EAI Endorsed Transactions
on Smart Cities

Research Article EAI.EU

Multimodal Sentiment Analysis in Natural Disaster Data on Social Media

Sefa Dursun[1] and Süleyman Eken[1,*]

[1]Kocaeli University, Department of Information Systems Engineering, İzmit 41001, Türkiye

Abstract

INTRODUCTION: With the development of the Internet, users tend to express their opinions and emotions through text, visual and/or audio content. This has increased the interest in multimodal analysis methods.
OBJECTIVES: This study addresses multimodal sentiment analysis (MSA) on tweets related to natural disasters by combining textual and visual embeddings.
METHODS: The use of textual representations together with the emotional expressions of the visual content provides a more comprehensive analysis. To investigate the impact of high-level visual and texual features, a three-layer neural network is used in the study, where the first two layers collect features from different modalities and the third layer is used to analyze sentiments.
RESULTS: According to experimental tests on our dataset, the highest performance values (77% Accuracy, 71% F1-score) are achieved by using the CLIP model in the image and the RoBERTa model in the text.
CONCLUSION: Such analyzes can be used in different application areas such as agencies, advertising, social/digital media content producers, humanitarian aid organizations and can provide important information in terms of social awareness.

Received on 22 04 2024; accepted on 03 11 2024; published on 13 11 2024

Keywords: Multimodal Learning, Sentiment Analysis, Natural Disaster, Natural Language Processing, Image Processing

doi:10.4108/eetsc.5860

1. Introduction

With the advent of the Internet, social media platforms in particular have become multimodal, with content containing text, audio, images, and videos to evoke different emotions of a user. The increasing popularity of social networks and the tendency of users to share their emotions, expressions, and thoughts in text, visual, and audio content have created new opportunities and challenges in sentiment analysis [1].

While sentiment analysis from texts has been widely researched in the literature [2–5], sentiment analysis from images and videos is relatively new. Within the scope of this study, we focused on the joint analysis of visual and textual content in a socially important area and examined the effect of different visual features together with contextual text representations for multimodal tweet sentiment classification.

Based on the approach of “a picture is worth a thousand words”, visuals are an effective tool to convey not only facts but also clues about feelings and emotions. Such clues representing emotions and thoughts can trigger similar emotions in the observer and help understand visual content beyond textual concepts in different application areas such as education, entertainment, advertising, journalism, and smart cities [6]. However, it is not entirely clear how such emotional cues can be evoked by visual content and, more importantly, how emotions derived from a scene can be expressed by an automatic algorithm. In this context, the proposed system and the analyses performed; It is thought to contribute to different stakeholders such as news publishers, agencies, social/digital media content producers, humanitarian organizations, and the general public [7].

In recent years, a number of major natural disasters such as earthquakes, hurricanes, forest fires, and floods have been experienced in different parts of the world,

[*]Corresponding author. Email: suleyman.eken@kocaeli.edu.tr

as well as in our country. In these difficult times, social media plays a key role in disseminating information about the damages incurred in these disasters [8]. Within the scope of this study, multi-modal sentiment analysis of disaster images shared on social media, which is considered to be an important area in terms of social sensitivity, and the sentiments in the texts shared with the images, was carried out with a three-layer deep learning model. The data set collected from social media within the scope of multimodality (image and text) will provide a benchmark for future research in this field, and the presented analysis model can be used in other fields through transfer learning.

The remaining part of the paper is organized as follows: in the 2nd section, the relevant studies are mentioned, in the 3rd section, the visual and textual features are explained. In the 4th section, the performed tests are given, and conclusion and future works are mentioned in the last section.

2. Related Works

In recent years, multimodal learning has trended upward in AI applications as researchers integrate data in different modes/types into modeling, such as text, images, speech, etc [9]. to achieve the best results. Matrix factorization methods are among the preferred methods for multimodal data setup. Some of the common methods that have been shown to be useful in learning representations of entities from a collection of matrices are collective matrix factorization, data fusion with matrix factorization, and deep collective matrix factorization. These methods learn entity-specific representations for entities across rows and columns of matrices and use them to reconstruct the matrices. However, these methods expect the input to be a collection of matrices and tensors are not supported. Jayagopal et al. [10] overcome this limitation in their proposed model by combining the reconstruction capabilities of matrix factorization and the ability of convolutional autoencoders to operate on tensor inputs. Thus, with the proposed architecture, it can jointly learn representations for entities based on inputs (matrices or tensors) of any size.

Joint representation learning in the multimodal domain occurs in different ways. The first group of works optimizes an image encoder and a language decoder for the image captioning task and transfers the learned visual representations to subsequent applications. The second group jointly learns multimodal pretext tasks, such as reconstructing masked image regions and language symbols/tokens, as well as directly estimating the alignment between image and text. The cross-modal attention modules that arise in these methods cause them to be less efficient in practical retrieval systems. The third group, closer to the contrast methodology in visual representation learning, uses dual encoder architecture to directly map image and text data into a common embedding space. Here, the agreement between paired samples is maximized while the agreement between unpaired samples is minimized. However, recent studies focus on data scale and model architectures, using state-of-the-art multimodal contrastive learning frameworks such as ConVIRT [11] and CLIP [12], which relax the driving force between negative pairs (relaxation of contrastiveness) in pretraining, are used [13].

Sentiment analysis studies conducted on social media data refer to a field that aims to determine the emotional state of text, visuals and other content shared on social media platforms. This type of analysis can serve many different purposes. For example, within marketing and advertising it can be used to understand people's emotional reactions to products, services or brands, to better target marketing strategies and increase customer satisfaction. Again, in the field of social monitoring, it is possible to monitor the general mood about a particular event or issue on social media and use this information as an alternative to public opinion surveys. Studies in this field are becoming increasingly important with the growth of social media platforms [14]. Sentiment analysis works often use natural language processing (NLP) techniques. These techniques process text data, trying to detect specific sentiments or sentiment intensity. Sentiment analysis has been extensively explored for textual social media data, with previous dictionary-based approaches evolving into statistical and machine learning-based classification over the last decade. SentiStrength [15] is a well-known dictionary-based approach for short texts created using words and expressions commonly used on social media. Later, SentiCircles [16] is developed for Twitter sentiment analysis by taking into account the co-occurrence of words in tweets in different contexts.

With the proliferation of deep learning, sequential models such as convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks have been successfully used for tweet sentiment classification. With the increase in image and video data on social media sites such as Instagram, Flickr and Twitter, visual sentiment analysis has recently begun to attract great attention. Techniques in this scope can be generally divided into two as mid-level and deep learning representations. Using the power of pre-trained CNNs, You et al. [17] fine-tuned networks for binary sentiment classification on Flickr and Twitter image datasets. They later extended the visual sensitivity problem to sentimental image content analysis to predict sentiments such as amusement, anger, awe, fear, sadness, and excitement [18]. Recently, Jiang et al. [19] proposed another attention mechanism in which they use both cross-modal attention fusion and modality-specific CNN-transitive feature extraction to learn a better representation. They used ImageNet [20] pre-trained ResNet [21] for visual features, and GloVe [22] and BERT [23] for textual features to achieve the best results on the MVSA dataset.

Bica et al. [24] combined textual and visual content in their study, examined geotagged images on social media published during two major earthquakes in Nepal in April-May 2015, and focused on identifying the damages associated with them. Sit et al. [25] analyzed tweets to identify and categorize fine-grained details about a disaster, such as affected individuals, damaged infrastructure, and interrupted services, as well as to distinguish domains and time periods and the relative importance of each category of disaster-related information in space and time. Hassan et al. [26] proposed a deep visual emotion analyzer that covers different aspects of visual sentiment analysis, starting from data collection, annotation, model selection, application and evaluations for disaster-related images. Competitions focusing on visual sentiment analysis on natural disasters are also organized [27]. Niu et al. [28] introduce a multi-view sentiment analysis dataset (MVSA) including a set of image-text pairs with manual annotations collected from Twitter. MVSA can be utilized as a valuable benchmark for both single-view and multi-view sentiment analysis. In this study, multimodal sentiment classification from natural disaster data on social media is discussed by combining text and visual content. The use of textual representations together with the sentimental expressions of the visual content provides a more comprehensive analysis. To investigate the impact of high-level visual features, a three-layer neural network is used in the study, where the first two layers collect features from different modalities and the third layer is used for classification of sentiments.

3. Material and Methods

The human brain consists of neural networks that can process multiple modalities simultaneously. For example, when a conversation is held, the brain's neural networks process multimodal input (sound, image, text, smell). After a deep subconscious modality fusion, we are able to reason about what our conversational interlocutor is saying, his emotional state, and our environment. This approach allows for a more holistic view and deeper understanding of the situation. Within the scope of artificial intelligence's quest to imitate human intelligence, it is an inevitable fact that artificial intelligence will learn to interpret, reason and combine multimodal information to match human intelligence. In this study, studies are carried out on a deep learning model that perceives the world more holistically with multi-modal analyzes within the scope of text + image. The main idea under multimodal classification is to use different types of high-level visual features and combine them with a textual model. The used architecture is shown in Figure 1 and it is basis on Se-MLNN [14]. It combines several visual features with contextual text features to predict the overall sentiment accurately. The detailed information of the architecture are explained in the subsections.

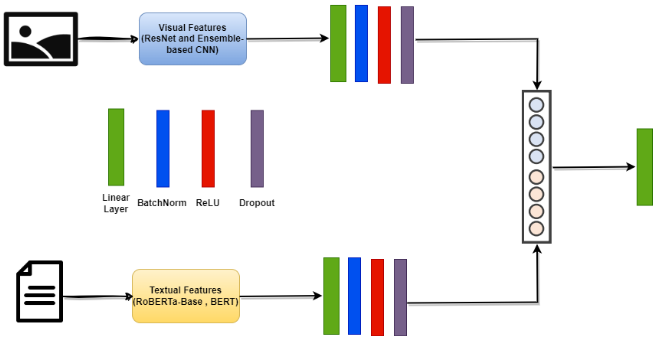

Visual Features (ResNet and Ensemble-based CNN)
Linear Layer
BatchNorm
ReLU
Dropout
Textual Features (RoBERTa-Base , BERT)

Figure 1. Architecture for multimodal sentiment learning

3.1. Visual features

Visual features within the scope of the study; It is examined in four categories: object features, place and scene features, facial expressions, and effective image content.

Object features. Object features refer to the attributes or characteristics that can be used to describe and identify an object. These features can be derived from various sensory inputs, including visual, tactile, auditory, and more. In the context of visual perception and computer vision, object features primarily involve characteristics that can be observed and analyzed from images or visual data. Different objects in an image can evoke a certain sentiment in a person. For example, while the destructive effect of the earthquake may evoke a negative emotion, objects that involve humanitarian aid organizations providing aid to earthquake-affected citizens (tents, aid parcels, volunteer helpful communities) may evoke a positive emotion. A pre-trained ResNet model on ImageNet is used to encode objects and overall image content. ResNet-50 and its final convolution layer are used instead of object categories (last layer) to extract features.

Place and scene features. Place and scene features in visual perception refer to the attributes that characterize entire environments or contexts, rather than individual objects. These features help in understanding the broader context of a visual scene, such as the type of location, the overall layout, and the relationships between various elements within the scene. A scene or place can also evoke different sentiments in people. As can be seen from the dataset prepared within the scope of the study, while people stranded in a flood disaster can evoke a negative emotion, the aid teams that come to save these people with boats can evoke a positive emotion. To encode the scene information of an image, ResNet-101 architecture pre-trained on Places365 [29] is used.

Facial expressions. Facial expressions are a subset of visual features that convey emotions, intentions, and reactions. They are crucial for social interactions and communication. The presence of faces and facial expressions (such as smiling and sad) in an image can also affect an observer's emotions. In the images in the prepared data set, it is observed that facial expressions have a direct effect on the intensity of emotion in data containing facial images. Pre-trained large ensemble-based convolutional neural networks [30] are used to encode facial expression information. Within the scope of these networks, ensembles with shared representations based on convolutional networks [31] have been studied to quantitatively and qualitatively demonstrate their data processing efficiency and scalability to large-scale facial expression datasets. In this study, facial expression analysis achieves human-level performance, outperforming state-of-the-art methods in facial expression recognition using emotion and affect concepts.

Effective image content. Effective image content as visual features refers to the attributes and elements within an image that make it engaging, informative, and aesthetically pleasing. These features contribute to the overall impact and communicative power of an image. Overall effective image content may also be important for multimodal emotion detection. This field has made rapid progress in recent years, with datasets taken from popular social media image sharing platforms such as Flickr and Instagram. To encode the overall emotion, a ResNet-50 ImageNet model is first fine-tuned on the publicly available FI (Flickr & Instagram) dataset, and last layer convolution features are extracted for object and scene embeddings. The dataset used consists of approximately 23,000 training images and eight emotion classes: amusement, anger, awe, satisfaction, disgust, excitement, fear, and sadness.

3.2. Textual features

Textual features refer to the characteristics and attributes of text that can be analyzed and used for various purposes such as natural language processing (NLP), text mining, information retrieval, and more. These features help in understanding, processing, and deriving insights from text data. Since the context and meaning of words are equally important for the emotional impact of the entire sentence, RoBERTa-Base [32] is used to extract contextual word embeddings.

Data preprocessing. Emojis, unnecessary non-ASCII characters, numbers, URL, hashtag sign and other characters in the tweet text column of the labeled dataset have been removed from the text content due to the inference that they do not help the model perform better.

Text data labeling. TextBlob library [33] is used to label the textual data in the dataset. TextBlob is a Python library for textual information that provides a simple API to access NLP activities. Using the tokenization, lemmatization, speech tagging and noun expression extraction features of the TextBlob library, each textual data line was labeled as 0 (neutral), 1 (positive), 2 (negative).

3.3. Multimodal features

Multimodal features refer to the combination of features from multiple types of data, such as text, images, audio, and video, to create a more comprehensive understanding of content. These features are essential in tasks that require the integration of information from different modalities to improve the accuracy and richness of the analysis. For the extraction of multimodal features, the multimodal CLIP [12] model, trained on 400 million image-text pairs collected from the internet, is used. The model is trained to predict which text goes with which image, and in doing so, it learns meaningful image representation without the need for millions of labeled training examples. Compared to multimodal transformers, the model uses binary learning over n-pair images and text, and no cross-attention mechanisms are used to learn multimodal features. This makes the model easier to use, as image and text embeddings can be calculated independently of the respective image and text encoders. Due to the diversity and large amount of data, the model demonstrates competitive zero-shot recognition performance on 30 different computer vision datasets compared to its supervised baselines. This shows that the amount and quality of visual information encoded in the model's visual features are much better than ImageNet and Places365 supervised pre-trained models. Both image and text features are extracted using an open-source CLIP model, where both image and text embeddings are 512-dimensional vectors.

4. Experimental Results

4.1. Test environment

The training and testing phases of the studies are carried out on Google Colab Pro on virtual servers with the following features: PC 1 (GPU Virtual Machine):

Operating system: Ubuntu 18.04.6 LTS, CPU: Intel(R) Xeon(R) CPU @ 2.20GHz, GPU : Tesla V100, RAM: 16 GB, Disk: 108 GB; PC 2 (TPU Virtual Machine): Operating system: Ubuntu 18.04.6 LTS, CPU: Intel(R) Xeon(R) CPU @ 2.20GHz, RAM: 16GB, Disk: 108 GB.

4.2. Dataset and training parameters

In the study, we create a new dataset containing 1,000 images and 1,000 texts by filtering the natural disaster hashtags and meaningful tweets containing images + text determined on Twitter. The determined natural disaster hashtags are as follows: earthquakes, landslides, droughts, famines, poverty, hurricanes, extreme precipitation and flood, wildfires. Table 1 shows the distribution in the dataset. Figure 2 contains examples of natural disasters in the dataset.

Table 1. Distribution of natural disasters in the dataset

natural disaster type	number of sample
Earthquake	245
Landslide	95
Droughts/famines	118
Poverty	92
Hurricane	136
Flood	180
Wildfire	134

Data is collected manually, via Python Tweepy service and Twitter APIs. Positive, negative and neutral categories are used in the labeling of the data. The class labels assigned to each pair (image+text) are collected under three cases: 1) If both have the same label, they are valid and the same. 2) If one label is positive or negative and the other is neutral, it is a polar (positive/negative) label. 3) If the image and text have opposite polarity tags, the tweet is a conflict. The distribution based on sentiment in the dataset, images and text, is as shown in Table 2.

In the training phase, Adam (adaptive moment estimation) is used to update the cross entropy and neural network parameters as the objective function. The learning rate is set to 2 × 10-5 and all models are trained for 100 epochs. If the validation loss does not decrease for five epochs, the learning rate is optimized to decrease by 10 times. To prevent overfitting, a dropout of 0.5 is applied after all intermediate linear layers. During training phase, the PyTorch library is used and object attributes are extracted from the publicly available ImageNet pre-trained ResNet-50 model. The ResNet-101 model, pre-trained from the Places365 dataset, is used to extract scene features.

Table 2. Distribution of sentiment in the dataset

	Positive	Negative	Neutral
Image	181	617	202
Text	289	478	233

4.3. Performance metrics

A 10-fold cross-validation metric is used in the training, validation and test set, where each bin resulted in an 8:1:1 ratio of data and the same label distribution. During the training phase, training loss, validation loss, training accuracy, validation accuracy and F1-score values are reported in each epoch.

4.4. Learning models

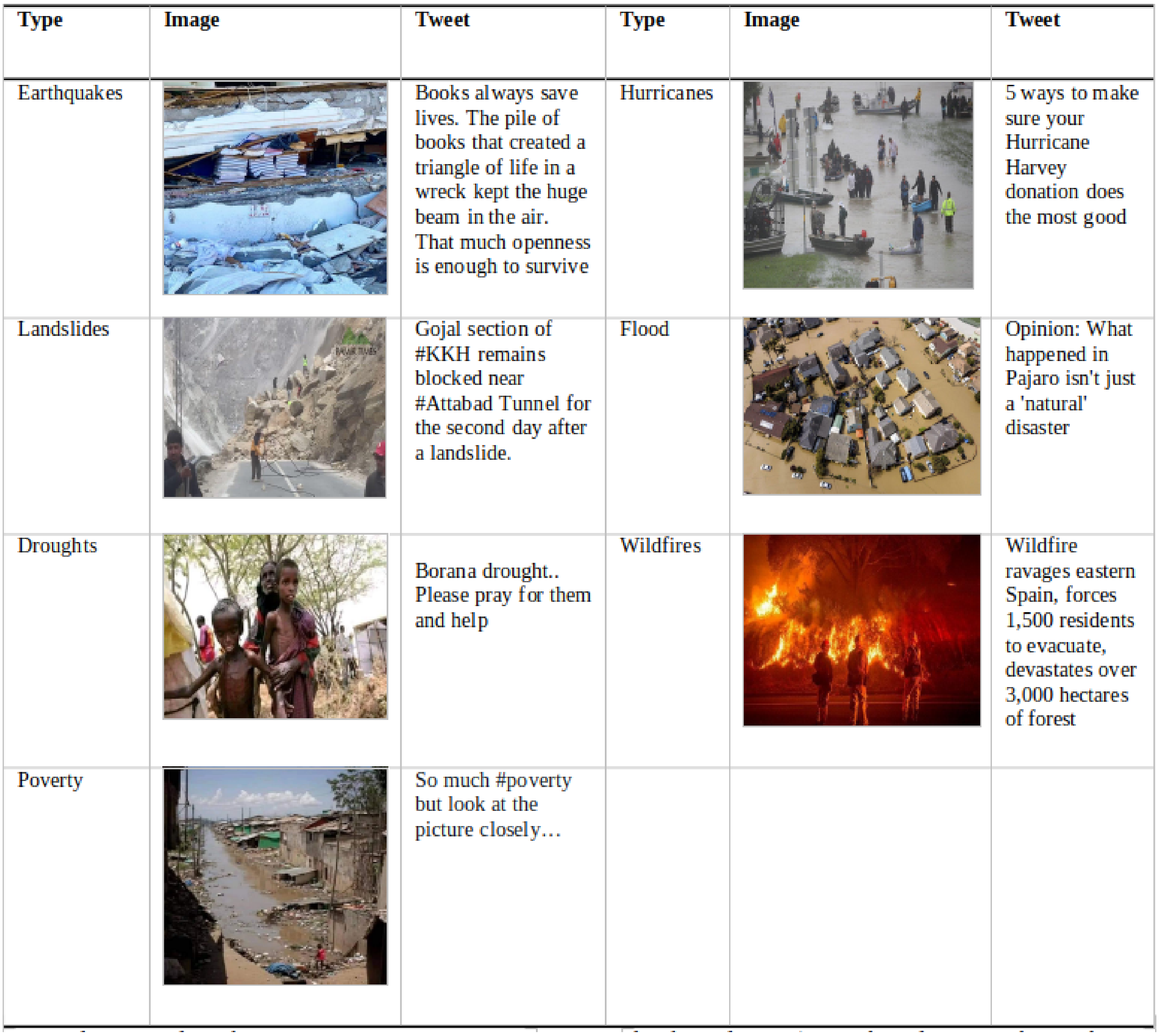
To encode objects and overall image content, a pre-trained ResNet-50 model is used to extract features on ImageNet. To encode scene information, features are extracted from a ResNet-101 model pre-trained on Places365. Pre-trained Wide Ensemble-based convolutional neural networks are used to encode facial expression information. Ensembles with shared representations (ESRs) based on convolutional networks have been studied to quantitatively and qualitatively demonstrate their scalability to facial expression datasets. RoBERTa-Base is used to extract contextual word embeddings. The multimodal CLIP model is used to extract multimodal features. Both image and text features are extracted using an open-source CLIP model, where both image and text embeddings are 512-dimensional vectors.

4.5. Performance results

The dataset is divided into 80% training, 10% validation/validation, and 10% test data in order to perform performance tests. The 10-fold cross validation method is used to verify the performance values. Table 3 presents the evaluation results of unimodal textual and visual features for the dataset. (Accuracy and F1-scores are averaged over 10-fold cross-validation.)

Table 3. Unimodal visual and textual feature results for sentiments on our own dataset

Features	Accuracy	F1-score
Object features	0.632	0.601
Facial expressions	0.615	0.502
Place and scene features	0.630	0.611
Effective image content	0.635	0.617
CLIP (Image)	0.715	**0.695**
CLIP (Text)	**0.708**	0.684
RoBERTa-Base (Text)	0.662	0.636



Type	Image	Tweet	Type	Image	Tweet
Earthquakes		Books always save lives. The pile of books that created a triangle of life in a wreck kept the huge beam in the air. That much openness is enough to survive	Hurricanes		5 ways to make sure your Hurricane Harvey donation does the most good
Landslides		Gojal section of #KKH remains blocked near #Attabad Tunnel for the second day after a landslide.	Flood		Opinion: What happened in Pajaro isn't just a 'natural' disaster
Droughts		Borana drought.. Please pray for them and help	Wildfires		Wildfire ravages eastern Spain, forces 1,500 residents to evacuate, devastates over 3,000 hectares of forest
Poverty		So much #poverty but look at the picture closely…			

Figure 2. Sample images and texts from natural disasters in the dataset

In order to compare the performance results, different models are used to extract image and text features and the results are summarized in Table 4. Figure 3 shows examples of unimodal and multimodal estimation results. The highlighted text parts in yellow are the word(s) that indicate the sentiment in the text.

As seen in Table 4, the highest performance values are achieved by using the CLIP model in the image and the RoBERTa model in the text (Accuracy 77.5%; F1-score 71%). The issues that caused this limited success performance can be attributed as follows:

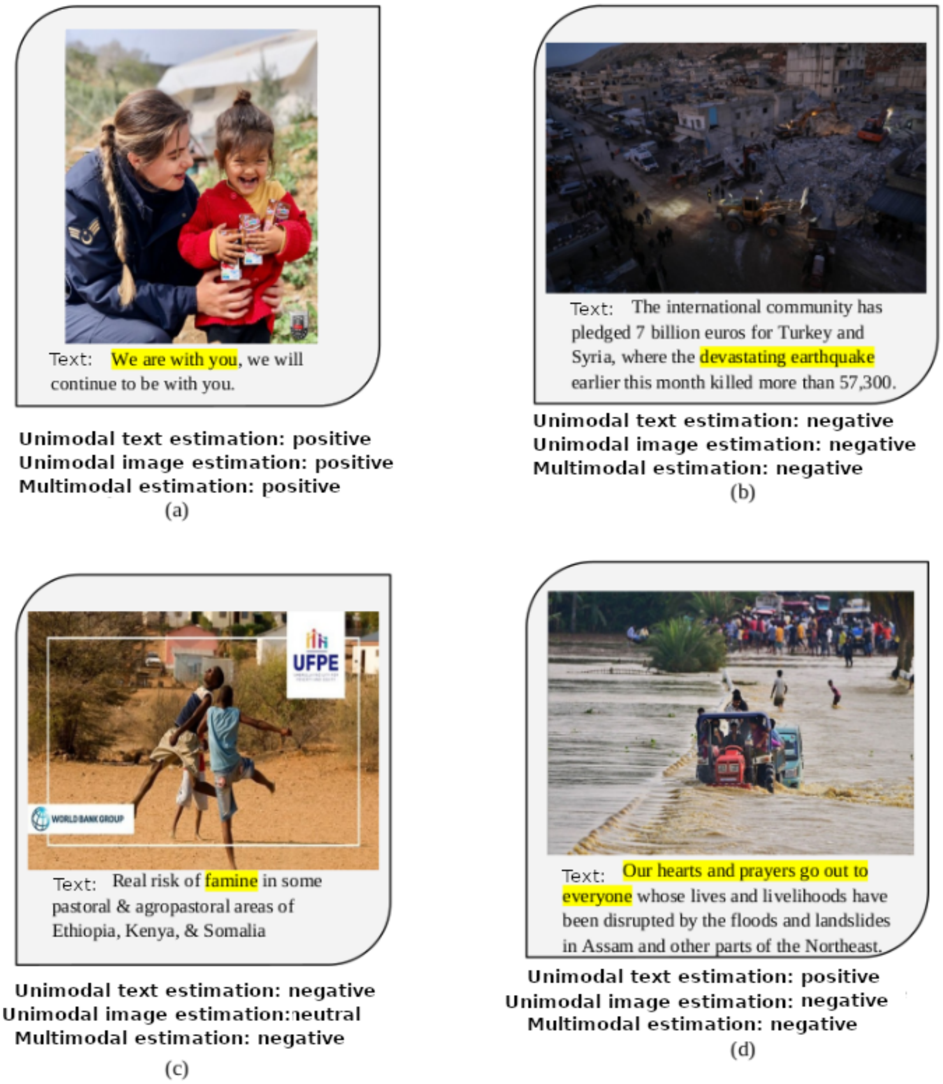
- Since the subject of the study is disaster data, the number of negative examples is quite high.
- Insufficient label pooling strategy used to pool labels, favoring positive or negative over neutral labels, resulting in a higher number of controversial labels.
- The model cannot capture interactions and cannot distinguish between neutral and polar samples.
- The data set is limited to 1,000 samples.

5. Conclusions and Future Works

In this study, an experimental evaluation of visual, textual and multimodal features is presented for MSA of natural disasters-themed tweets collected on Twitter. Experience has revealed that CLIP embeddings can

Table 4. Multimodal visual and textual feature results for sentiments on our own dataset

	Accuracy	F1-score	Accuracy	F1-score
	RoBERTa		CLIP (Text)	
Effective image content (ImageNET)	0.700	0.685	0.729	0.700
CLIP (Image)	**0.775**	0.710	0.701	0.750



Text: We are with you, we will continue to be with you.

Unimodal text estimation: positive
Unimodal image estimation: positive
Multimodal estimation: positive
(a)

Text: The international community has pledged 7 billion euros for Turkey and Syria, where the devastating earthquake earlier this month killed more than 57,300.

Unimodal text estimation: negative
Unimodal image estimation: negative
Multimodal estimation: negative
(b)

Text: Real risk of famine in some pastoral & agropastoral areas of Ethiopia, Kenya, & Somalia

Unimodal text estimation: negative
Unimodal image estimation:neutral
Multimodal estimation: negative
(c)

Text: Our hearts and prayers go out to everyone whose lives and livelihoods have been disrupted by the floods and landslides in Assam and other parts of the Northeast.

Unimodal text estimation: positive
Unimodal image estimation: negative
Multimodal estimation: negative
(d)

Figure 3. Examples for unimodal and multimodal prediction results

serve as a strong basis for the task of multimodal sentiment prediction in tweets. It is envisaged that the natural disasters themed dataset prepared by collecting from social media within the scope of multimodality (image and text) may be useful for future research in this field. While MSA can provide richer insights compared to unimodal analysis, it also faces several limitations. Social media data is noisy, with many irrelevant or low-quality posts. There is often a lack of labeled datasets that contain multimodal information specific to natural disasters. Sentiments during a natural disaster can change rapidly, requiring models to account for temporal dynamics. Using social media data raises concerns about user privacy and data security. Future work can address these challenges to improve the effectiveness and reliability of MSA in this context. In the future, it is planned to collect more images and text from different social media platforms and carry out activities to reduce bias. Also, we will develop models that combine early, late, and hybrid fusion techniques to better integrate multimodal features. Moreover, it is important to capture temporal dynamics and changes in sentiment over time.

Data and code availability

Code underlying this paper are available at Github repo.

References

[1] Chandrasekaran, G., Nguyen, T.N. and Hemanth D, J. (2021) Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11**(5): e1415.

[2] Yavuz, A. and Eken, S. (2023) Gold returns prediction: Assessment based on major events. *EAI Endorsed Transactions on Scalable Information Systems* **10**(5).

[3] Balta Kaç, S. and Eken, S. (2023) Customer complaints-based water quality analysis. *Water* **15**(18): 3171.

[4] Yurtsever, M.M.E., Shiraz, M., Ekinci, E. and Eken, S. (2023) Comparing covid-19 vaccine passports attitudes across countries by analysing reddit comments. *Journal of Information Science* : 01655515221148356.

[5] Yurtsever, M.M.E., Ekinci, E. and Eken, S. (2023) Covid-19 and behavioral analytics: Deep learning-based work-from-home sensing from reddit comments. In *International Conference on Computing, Intelligence and Data Analytics* (Springer): 143–155.

[6] Köroğlu, F.E., Çakmak, S., Yurtsever, M.M.E. and Eken, S. (2024) Smart waste management: A case study on garbage container detection. In *6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence, MedPRAI 2024, İstanbul, Türkiye, October 18-19, 2024, Proceedings* (Springer).

[7] Medhat, W., Hassan, A. and Korashy, H. (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* **5**(4): 1093–1113.

[8] Alam, F., Ofli, F. and Imran, M. (2018) Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, **12**.

[9] Kaç, S.B., Eken, S., Balta, D.D., Balta, M., İskefiyeli, M. and Özçelik, İ. (2024) Image-based security techniques for water critical infrastructure surveillance. *Applied Soft Computing* **161**: 111730.

[10] Jayagopal, A., Aiswarya, A.M., Garg, A. and Nandakumar, S.K. (2022) Multimodal representation learning with text and images. *arXiv preprint arXiv:2205.00142* .

[11] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D. and Langlotz, C.P. (2022) Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference* (PMLR): 2–25.

[12] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G. *et al.* (2021) Learning transferable visual models from natural language supervision. In *International conference on machine learning* (PMLR): 8748–8763.

[13] Lin, Z., Bas, E., Singh, K.Y., Swaminathan, G. and Bhotika, R. (2023) Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*: 2227–2236.

[14] Cheema, G.S., Hakimov, S., Müller-Budack, E. and Ewerth, R. (2021) A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*: 37–45.

[15] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010) Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* **61**(12): 2544–2558.

[16] Saif, H., Bashevoy, M., Taylor, S., Fernandez, M. and Alani, H. (2016) Senticircles: A platform for contextual and conceptual sentiment analysis. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13* (Springer): 140–145.

[17] You, Q., Luo, J., Jin, H. and Yang, J. (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, **29**.

[18] You, Q., Luo, J., Jin, H. and Yang, J. (2016) Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, **30**.

[19] Jiang, T., Wang, J., Liu, Z. and Ling, Y. (2020) Fusion-extraction network for multimodal sentiment analysis. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24* (Springer): 785–797.

[20] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z. *et al.* (2015) Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**: 211–252.

[21] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 770–778.

[22] Pennington, J., Socher, R. and Manning, C.D. (2014) Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*: 1532–1543.

[23] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

[24] Bica, M., Palen, L. and Bopp, C. (2017) Visual representations of disaster. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*: 1262–1276.

[25] Sit, M.A., Koylu, C. and Demir, I. (2020) Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma. In *Social Sensing and Big Data Computing for*

Disaster Management (Routledge), 8–32.

[26] Hassan, S.Z., Ahmad, K., Hicks, S., Halvorsen, P., Al-Fuqaha, A., Conci, N. and Riegler, M. (2022) Visual sentiment analysis from disaster images in social media. *Sensors* **22**(10): 3628.

[27] Hassan, S.Z., Ahmad, K., Riegler, M.A., Hicks, S., Conci, N., Halvorsen, P. and Al-Fuqaha, A. (2021) Visual sentiment analysis: A natural disasteruse-case task at mediaeval 2021. *arXiv preprint arXiv:2111.11471* .

[28] Niu, T., Zhu, S., Pang, L. and El Saddik, A. (2016) Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22* (Springer): 15–27.

[29] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. and Torralba, A. (2017) Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **40**(6): 1452–1464.

[30] Siqueira, H., Magg, S. and Wermter, S. (2020) Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, **34**: 5800–5809.

[31] Yi, D., Lei, Z. and Li, S.Z. (2015) Shared representation learning for heterogenous face recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE), **1**: 1–7.

[32] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O. *et al.* (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

[33] Loria, S. *et al.* (2018) textblob documentation. *Release 0.15* **2**(8): 269.