and image tweets.

Table 1

Dataset: Text and Images for Classification of Informative Vs Non-Informative Categories

	Text Modality			Image Modality		
	Informative	Non-Informative	Total	Informative	Non-Informative	Total
Train	5546	2747	8293	6345	3256	9601
Validation	1056	517	1573	1056	517	1573
Test	1030	504	1534	1030	504	1534
Total	7632	3768	11400	8431	4277	12708

Table 2

Dataset: Fusion of Text and Images for Classification of Informative Vs Non-Informative Categories

	Text + Image Modality		
	Informative	Non-Informative	Total
Train	6345	3256	9601
Validation	1056	517	1573
Test	1030	504	1534
Total	8431	4277	12708

Table 3

Dataset: Humanitarian Categories with Both Text and Image Tweets

	Train	Validation	Test	Total
Rescue volunteering or donation effort (R)	912	149	126	1187
Not-humanitarian (N)	3252	521	504	4277
Infrastructure and utility damage (I)	612	80	81	773
Other relevant information (O)	1279	239	235	1753
Affected individuals (A)	71	9	9	89
Total	6126	998	955	8079

3.2.1. Experiment #1: Text Modality for Disaster Response Assessment

Pre-trained DistilBert model [19] was used to classify tweet texts into Informative vs Non-Informative and Humanitarian categories. The Bert model was pretrained on the concatenation of two corpora: BookCorpus [20] and English Wikipedia [21]. BookCorpus was a massive collection of free novel books created by unpublished writers with 11,038 novels (about 74M phrases and 1 billion words) divided into 16 distinct sub-genres. English Wikipedia dataset contained clean articles of all languages, which was built from a Wikipedia dump with one split per language. Even though we used Bert pretrained on English Wikipedia and BookCorpus, this research work can be extended with Bert pretrained on Twitter corpus[22] to improve performance. Tweets on social media are typically cluttered with many icons, emoticons, and unseen characters. In order to remove stop words, HTML tags, URLs, alphanumeric letters, hash tags, and special characters from each tweet, we used NLTK. After preprocessing, based on empirical evidence and existing literature, each tweet text was then tokenized to a maximum length of 24 and then converted into feature vector of length 768 using the pre-trained DistilBert model(base uncased) with a batch size of 32, dropout rate of 0.2 and attention rate of 0.2 with

output hidden state True. Features vectors were then passed through a fully connected hidden layers architecture (512 – 256 – 64) followed by an output layer. RELU was used as an activation function between fully connected layers and for output layer sigmoid was used for Informative vs Non-informative classes and softmax for Humanitarian categories.

3.2.2. Experiment #2: Image Modality for Disaster Response Assessment

In order to take advantage of features from ImageNet, transfer learning was used to extract features from the images. Models were initialized with pretrained ImageNet [23] weights. CNN models like VGG16, VGG19, ResNet50, DenseNet121 and RegNetY320 models were used for classifying tweet images into Informative vs Non-informative and Humanitarian categories. For models like VGG16 and VGG19 [11], feature vector of length 4096 were extracted from fully connected layer (FC2) and for ResNet50 [12], DenseNet121 [13], RegNetY320 [14] feature vector of length 2048, 1024, 3712 respectively were extracted. Features vectors were then passed through fully connected Dense layers (2048 - 1024 - 256 - 64) followed by an output layer. Relu was used as an activation in fully connected layers and Sigmoid for output layer for Informative vs Non-informative and softmax for Humanitarian categories.

3.2.3. Experiment #3: Text and Image Modality Fusion For Disaster Response Assessment

Text feature vectors extracted from Pretrained DistilBert for tweet texts and Image feature vectors extracted from different transfer learning based CNN models for tweet images were concatenated. For example, model like Bert+VGG16 (768 + 4096), Bert+VGG19 (768 + 4096), Bert+ResNet50 (768 + 2048), Bert+DenseNet121 (768 + 1024) , Bert+RegNetY320 (768 + 3712), features vectors were then passed through a fully connected hidden layers architecture (2048 - 1024 - 256 - 64) followed by an output layer. ReLU was used as an activation between fully connected layers and for output layer as sigmoid for Informative vs Non-informative and softmax for Humanitarian categories. Adam optimizer, activation function ReLU, 100 epochs, Batch size of 32, 64, learning rates of $3e^{-4}$, $2e^{-4}$, $2e^{-3}$, $3e^{-3}$ were used as hyperparameters while conducting the following experiments with an early stopping criteria.

4. Results and Discussion

To measure the performance of algorithms used in all the three experiments, accuracy, precision, recall and F1 score are used as metrics. There have been three sets of experiments performed on text, image and text + image modalities.

4.1. Experiment #1: Performance Evaluation of DistilBERT model for Text modality

Table-4 lists the results of the performance of DistilBERT on text tweets in the classification of Informative vs Non-Informative and Humanitarian Categories. It can be observed that DistilBERT on text modality performed better than most of the state-of-the-art deep learning

Table 4

Text modality results for Informative and Humanitarian categories

Text Modality	Informative vs Non-Informative (%)	Humanitarian Categories(%)
Train Accuracy	99.98	100
Validation Accuracy	79.98	77.15
Test Accuracy	82.65	75.39
Precision	86.31	75.39
Recall	88.15	75.39
F1 Score	87.22	75.39

Table 5

Accuracy (%) for Informative vs Non-informative using Image modality

Model	VGG16	VGG19	ResNet50	DenseNet121	RegNetY320
Train Accuracy	99.86	99.80	99.67	99.79	99.90
Validation Accuracy	83.34	83.28	83.53	82.64	84.23
Test Accuracy	83.11	83.18	83.11	83.63	85.13
Precision	86.12	86.07	89.53	85.12	87.19
Recall	89.22	89.41	84.75	91.65	91.26
F1 Score	87.64	87.71	87.08	88.26	89.18

methods by providing a test accuracy of 82.65%, precision of 86.31, recall of 88.15, F1 score of 87.22 on Informative vs Non-Informative categories and 75.39% test accuracy, precision of 75.39, recall of 75.39 and F1 score of 75.39 on Humanitarian categories.

4.2. Experiment #2: Performance Evaluation of CNN models on Image modality

Table 5 summarizes the results of various deep-learning based techniques to classify social media images into Informative vs Non-Informative categories. RegNetY320 gave the best test accuracy of 85.13%. When comparing the precision and recall, ResNet50 gave a better precision of 89.53%, DenseNet121 gave a better recall of 91.65% and RegNetY320 gave an F1 score of 89.18. Table-6 shows the results of various deep learning methods to classify social media images into various humanitarian categories. While classifying Humanitarian categories using image modality. Using the pretrained deep learning models, it can be concluded that RegNetY320 gave better performance compared to other CNN models with test accuracy of 80.20%, precision of 80.20, recall of 80.20 and F1 score of 80.20.

4.3. Experiment #3: Performance Evaluation of Fusion of the Text and Image modality

Table-7 shows train accuracy, validation accuracy, test accuracy, precision, recall and F1 score of deep learning models on fusion of text and image modality in classifying informative Vs non-informative. It can be observed that Bert model on text combined with RegNetY320 on

Table 6

Accuracy (%) for Humanitarian Categories using Image modality

Model	VGG16	VGG19	ResNet50	DenseNet121	RegNetY320
Train Accuracy	99.49	99.60	98.41	99.37	99.77
Validation Accuracy	73.74	75.15	73.44	74.74	77.65
Test Accuracy	73.5	76.23	77.17	76.12	80.20
Precision	73.5	76.23	77.17	76.12	80.20
Recall	73.5	76.23	77.17	76.12	80.20
F1 Score	73.5	76.23	77.17	76.12	80.20

Table 7

Accuracy (%) for Informative vs Non-Informative using Text + Image Modality

Model	Bert + VGG16	Bert + VGG19	Bert + ResNet50	Bert + Dense121	Bert + RegNetY320
Train Accuracy	99.94	99.96	100	100	100
Validation Accuracy	86.01	86.52	89.06	88.55	89.63
Test Accuracy	86.5	86.83	88.00	89.50	89.63
Precision	86.5	86.83	88.00	89.50	89.63
Recall	86.5	86.83	88.00	89.50	89.63
F1 Score	86.5	86.83	88.00	89.50	89.63

Table 8

Accuracy(%) for Humanitarian categories using Text + Image Modality

Model	Bert + VGG16	Bert + VGG19	Bert + ResNet50	Bert + Dense121	Bert + RegNetY320
Train Accuracy	100	100	100	100	100
Validation Accuracy	80.96	81.36	82.26	83.76	84.16
Test Accuracy	82.40	81.67	84.71	85.96	86.59
Precision	82.40	81.67	84.71	85.86	86.59
Recall	82.40	81.67	84.71	85.96	86.59
F1 Score	82.40	81.67	84.71	85.96	86.59

images performs better than text only or image only results with a with a test accuracy of 89.63%, precision of 89.63, recall of 89.63 and f1 score of 89.63. Table-8 shows train accuracy, validation accuracy, test accuracy, precision, recall and F1 score of deep learning models on fusion of text and image modality in classifying Humanitarian categories. The results shows that Bert model on text combined with RegNetY320 on images performs better compared to text only or image only results with a test accuracy of 86.59%, precision of 86.59, recall of 86.59 and f1 score of 86.59. **Table-9 shows a comparison of test accuracy produced by our methods and the state-of-the-art methods which uses CrisisMMD.** Bert + RegNetY320 gave 5.23% and 8.19% increase in informative vs non-informative and humanitarian category test accuracy when compared with [5].

Table 9

Performance comparison with the State-of-the-art methods evaluated on CrisisMMD dataset

Model	Informativeness Classification Test accuracy (%)	Humanitarian Classification Test Accuracy (%)
Bert + RegNetY320	89.63	86.59
Ofli et al. [5]	84.40	78.40

5. Conclusions

The abundance of social media data clearly indicates the possibility of image processing research mainly by assisting humanitarian aid workers. This paper proposes multi-modal deep learning methodology for analyzing tweets using both textual and image tweets. Experimental results suggest that fusion of text and image tweets using Multi-modal deep learning model on CrisisMMD dataset performs better than either the single text or image modality. As part of future work, we will explore advanced fusion-techniques for combining text and image modalities and advanced deep learning models such as those based on attention mechanism to improve the classification performance.

6. Acknowledgement

We acknowledge support from Wichita State University President’s Convergent Science Initiative for conducting the research described in this paper.

References

- [1] X. Dong, D. Mavroeidis, F. Calabrese, P. Frossard, Multiscale event detection in social media, *Data Mining and Knowledge Discovery* 29 (2015) 1374–1405.
- [2] USAID, Us aid from the american people, 2021. URL: <https://www.usaid.gov/humanitarian-assistance>.
- [3] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, *IEEE Intelligent Systems* 28 (2013) 46–53.
- [4] F. Alam, F. Ofli, M. Imran, Processing social media images by combining human and machine computing during crises, *International Journal of Human–Computer Interaction* 34 (2018) 311–327.
- [5] F. Ofli, F. Alam, M. Imran, Analysis of social media data using multimodal deep learning for disaster response, *arXiv preprint arXiv:2004.11838* (2020).
- [6] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, H.-M. Wang, Audio-visual speech enhancement using multimodal deep convolutional neural networks, *IEEE Transactions on Emerging Topics in Computational Intelligence* 2 (2018) 117–128.
- [7] A. Bhoi, S. P. Pujari, R. C. Balabantaray, A deep learning-based social media text analysis

framework for disaster resource management, *Social Network Analysis and Mining* 10 (2020) 1–14.

- [8] S. Kumar, G. Barbier, M. Abbasi, H. Liu, Tweettracker: An analysis tool for humanitarian and disaster relief, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011, pp. 661–662.
- [9] X. Wu, J. Mao, H. Xie, G. Li, Identifying humanitarian information for emergency response by modeling the correlation and independence between text and images, *Information Processing & Management* 59 (2022) 102977.
- [10] F. Alam, F. Ofli, M. Imran, Crisismmd: Multimodal twitter datasets from natural disasters, in: *Twelfth international AAAI conference on web and social media*, 2018.
- [11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [13] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [14] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10428–10436.
- [15] S. Daly, J. A. Thom, Mining and classifying image posts on social media to analyse fires., in: *ISCRAM, Citeseer*, 2016, pp. 1–14.
- [16] T. Chen, D. Lu, M.-Y. Kan, P. Cui, Understanding and classifying image tweets, in: *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 781–784.
- [17] H. Mouzannar, Y. Rizk, M. Awad, Damage identification in social media posts using multimodal deep learning., in: *ISCRAM*, 2018.
- [18] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, R. R. Shah, Multimodal analysis of disaster tweets, in: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2019, pp. 94–103.
- [19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* (2019).
- [20] BookCorpus, BookCorpus, <https://www.smashwords.com/>, 2008. [Online; accessed 12-May-2022].
- [21] Wikipedia, Wikipedia, <https://dumps.wikimedia.org/>, 2008. [Online; accessed 12-May-2022].
- [22] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 9–14.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.