





Article

Visual Sentiment Analysis from Disaster Images in Social Media

Syed Zohaib Hassan ¹, Kashif Ahmad ^{2,*}, Steven Hicks ¹, Pål Halvorsen ¹, Ala Al-Fuqaha ²
and Nicola Conci ³ and Michael Riegler ¹

¹ SimulaMet, 0167 Oslo, Norway; syed@simula.no (S.Z.H.); steven@simula.no (S.H.); paalh@simula.no (P.H.); michael@simula.no (M.R.)

² Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha 34110, Qatar; aalfuqaha@hbku.edu.qa

³ Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy; nicola.conci@unitn.it

* Correspondence: kahmad@hbku.edu.qa

Abstract: The increasing popularity of social networks and users' tendency towards sharing their feelings, expressions, and opinions in text, visual, and audio content have opened new opportunities and challenges in sentiment analysis. While sentiment analysis of text streams has been widely explored in the literature, sentiment analysis from images and videos is relatively new. This article focuses on visual sentiment analysis in a societally important domain, namely disaster analysis in social media. To this aim, we propose a deep visual sentiment analyzer for disaster-related images, covering different aspects of visual sentiment analysis starting from data collection, annotation, model selection, implementation, and evaluations. For data annotation and analyzing people's sentiments towards natural disasters and associated images in social media, a crowd-sourcing study has been conducted with a large number of participants worldwide. The crowd-sourcing study resulted in a large-scale benchmark dataset with four different sets of annotations, each aiming at a separate task. The presented analysis and the associated dataset, which is made public, will provide a baseline/benchmark for future research in the domain. We believe the proposed system can contribute toward more livable communities by helping different stakeholders, such as news broadcasters, humanitarian organizations, as well as the general public.

Keywords: sentiment analysis; emotions; deep learning; multimedia retrieval; natural disasters



Citation: Hassan, S.Z.; Ahmad, K.; Hicks, S.; Halvorsen, P.; Al-Fuqaha, A.; Conci, N.; Riegler, M. Visual Sentiment Analysis from Disaster Images in Social Media. *Sensors* **2022**, *22*, 3628. <https://doi.org/10.3390/s22103628>

Academic Editor: Leon Rothkrantz

Received: 23 November 2021

Accepted: 16 December 2021

Published: 10 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sentiment analysis aims to analyze and extract opinions, views, and perceptions about an entity (e.g., product, service, or an action). It has been widely adopted by businesses helping them to understand consumers' perceptions about their products and services. The recent development and popularity of social media helped researchers to extend the scope of sentiment analysis to other interesting applications. A recent example is reported by Ozturk et al. [1], where computational sentiment analysis is applied to the leading media sources, as well as social media, to extract sentiments on the Syrian refugee crisis. Another example is reported by Kuvsen et al. [2], where the neutrality of tweets and other reports from the winner of the Austrian presidential election were analyzed and compared to the opponents' content on social media.

The concept of sentiment analysis has been widely utilized in Natural Language Processing (NLP), where several techniques have been employed to extract sentiments from text streams in terms of positive, negative, and neutral perception/opinion. With the recent advancement in NLP, an in-depth analysis of text streams from different sources is possible in different application domains, such as education, entertainment, hosteling, and other businesses [3]. More recently, several efforts have been made to analyze visual

content to derive sentiments. The vast majority of the literature on visual sentiment analysis focuses on close-up facial images where facial expressions are used as visual cues to derive sentiments and predict emotions [4]. Attempts have also been made to extend the visual approach to more complex images, including, for example, multiple objects and background details. The recent developments in machine learning and, in particular, deep learning has contributed to significantly boost the results in this research area [5]. However, extracting sentiments from visual content is not straightforward, and several factors need to be considered.

In this article, we analyze the problem of visual sentiment analysis from different perspectives with a particular focus on the challenges, opportunities, and potential applications of visual sentiment analysis of challenging disaster-related images from social media. Disaster analysis in social media content has received great attention in the community in recent years [6–8]. We believe visual sentiment analysis of disaster-related images is an exciting research domain that will benefit users and the community in a diversified set of applications. To this aim, we propose a deep sentiment analyzer, and discuss the processing pipeline of visual sentiment analysis starting from data collection and annotation via a crowd-sourcing study, and conclude with the development and training of deep learning models. The work is motivated by our initial efforts in the domain [9], where an initial crowd-sourcing study was conducted with a few volunteers to test the viability of the approach.

To the best of our knowledge, this is the first attempt to develop a large-scale benchmark for sentiment analysis of disaster-related visual content. Disaster-related images are complex and generally involve several objects, as well as significant details in their backgrounds. We believe such a challenging use case is quintessential, being an opportunity to discuss the processing pipeline of visual sentiment analysis and provide a baseline for future research in the domain. Moreover, the visual sentiment analysis of disasters has several applications and can contribute toward more livable communities. It can also help news agencies to cover such adverse events from different angles and perspectives. Similarly, humanitarian organizations can benefit from such a framework to spread the information on a wider scale, focusing on the visual content that best demonstrates the evidence of a certain event. In order to facilitate future work in the domain, a large-scale dataset is collected, annotated, and made publicly available (<https://datasets.simula.no/image-sentiment/>, accessed on 10 December 2021). For the annotation of the dataset, a crowd-sourcing activity with a large number of participants has been conducted.

The main contributions of the work can be summarized as follows:

- We extend the concept of visual sentiment analysis to a more challenging and crucial task of disaster analysis, generally involving multiple objects and other relevant information in the background of images, and propose a deep architecture-based visual sentiment analyzer for an automatic sentiment analysis of natural disaster-related images from social media.
- Assuming that the available deep architectures respond differently to an image by extracting diverse but complementary image features, we evaluate the performance of several deep architectures pre-trained on ImageNet and Places dataset both individually and in combination.
- We conduct a crowd-sourcing study to analyze people's sentiments towards disasters and disaster-related content and annotate the training data. In the study, a total of 2338 users participated in analyzing and annotating 4003 disaster-related images (All images are Creative Commons licensed).
- We provide a benchmark visual sentiment analysis dataset with four different sets of annotations, each aimed at solving a separate task, which is expected to be proven as a useful resource for future work in the domain. To the best of our knowledge, this is the first attempt on the subject.

The rest of the paper is organized as follows: Section 2 motivates the work by differentiating it from the related concepts, such as emotion and facial recognition, as well

as textual sentiment analysis, and emphasizing on the opportunities, challenges, and potential applications. Section 4 describes the proposed pipeline for visual sentiment analysis of natural disaster-related images. Section 5 provides the statistics of the crowdsourcing study along with the experimental results of the proposed deep sentiment analyzer. Section 6 concludes this study and provides directions for future research.

2. Motivation, Concepts, Challenges and Applications

As implied by the popular proverb “a picture is worth a thousand words,” visual content is an effective means to convey not only facts but also cues about sentiments and emotions. Such cues representing the emotions and sentiments of the photographers may trigger similar feelings from the observer and could be of help in understanding visual content beyond semantic concepts in different application domains, such as education, entertainment, advertisement, and journalism. To this aim, masters of photography have always utilized smart choices, especially in terms of scenes, perspective, angle of shooting, and color filtering, to let the underlying information smoothly flow to the general public. Similarly, every user aiming to increase in popularity on the Internet will utilize the same tricks [10]. However, it is not fully clear how such emotional cues can be evoked by visual content, and more importantly, how the sentiments derived from a scene by an automatic algorithm can be expressed. This opens an interesting line of research to interpret emotions and sentiments perceived by users viewing visual content.

In the literature, emotions, opinion mining, feelings, and sentiment analysis have been used interchangeably [4,11]. In practice, there is a significant difference among those terms. Sentiments are influenced by emotions, and they allow individuals to show their emotions through expressions. In short, sentiments can be defined as a combination of emotions and cognition. Therefore, sentiments reveal underlying emotions through ways that require cognition (e.g., speech, actions, or written content).

The categorical representation of those concepts (i.e., emotions, sentiments) can be different, although the visual cues representing them are closely related. For instance, emotion recognition, opinion mining, and sentiment analysis are generally expressed by three main classes: *happy*, *sad*, and *neutral* or, similarly, *positive*, *negative*, and *neutral* [12]. However, similar types of visual features are used to infer those states [13]. For instance, facial expressions have been widely explored for both emotion recognition and visual sentiment analysis in close-up images [4,14]; though it would be simplistic to limit the capability of recognizing emotions and sentiments in close-up facial images. There are several application domains where more complex images need to be analyzed. This is exactly the case of the aforementioned scenario of disaster-related images, in which the background information is often crucial to evoke someone’s emotions and sentiments. Figure 1 provides samples of disaster-related images, highlighting the diversity in terms of content that needs to be examined. In addition, it is also important to mention that emotions and feelings can be different from subject to subject and based on experience.

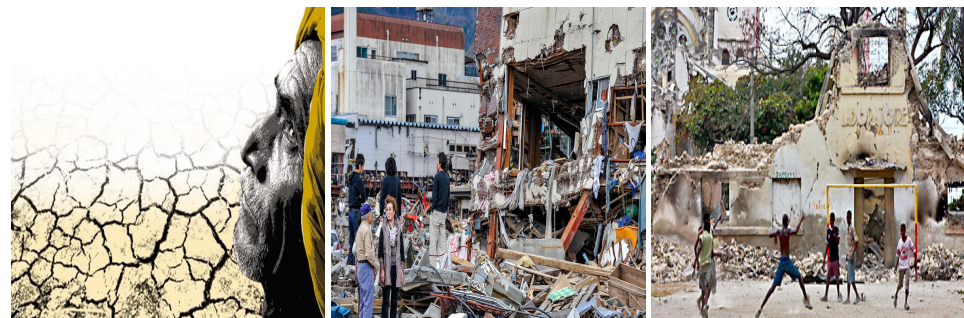


Figure 1. Sample images of natural disasters for sentiment analysis showing the diversity in the content and information to be extracted.

In contrast to textual sentiment analysis, visual sentiment analysis is a nascent area of research, and several aspects still need to be investigated. The following are some of the key open research challenges in visual sentiment analysis in general and disaster-related content in particular that need to be addressed:

- Defining/identifying sentiments—The biggest challenge in this domain is defining sentiments and identifying the one that better suits the given visual content. Sentiments are very subjective and vary from person to person. Moreover, the intensity of the sentiments conveyed by an image is another item to be tackled.
- Semantic gap—One of the open questions that researchers have thoroughly investigated in the past decades is the semantic gap between visual features and cognition [13]. The selection of visual features is very crucial in multimedia analysis in general and in sentiment analysis in particular. We believe object and scene-level features could help in extracting such visual cues.
- Data collection and annotation—Image sources, sentiment labels, and feature selection are application-dependent. For example, an entertainment or education context is completely different from the humanitarian one. Such diversity makes it difficult to collect benchmark datasets from which knowledge can be transferred, thus requiring ad-hoc data crawling and annotation.

3. Related Work

Natural language processing has made great strides in accurately determining the sentiment of a given spoken text, with reference to users' reviews on movies and products [15–17]. When looking at the inference of sentiments from visual data, the literature is rather limited [18]. However, being a new and challenging task, the lack of openly available datasets makes it difficult to create a common benchmark on which a solid state-of-the-art can be built. Machajdik et al. [19] performed a study on using extracted features based on psychology and art theory to classify the emotional response of a given image. The features were grouped by color, texture, composition, and content and then classified by a naive Bayes-based classifier. Although the work achieved good results for the time, the extracted features have a hard time capturing the complex relationship between human emotion and the content of an image. Thus, more recent works have relied on reaching some middle-ground by extracting adjective-noun pairs (ANPs), such as *funny dog* or *sad monkey*, which then may be used to infer the sentiment of the image. Borth et al. [20] released a dataset consisting of over 3000 ANPs, aimed to help researchers contribute to the field. Their work also includes a set of baseline models and is commonly used to benchmark methods based on ANPs [21]. Another widely used approach that bypasses the need for large sentiment datasets consists of using deep neural networks and transfer learning from models trained on large-scale classification datasets, such as ImageNet [22]. For instance, Chandrasekaran et al. [23] fine-tuned an existing pre-trained model, namely VggNet on a Twitter dataset, to extract and classify the evoked emotions. There are also some efforts where the domain datasets are utilized for training visual sentiment analysis models [24]. Al-Halah et al. [25] developed a method for predicting emoticons (emojis) based on a given image. The emojis act as a proxy for the emotional response of an image. They collected a dataset containing over 4 million images and emoticon pairs from Twitter, which was used to train a novel CNN architecture named SimleyNet [25]. Some works also employed the text associated with images for visual sentiment analysis. For instance, in [26], an attention-based network, namely Attention-based Modality-Gated Networks (AMGN), has been proposed to exploit the correlation between visual and textual information for sentiment analysis. There are also some recent efforts for learning subjective attributes of images, such as emotions, from auxiliary sources of information (i.e., users' interactions on social media) [27]. More recently, several attempts have been made for multi-level and multi-scale representation to extract visual cues of sentiment analysis. For instance, You et al. [28] analyzed the importance of local image regions in visual sentiment analysis through an attention model. Similarly, Ou et al. [29] proposed a multi-level context pyra-

mid network (MCPNet) architecture to combine local and global features in a cross-layer feature fusion scheme. The basic motivation behind the multi-level representation is to utilize both the local features and global context as the size and position of relevant visual cues (i.e., objects and background information) in images is diverse. Aside from diversity in size and position, visual cues of sentiments and emotions could also comprise multiple objects. Thus, the relation/interaction among different objects in an image is also important to be considered in visual sentiment analysis. In order to incorporate the interactive characteristics of objects, Wu et al. [30] proposed a Graph Convolutional Network (GCN)-based solution to extract interaction features. To this aim, firstly, a CNN model is employed to detect objects in an image. The detected objects are then connected via a graph where visual features represent nodes while the emotional distances between objects correspond to the edges of the graph. Table 1 summarizes the existing works on visual sentiment analysis. As can be seen in the table, the majority of existing works on visual sentiment analysis focus on general close-up facial images or images of natural scenes and objects, such as flowers or dogs, where facial expressions, objects, the scene, and color features are used as visual cues to derive sentiments and predict emotions. This paper demonstrates that the concept of visual sentiment analysis can be extended to more complex images of disasters, which generally contain multiple objects and background details.

Table 1. A summary and comparative analysis of existing works on the topic.

Refs.	Dataset	Application	Features/Model	Main Focus
[19]	CMUMOSEI [19]	Generic	Color and texture features, such as Tamura	Relies on features based on psychology and art theory to classify the emotional response of a given image. The features are grouped by color, texture, composition, and content, and then classified by a naive Bayes-based classifier.
[25]	SimleyNet [25]	Emojis	CNNs	Mainly focuses on detection and classification of emojis, which act as a proxy for the emotional response of an image. Moreover, also proposes a dataset containing over 4 million images and emoticon pairs from Twitter.
[31]	Twitter dataset [32], Flickr and Instagram dataset [33], CMUMOSEI [19]	Generic	CNNs	Proposes a residual attention-based deep learning network (RA-DLNet) aiming to learn the spatial hierarchies of image features extracted through CNNs. Moreover, analyses of the performance of seven state-of-the-art CNN architectures on several datasets.
[34]	Flickr dataset [35]	Generic	CNNs, handcrafted features (GIST, BoW)	Jointly utilize text (objective text description of images obtained/extracted from the visual content of images through CNNs model) and visual features in an embedding space obtained with Canonical Correlation Analysis (CCA).
[26]	Self-collected dataset	Generic	CNNs, LSTM	Proposes an attention-based network, namely Attention-based Modality-Gated Networks (AMGN) to exploit the correlation between visual and textual information for sentiment analysis.
[28]	VSO [20]	Generic	CNNs	Aims to explore the role of local image regions in visual sentiment analysis through an attention mechanism-based model.
[29]	Twitter [36], Flickr and Instagram dataset [33], Emotion ROI [37]	Generic	CNNs	Proposes a multi-level context pyramid network (MCPNet) aiming to combine local and global features in cross-layer feature fusion scheme for better representation of visual cues.
[30]	Emotion ROI [37], Twitter [36]	Generic	CNNs and GCN	Proposes a Graph Convolutional Network (GCN)-based framework to incorporate the interaction features among different objects in an image. The visual features of the objects represent nodes, while the emotional distances between objects correspond to the edges of the graph.
This Work	Self-collected	Natural Disasters	CNNs	Explores a new application of visual sentiment analysis by collecting and sharing a benchmark dataset with baseline experimental results. Moreover, it also highlights the key research challenges, potential applications, and stack-holders.

4. Proposed Visual Sentiment Analysis Processing Pipeline

Figure 2 provides the block diagram of the proposed architecture for visual sentiment analysis. The pipeline is composed of five phases. The process starts with crawling social media platforms for disaster-related images, followed by sentiment tags/categories selection to be associated with the disaster-related images in the crowd-sourcing study. Before conducting the crowd-sourcing study, we manually analyzed the images and removed irrelevant images. In the crowd-sourcing study, a subset of the downloaded images, after removing the irrelevant images, are annotated by human participants. A CNN and a transfer learning-based method are then used for multi-label classification and to automatically assign sentiments/tags to the images. In the next subsections, we provide a detailed description of each component.

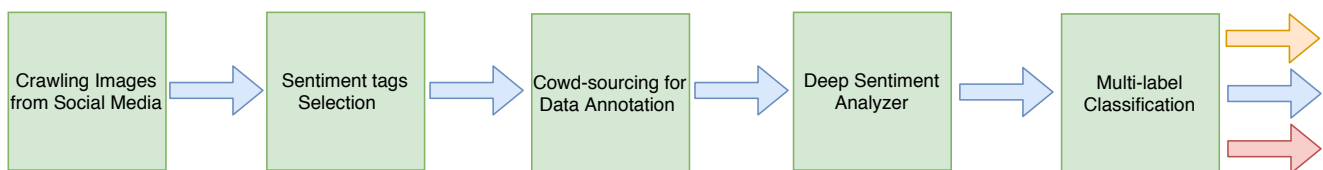


Figure 2. Block diagram of the proposed visual sentiment analysis processing pipeline.

4.1. Data Collection and Sentiment Category Selection

At the beginning of the processing pipeline, social media platforms, such as Twitter and Flickr, and Google API are crawled to collect images to analyze. All the downloaded images have been selected, paying attention to the licensing policies in terms of free usage and sharing. The images have been selected according to a set of keywords, such as *floods*, *hurricanes*, *wildfires*, *droughts*, *landslides*, and *earthquakes*, and enriched, with additional relevant information, as, for example, location (*cyclones in Fiji* or *floods in Pakistan*), and accessing the list of recent natural disasters made available from EM-DAT (<https://www.emdat.be/>, accessed on 10 December 2021). EM-DAT is a platform maintained by the United Nations providing statistics on worldwide disasters.

The selection of labels for the crowd-sourcing study was one of the challenging and perhaps most crucial phases of the work, as discussed above. In the literature, sentiments are generally represented as *Positive*, *Negative*, and *Neutral* [13]. However, considering the nature and potential applications of the proposed deep sentiment analysis processing pipeline, we aim to target sentiments that are more specific to disaster-related content. For instance, terms such as *sadness*, *fear*, and *destruction* are more commonly used with disaster-related content. Moreover, we also considered the potential stakeholders and users of the proposed system in the tag selection. In order to choose more relevant and representative labels for our deep sentiment analyzer, we choose four different sets of tags, including the most commonly used ones and two sets obtained from recent work in Psychology [38], reporting 27 different types of emotions. In total, we annotated every image with four different sets of labels. The first set is composed of three tags, namely *positive*, *negative*, and *neutral*. The second set also contains three tags, namely *relax/calm*, *normal*, and *stimulated/excited*. The third set is composed of seven tags, namely *joy*, *sadness*, *fear*, *disgust*, *anger*, *surprise*, and *neutral*. The last set of tags is composed of ten tags, namely *anger*, *anxiety*, *craving*, *empathetic pain*, *fear*, *horror*, *joy*, *relief*, *sadness*, and *surprise*. Table 2 lists the tags used in each question of the crowd-sourcing. The third set is closely related to Ekman's basic emotions model, which represents human emotion in six categories, including *anger*, *surprise*, *disgust*, *enjoyment*, *fear*, and *sadness*. The final set provides a deeper categorization of the sentiments/emotions, which further increases the complexity of the task. The basic motivation for the four different sets of labels is to cover different aspects of the task and analyze how the complexity of the task varies by going deeper into the sentiments hierarchy.

Table 2. List of tags used in the crowd-sourcing study in the four sets.

Sets	Tags
Set 1	Positive, Negative, Neutral
Set 2	Relax, Stimulated, Normal
Set 3	Joy, Sadness, Fear, Disgust, Anger, Surprise, and Neutral
Set 4	Anger, Anxiety, craving, Empathetic pain, Fear, Horror, Joy, Relief, Sadness, and Surprise

4.2. The Crowd-Sourcing Study

The crowd-sourcing study aims to develop ground truth for the proposed deep sentiment analyzer by collecting human perceptions and sentiments about disasters and the associated visual content. The crowd-sourcing study was conducted using Microworkers (<https://www.microworkers.com>, accessed on 10 December 2021), where the selected images were presented to the participants to be annotated. In total, 4003 images were analyzed during the study. In order to assure the quality of the annotations, at least five different participants were assigned to analyze an image. The final tag/tags were chosen based on the majority votes from the five participants assigned to it. In total, 10,010 different responses were obtained during the study from 2338 different participants. Figure 3 provides the statistics of participants' demographics in terms of percentages of participants belonging to different regions of the world and the percentages of responses obtained from different regions. As can be seen, the majority of the participants and responses are from South Asian countries. However, there is participation from other parts of the world too. One of the potential reasons for the high participation of South Asian countries is the more use of such platforms in the region. The participants included individuals from different age groups and 98 different countries. We also noted the time spent by a participant on an image, which helped in filtering out the careless or inappropriate responses from the participants. The average response time recorded per image during the study is 139 s. Before the final study was conducted, two trial studies were performed to fine-tune the test, correct errors, and improve clarity and readability. The HTML version of the crowd-sourcing study template has been made available as a part of the dataset.

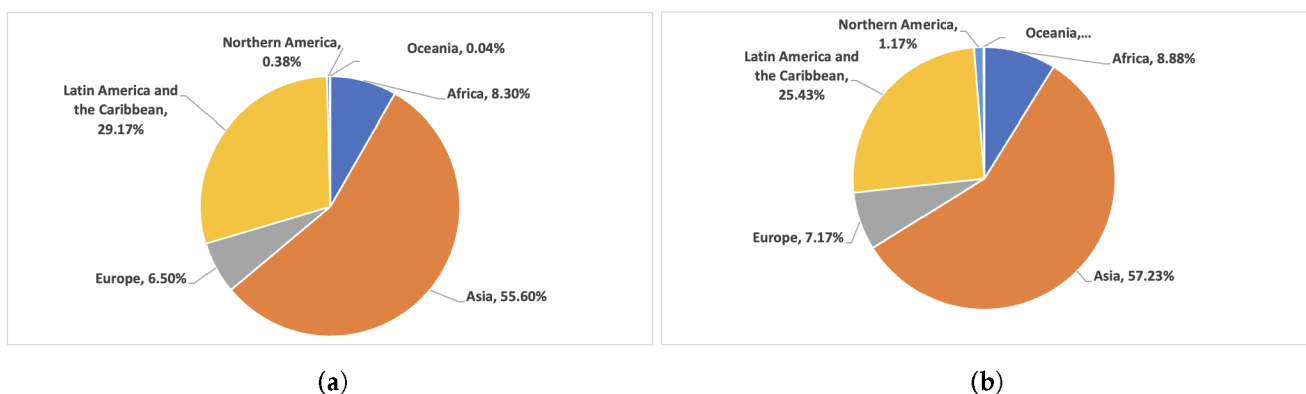


Figure 3. Statistics of the crowd-sourcing study participants' demographics. (a) Statistics of the participants' demographics in terms of the percentage of participants from different regions of the world. (b) Statistics of the participant demographics in terms of percentages of responses received from participants belonging to different regions of the world.

Figure 4 provides a block diagram of the layout of the crowd-sourcing study platform. The participants were provided with a disaster-related image followed by five different questions. In the first four questions, the participants are asked to annotate the image with different sets of labels (Table 2), each aiming to prepare training data for a separate task. In the first question, we asked the participants to rate their evoked emotions from 1 to

10 (1 = very negative, 5 = neutral, and 10 = very positive), after seeing the image. This question aims to analyze the degree of sentiments conveyed by an image. The second question is similar to the first one, except the labels focus on feelings in terms of *calm/relaxed*, *normal*, and *excited*. In the third and fourth questions, the participants are asked to assign one or more labels from a list of seven and ten tags, respectively. In these two questions, the participants were also encouraged to provide their own tags if they felt the provided lists were not representative enough. The fifth question aims to highlight the image features, at the scene or object level, which influence human emotions.

The Data Annotation Task



Given Image

Questions

- Q.1: Your evoked emotion after seeing this picture is: (Likert 1-9: 1-very negative, 5-neutral, 10-very positive)
- Q.2: Confronted with the picture, you are feeling: (Likert 1-9: 1-calm/relaxed, 10-excited/stimulated)
- Q.3: Which one(s) of the following major emotion keywords best describe your evoked emotion after seeing this picture? (Choose at least one or more keywords that are suitable: 1-joy, 2-sadness, 3-fear, 4-disgust, 5-anger, 6-surprise, 7-neutral)
- Q.4: Select the specific emotion keywords that can describe your evoked emotion after seeing this picture? (Choose all the keywords that are relevant to your emotion)
- | | |
|---|--|
| 1-Anger(anger, angry, disgust, boiling with anger) | 2-Anxiety (anxiety, fear, nervousness) |
| 3-Craving (hunger, desire, a situation of hunger) | 4-Emphatic pain (pain, empathic pain, shock) |
| 5-Fear (fear, feeling scared, extreme fear) | 6-Horror (shock, horror, feeling scared) |
| 7-Joy (happiness, extreme happiness, love) | 8-Relief (relief, deep relief, sense of narrow escape) |
| 9-Sadness (sadness, extreme sadness, sympathy) | 10-Surprise (surprise, shock, amazement) |
| 11-Others/comments | |
- Q.5: What kind of information of the image influences your evoked emotion the most?
- | | |
|--|--|
| 1-Human facial expression, post or gesture | 2-Image color, contrast, saturation, etc. |
| 3-Image background (scene, landmark, etc.) | 4-Objects in image (gadgets, clothes, animals, etc.) |
| 5-Texts in image | 6-Emoji sticker |
| 7-Halo effect | 8-Others/comments |

Figure 4. An illustration of the web application used for the crowd-sourcing study. A disaster-related image is provided to the users who are asked to provide options/tags. In case, additional tags/comments can also be reported.

The resulting dataset is named image-sentiment dataset and can be downloaded via <https://datasets.simula.no/image-sentiment/>, accessed on 10 December 2021. Details about the dataset can be found in Tables 3–5.

Table 3. Statistics of the dataset for task 1.

Tags	# Samples
Positive	803
Negative	2297
Neutral	579

Table 4. Statistics of the dataset for task 2.

Tags	# Samples	Tags	# Samples
Joy	1207	Sadness	3336
Fear	2797	Disgust	1428
Anger	1419	Surprise	2233
Neutral	1892	-	-

Table 5. Statistics of the dataset for task 3.

Tags	# Samples	Tags	# Samples
Anger	2108	Anxiety	2716
Craving	1100	Pain	2544
Fear	2803	Horror	2042
Joy	1181	Relief	1356
Sadness	3300	Surprise	1975

4.3. Deep Visual Sentiment Analyzer

Our proposed multi-label deep visual sentiment analyzer is mainly based on a convolutional neural network (CNN) and transfer learning. Based on the participants' responses in the fifth question, where they were asked to highlight the image features/information that influences their emotions and sentiments, we believe both object and scene-level features could be useful for the classification task. Thus, we opted for both object and scene-level features extracted through existing deep models pre-trained on the ImageNet [22] and Places [39] datasets, respectively. The model pre-trained on ImageNet extracts object-level information, while the one pre-trained on the Places dataset covers the background details [40]. In this work, we employed several state-of-the-art deep models, namely AlexNet [41], VGGNet [42], ResNet [43], Inception v3 [44], DenseNet [45], and EfficientNet [46]. These models are fine-tuned on the newly collected dataset for visual sentiment analysis of disaster-related images. The object and scene-level features are also combined using early fusion by including a concatenation layer in our framework, where features extracted from the final fully connected layers from models pre-trained on the ImageNet and Places datasets are combined before the classification layer. In the current implementation, we rely on a simple fusion technique aiming to identify and analyze the potential improvement by combining both object and background details. We note that in the current implementation, our deep sentiment analyzer relies on object and background information only, which accumulates a high percentage of the visual cues influencing participants' decisions in the crowd-sourcing study, as depicted in Figure 5. For future work, we plan to extend our framework to incorporate color and other information highlighted by the participants in the fifth question of the crowd-sourcing study (see Figure 5). In addition, in order to deal with class imbalance, as will be detailed in Section 5.2, we also used an oversampling technique to adjust the class distribution of the dataset. For the single-label classification (i.e., first task), we used an open-source library, namely *imblearn* [47], while for the multi-label problem (i.e., second and third tasks), we developed our own function. In fact, in the multi-label tasks, the classes are not independent, thus the naïve approach

(i.e., *imblearn*) could not be applied. In order to deal with it, we divided the classes into two groups based on positive and negative correlation with the majority class occurrence. Then, each group was sorted in descending order based on the number of samples in each class. We then iterate over each group and oversample the minority classes.

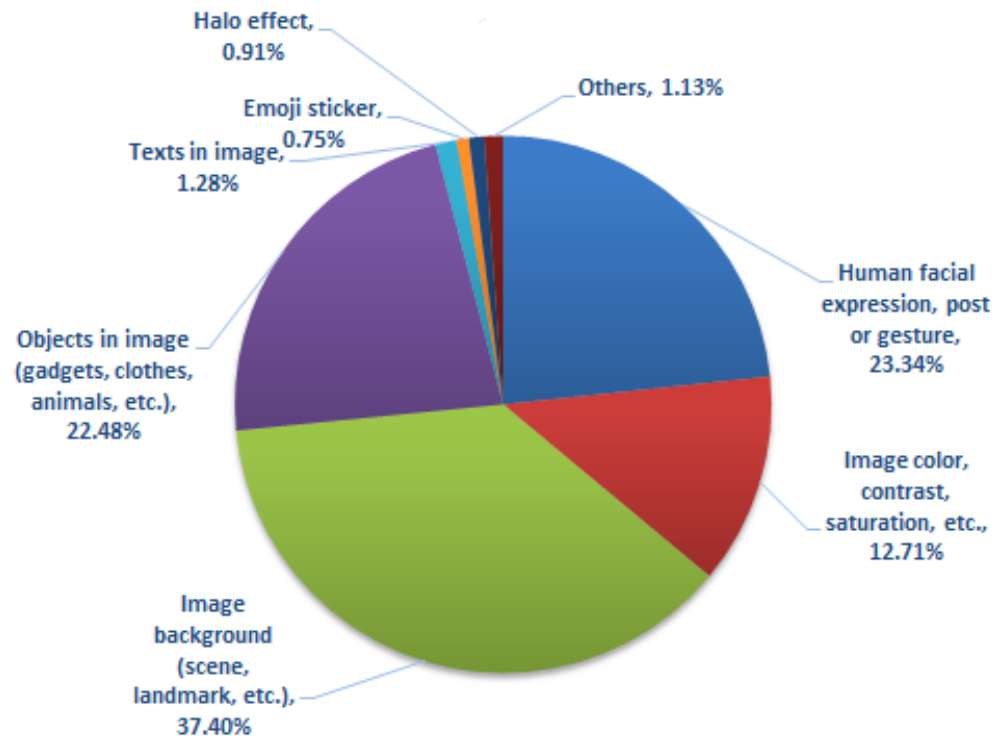


Figure 5. Statistics of the fifth question of the crowd-sourcing study in terms of what kind of information in the images influence users' emotion most.

Furthermore, for the multi-label analysis, we made several changes in the framework to adapt the pre-trained models for the task at hand. As a first step, a vector of the ground truth having all the possible labels has been created with the corresponding changes in the models. For instance, the top layer of the model has been modified to support multi-label classification by replacing the soft-max function with a sigmoid function. The sigmoid function turns out to be helpful, as it presents the results for each label in probabilistic terms, while the soft-max function holds the probability law and squashes all the values of a vector into a $[0, 1]$ range. Similar changes (i.e., replacing softmax with sigmoid function) are made in the formulation of the cross-entropy to properly fine-tune the pre-trained models.

5. Experiments and Evaluations

In this section, we provide a detailed analysis of the outcomes of the crowd-sourcing study and achieved experimental results.

5.1. Statistics of the Crowd-Sourcing Study and Dataset

Figure 6 provides the statistics of the first four questions of the crowd-sourcing study. Figure 6a (where tags 1 to 4 correspond to *negative* sentiments, 5 to *neutral*, and tags 6 to 9 represent *positive* sentiments) shows that the majority of the images analyzed in the crowd-sourcing study evoked *negative* sentiments. Looking at the remaining responses, we noticed that images labeled as *positive* are mostly captured during the rescue and rehabilitation process. Figure 6b provides the statistics of the second question, which is based on a different set of labels: *calm/relaxed*, *normal*, and *stimulated/excited*. The emotions here are quite evenly distributed across the entire spectrum, ranging from negative to

positive. Figure 6c provides the statistics of the third question of the study in terms of how frequently different tags are assigned with the images by the participants. As expected, *sadness* and *fear* are the most frequently used tags. The statistics of Question 4, which further extends the tags' list by going deeper in the hierarchy, are shown in Figure 6d. Similar to Question 3, higher percentages have been observed for *sadness* and *fear*.

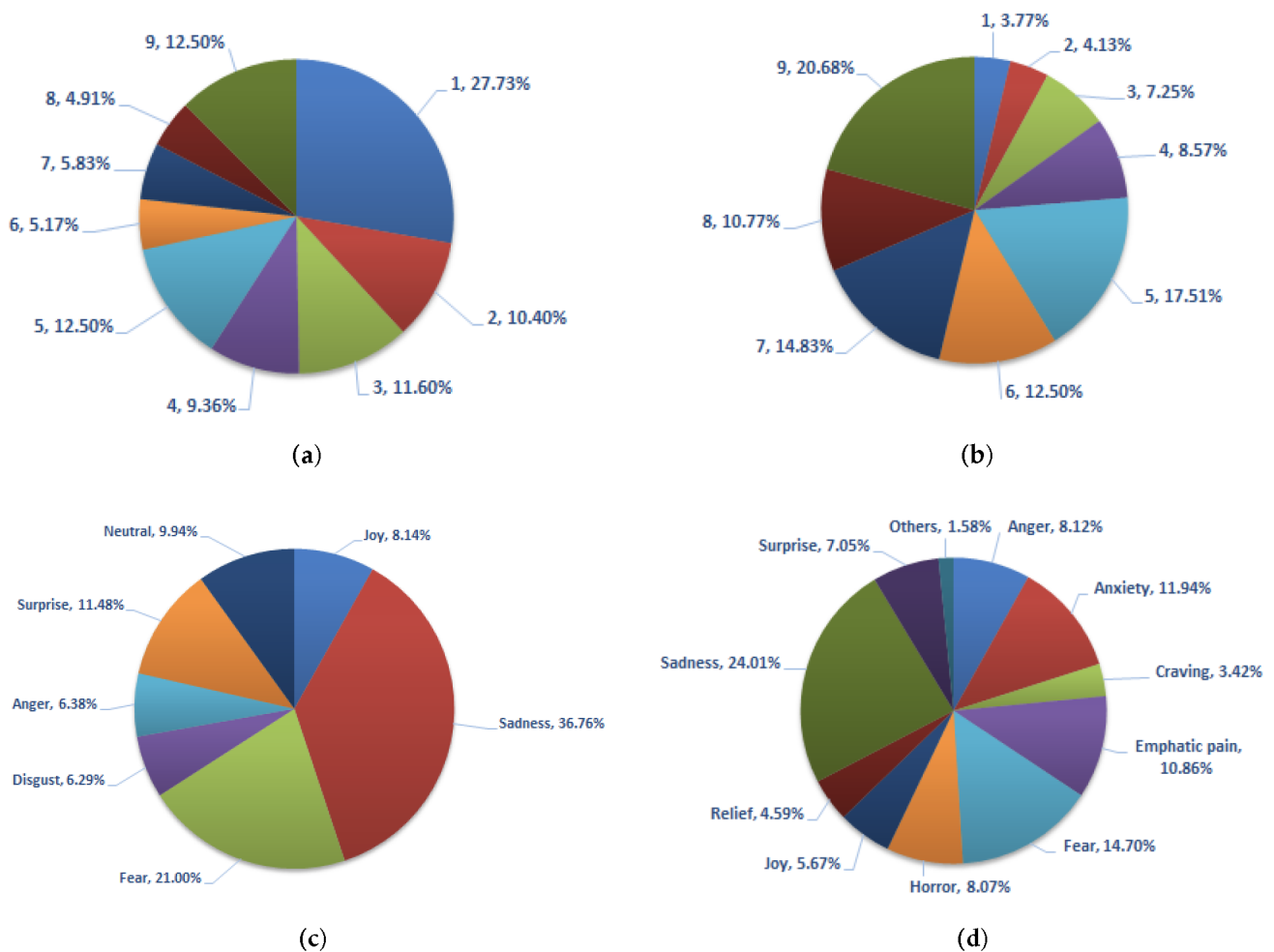
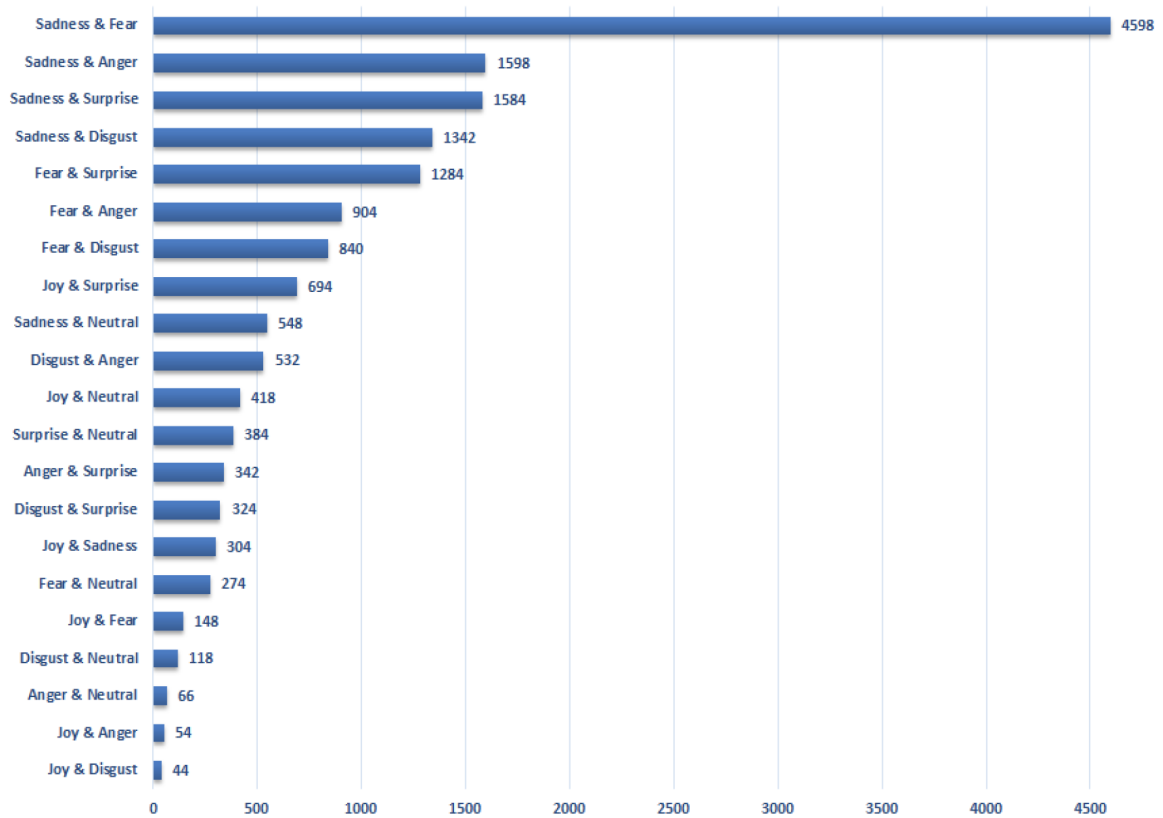
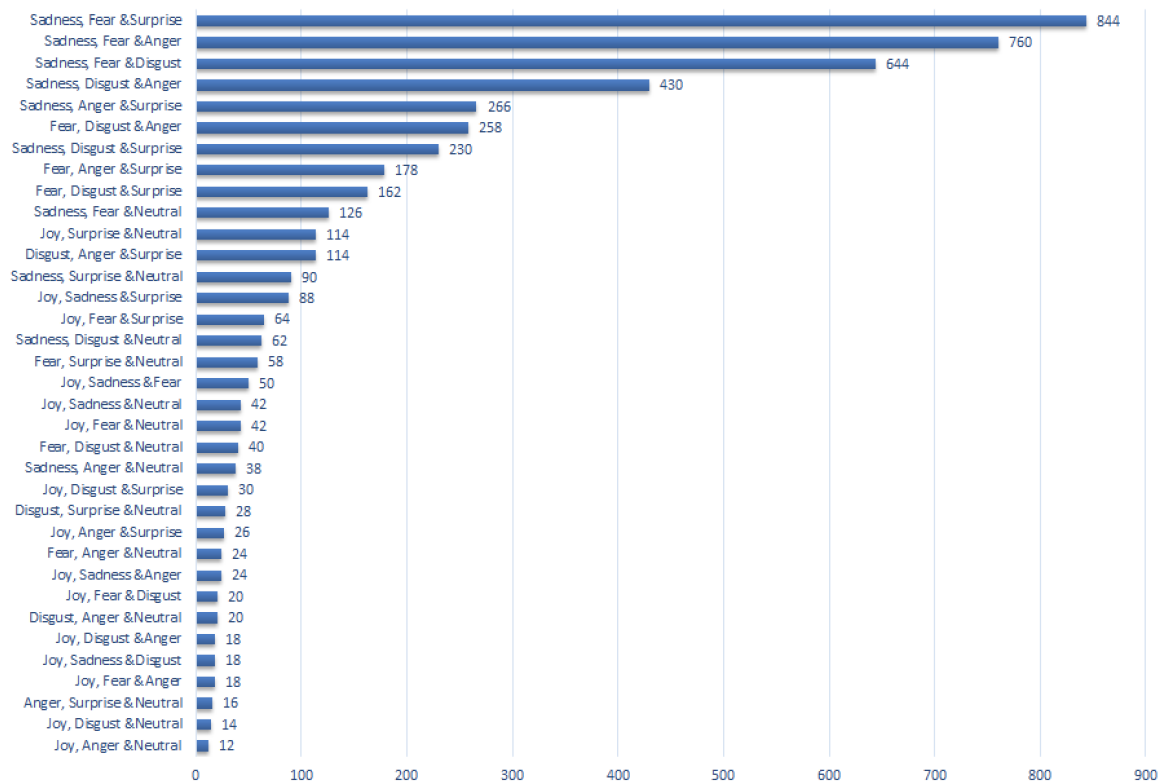


Figure 6. Statistics of the first four questions of the crowd-sourcing study. (a) Statistics of the responses for the first question. Tags 1 to 4 represent negative sentiments while tag 5 represents neutral, and tags 6 to 9 show positive sentiments. (b) Statistics of the responses for the second question. Tags 1 to 4 represent calm/relaxed emotion, tag 5 shows a normal condition, while tags 6 to 9 depict excited/stimulated status. (c) Statistics of the responses for the third question. (d) Statistics of the responses for the fourth question.

Another important aspect of the crowd-sourcing study is the analysis related to how frequently different tags are used jointly. In Figure 7a,b, we show this association in pairs, and groups of three, respectively. As can be seen, *sadness* is most frequently used with *fear*, *anger*, and *disgust*. Similarly, *fear* is also frequently used with *disgust*, *anger*, and *surprise*. As for the positive tags, *joy*, *surprise*, and *neutral* are jointly used. A similar trend has been observed in the group of three tags.



(a)



(b)

Figure 7. Statistics of the crowd-sourcing study in terms of how different tags are jointly associated with the images. (a) Tags jointly used in pairs. (b) Tags jointly used in a group of three.

Figure 5 provides the statistics of the final question of the study, where we asked the participants to highlight the image features/information that influences their emotions and tag selection for a given image. This question is expected to provide useful information from a methodological point of view. As can be seen, the image background (i.e., scene, landmarks) has been proven the most influential piece of information for evoking people's emotions (37.40% of the responses). Human expressions, gestures, and poses also seem very crucial (23%). Other factors, such as object-level information in images, and image color and contrast, contributed 22.48% and 12.71%, respectively.

5.2. Datasets

In this section, we provide the details of the datasets we have collected and adopted for the crowd-sourcing study. Tables 3–5 report the statistics of the dataset, in terms of a total number of samples per class used for each of the three tasks. For the first task, images are arranged in three different classes, namely *positive*, *negative*, and *neutral*, with a bias towards the negative samples, due to the topic taken into consideration. For this task, the dataset is single-label. In tasks 2 and 3, we use instead a multi-label annotation, and the dataset contains images from seven and ten different classes, respectively. As can be seen in Table 4, the majority of the classes have a higher number of images, and some of the classes have a similar range. For instance, *anger*, *joy*, and *disgust* have samples in the same range, while *neutral* and *surprise* and *sadness* and *fear* have almost the same amount of samples. One of the reasons for the pattern is the joint association of the tags with images by the participants of the crowd-sourcing study. A similar pattern can be observed in Table 5 for task 3, where sentiment classes, such as *craving*, *joy*, and *relief*, have a number of samples in the same range. Similarly, *anger*, *horror*, and *surprise* have the same range of a number of samples. On the other hand, the number of samples in *fear*, *sadness*, and *anxiety* are in the same range.

5.3. Experimental Setup

For all experiments, we used 70% of the data for training, 10% for validation, and 20% for testing. The experiments were carried out on Intel(R) machine Core(TM) i7-8700 with GPU GeForce RTX 2070 (8 GB) and 62 GB of RAM. For the evaluation of the methods, we are using accuracy, precision, recall, and F1-Score. We note that the overall results are measured in terms of weighted average, i.e., weighted precision, recall, and F1-Score.

5.4. Experimental Results

Tables 6–8 provide experimental results of the proposed deep sentiment analyzer on the three tasks. Since one of the main motivations behind the experiments is to provide a baseline for future work in the domain, we evaluate the proposed single and multi-label frameworks with several existing deep models pre-trained on ImageNet and Places datasets.

For the first task, we evaluate the performance of the proposed single-label framework with several state-of-the-art models in differentiating in *positive*, *negative*, and *neutral* sentiments. As shown in Table 6, we obtain encouraging results in terms of accuracy, recall, precision, and F1-score. Surprisingly, better results have been observed for the smaller architecture (VGGNet) compared to the most recent models, such as EfficientNet and DenseNet. As far as the contribution of the object and scene-level features is concerned, both types of features could turn out to be useful for the classification task. We also combined the object and scene-level feature following an early fusion approach; no significant improvement has been observed.

Table 7 provides the experimental results of the proposed multi-label framework, where the system needs to automatically associate to an image one or more labels from seven tags, namely *sadness*, *fear*, *disgust*, *joy*, *anger*, *surprise*, and *neutral*. Furthermore, in this case, the results are encouraging, especially considering the complexity of the tags in

terms of inter and intra-class variation. In this case, the fusion of object and scene-level features outperforms the individual models in terms of accuracy, recall, and F1-score.

Table 8 provides the results of the third task, where we go deeper in the sentiments hierarchy with a total of ten tags. Moreover, similar to previous tasks, the results are also encouraging on the more complex task where the sentiments' categories are increased further.

Table 6. Evaluation of the proposed visual sentiment analyzer on task 1 (i.e., single-label classification of three classes, namely negative, neutral, and positive).

Model	Accuracy	Precision	Recall	F-Score
VGGNet (ImageNet)	92.12	88.64	87.63	87.89
VGGNet (Places)	92.88	89.92	88.43	89.07
Inception-v3 (ImageNet)	82.59	76.38	68.81	71.60
ResNet-50 (ImageNet)	90.61	86.32	85.18	85.63
ResNet-101 (ImageNet)	90.90	86.79	85.84	86.01
DenseNet (ImageNet)	85.77	79.39	78.53	78.20
EfficientNet (ImageNet)	91.31	87.00	86.94	86.70
VGGNet (places + ImageNet)	92.83	89.67	88.65	88.97

Table 7. Evaluation of the proposed visual sentiment analyzer on task 2 (i.e., multi-label classification of seven classes, namely sadness, fear, disgust, joy, anger, surprise, and neutral).

Model	Accuracy	Precision	Recall	F-Score
VGGNet (ImageNet)	82.61	84.12	80.28	81.66
VGGNet (Places)	82.94	82.87	82.30	82.28
Inception-v3 (ImageNet)	80.67	80.98	82.98	80.72
ResNet-50 (ImageNet)	82.48	84.33	79.41	81.38
ResNet-101 (ImageNet)	82.70	82.92	82.04	82.20
DenseNet (ImageNet)	81.99	83.43	81.30	81.51
EfficientNet (ImageNet)	82.08	82.80	81.31	81.51
VGGNet (places + ImageNet)	83.18	83.13	83.04	82.57

Table 8. Evaluation of the proposed visual sentiment analyzer on task 3 (i.e., multi-label classification of seven classes, namely anger, anxiety, craving, empathetic pain, fear, horror, joy, relief, sadness, and surprise).

Model	Accuracy	Precision	Recall	F-Score
VGGNet (ImageNet)	82.74	80.43	85.61	82.14
VGGNet (Places)	81.55	79.26	85.08	81.16
Inception-v3 (ImageNet)	81.53	78.21	89.30	82.27
ResNet-50 (ImageNet)	82.30	79.90	84.18	81.60
ResNet-101 (ImageNet)	82.56	80.25	84.51	81.80
DenseNet (ImageNet)	81.72	79.40	85.35	81.63
EfficientNet (ImageNet)	82.25	80.83	82.70	81.39
VGGNet (places + ImageNet)	82.08	79.36	87.25	81.99

For completeness, we also provide experimental results of the proposed methods in terms of accuracy, precision, recall, and F1-score per class. Table 9 provides the experimental results of task 1. Looking at the performances of the model on the individual classes, we can notice that some have performed comparably better than others. For instance, in the *positive* class, VGGNet pre-trained on the Places dataset performed better compared to all the other models.

Table 9. Experimental results of the proposed visual sentiment analyzer on task 1 in terms of accuracy, precision, recall, and F1-score per class. *P* represents the version of the model pre-trained on the Places dataset while the rest are pre-trained on the ImageNet dataset.

Model	Metric	Negative	Neutral	Positive
VGGNet	Accuracy	88.61	95.36	91.66
	Precision	88.45	93.20	84.56
	Recall	74.59	93.29	91.83
	F1-Score	80.85	93.22	88.04
VGGNet (p)	Accuracy	90.07	94.88	93.21
	Precision	88.63	91.13	89.87
	Recall	79.52	94.21	89.88
	F1-Score	83.79	92.64	89.85
Inception V-3	Accuracy	76.48	86.51	82.28
	Precision	70.64	79.34	78.25
	Recall	45.76	82.51	66.86
	F1-Score	55.46	80.85	71.41
ResNet-50	Accuracy	86.95	92.22	92.07
	Precision	83.40	87.15	88.14
	Recall	74.51	90.68	88.29
	F1-Score	78.65	88.86	88.170
ResNet-101	Accuracy	87.16	92.31	92.29
	Precision	86.57	86.07	87.99
	Recall	71.38	92.80	89.15
	F1-Score	78.11	89.25	88.54
DenseNet	Accuracy	80.59	87.84	87.72
	Precision	76.98	80.33	83.04
	Recall	60.16	87.01	79.54
	F1-Score	66.15	83.10	81.18
EfficientNet	Accuracy	87.50	93.91	91.66
	Precision	86.41	93.91	84.87
	Recall	72.87	92.58	91.68
	F1-Score	78.96	91.24	88.07
VGGNet (P+I)	Accuracy	89.94	94.90	92.99
	Precision	88.99	90.62	89.62
	Recall	78.44	95.17	89.58
	F1-Score	83.15	92.81	89.58

Table 10 provides the experimental results in terms of accuracy, precision, recall, and F1-score per class of task 2, where seven different categories of sentiments are considered. Overall better results are observed on class *sadness*, while the lowest performance has been observed on class *anger*, where precision and recall are generally on the lower side for most of the models.

Table 10. Experimental results of the proposed visual sentiment analyzer on task 2 in terms of accuracy, precision, recall, and F1-score per class. *P* represents the version of the model pre-trained on the Places dataset while the rest are pre-trained on the ImageNet dataset.

Model	Metric	Joy	Sadness	Fear	Digest	Anger	Surprise	Neutral
VGGNet	Accuracy	83.37	95.32	88.24	76.67	76.86	75.29	75.78
	Precision	92.17	92.46	85.09	76.78	82.13	76.96	80.31
	Recall	76.78	99.12	94.83	60.63	56.71	77.22	73.32
	F1-Score	83.77	95.67	89.68	67.68	66.99	77.07	76.35
VGGNet (p)	Accuracy	84.59	95.67	88.86	76.07	77.43	75.99	77.21
	Precision	92.44	93.47	85.47	71.19	76.23	75.21	81.58
	Recall	78.65	98.60	95.46	67.15	62.18	82.27	75.64
	F1-Score	84.99	95.97	90.19	68.78	68.33	78.58	78.43
Inception V-3	Accuracy	79.81	90.51	85.40	76.26	75.51	76.21	75.51
	Precision	89.81	86.77	81.19	86.36	86.12	71.72	75.66
	Recall	72.09	96.53	94.88	49.30	49.87	92.06	80.48
	F1-Score	79.94	91.39	87.50	62.57	62.64	80.62	77.84
ResNet-50	Accuracy	85.59	95.03	87.97	75.16	77.64	73.75	75.72
	Precision	94.16	92.71	86.18	73.83	81.89	79.31	78.49
	Recall	79.15	98.19	92.52	61.23	59.70	69.63	75.59
	F1-Score	85.99	95.37	89.22	66.43	68.83	73.92	76.91
ResNet-101	Accuracy	85.10	95.30	88.38	76.13	76.10	75.43	77.24
	Precision	88.15	93.59	86.84	76.67	76.90	74.76	79.28
	Recall	84.86	97.67	92.42	58.95	61.13	82.00	77.96
	F1-Score	86.42	95.59	89.54	66.49	67.94	78.17	78.60
DenseNet	Accuracy	83.81	93.51	87.32	76.24	76.48	75.78	75.43
	Precision	91.41	91.79	85.50	81.24	87.74	73.15	77.47
	Recall	78.47	96.16	92.07	53.72	50.52	86.83	76.85
	F1-Score	84.41	93.92	88.66	64.52	64.02	79.38	76.94
EfficientNet	Accuracy	84.40	94.84	88.38	75.70	75.78	74.83	75.67
	Precision	91.44	93.04	86.16	75.49	78.24	74.24	80.47
	Recall	79.73	97.41	93.52	63.65	59.57	81.50	72.67
	F1-Score	85.09	95.16	89.63	67.57	66.82	77.65	76.13
VGGNet (P+I)	Accuracy	83.09	95.62	89.11	77.05	77.72	77.18	77.53
	Precision	95.89	93.30	84.66	73.57	76.36	74.31	82.91
	Recall	72.66	98.71	97.33	65.50	63.15	87.78	74.46
	F1-Score	82.65	95.93	90.55	69.24	68.91	80.45	78.41

In line with the previous scenario, Table 11 provides the results in terms of accuracy, precision, and recall per class.

Table 11. Experimental results of the proposed visual sentiment analyzer on task 3 in terms of accuracy, precision, recall, and F1-score per class.

Model	Metric	Anger	Anxiety	Craving	Pain	Fear	Horror	Joy	Relief	Sadness	Surprise
VGGNet	Accuracy	73.87	86.16	80.73	82.29	87.52	79.12	84.21	81.23	95.22	70.70
	Precision	63.52	82.04	61.14	76.15	81.46	67.87	95.11	92.04	92.28	79.50
	Recall	80.28	95.39	28.40	93.73	97.88	83.85	75.09	71.63	99.59	68.24
	F1-Score	70.83	88.20	38.65	83.99	88.91	75.00	83.88	80.56	95.80	72.60
VGGNet (p)	Accuracy	74.81	83.76	79.57	80.34	86.38	78.23	82.12	77.31	95.16	72.81
	Precision	64.38	77.37	60.14	73.74	80.73	69.92	95.22	93.12	92.37	75.15
	Recall	82.81	97.11	29.77	92.17	96.83	77.25	72.07	65.11	99.39	78.35
	F1-Score	72.39	86.12	39.52	81.90	88.04	73.37	82.00	76.63	95.75	76.61
Inception V-3	Accuracy	75.76	85.29	81.40	80.96	86.24	75.70	82.43	79.46	94.36	73.03
	Precision	63.38	80.14	91.79	73.47	80.12	62.47	94.86	92.68	91.84	73.18
	Recall	92.00	96.93	14.66	96.47	97.23	87.80	71.99	67.74	98.42	84.89
	F1-Score	75.04	87.73	25.24	83.41	87.83	72.95	81.79	78.18	95.02	78.47
ResNet-50	Accuracy	72.81	85.38	79.93	81.21	86.79	79.12	85.10	81.79	94.72	71.48
	Precision	63.91	82.12	55.65	76.23	82.89	69.13	90.09	89.59	92.78	75.17
	Recall	71.93	93.39	32.81	90.31	93.53	79.91	81.94	75.27	97.97	76.37
	F1-Score	67.41	87.39	41.13	82.67	87.87	74.08	85.77	81.78	95.30	75.59
ResNet-101	Accuracy	73.09	85.40	79.84	82.71	87.46	78.62	85.29	80.87	94.72	72.31
	Precision	63.08	81.02	55.52	76.32	83.32	70.46	93.01	90.90	92.54	76.84
	Recall	77.40	95.49	32.30	94.51	94.40	74.11	79.30	72.04	98.27	75.08
	F1-Score	69.49	87.66	40.72	84.44	88.50	72.14	85.52	80.36	95.32	75.90
DenseNet	Accuracy	73.31	84.88	80.73	81.10	87.21	77.95	82.60	81.26	93.41	72.17
	Precision	63.80	81.41	67.75	74.75	83.03	66.11	90.24	89.92	92.33	74.64
	Recall	75.51	93.50	19.33	93.50	94.29	84.60	76.76	73.84	95.93	79.10
	F1-Score	67.20	87.92	38.81	84.44	88.22	75.20	83.16	80.12	95.37	77.37
EfficientNet	Accuracy	74.46	86.34	81.40	83.18	88.12	77.62	82.72	78.24	94.91	72.38
	Precision	62.07	82.16	65.22	76.69	84.58	71.05	91.86	91.26	92.43	79.86
	Recall	79.08	95.43	31.80	95.71	94.55	70.48	75.96	68.16	98.56	65.82
	F1-Score	69.06	87.03	30.02	83.07	88.28	74.09	82.85	81.05	94.08	76.72
VGGNet (P+I)	Accuracy	75.90	84.24	79.96	80.43	87.24	78.23	82.35	78.32	95.47	73.56
	Precision	64.85	77.44	66.01	72.59	81.06	67.11	95.87	94.75	92.61	77.24
	Recall	86.71	98.23	24.85	95.57	98.34	86.27	71.93	65.69	99.70	76.10
	F1-Score	74.07	86.60	35.68	82.50	88.87	75.49	82.16	77.59	96.02	76.55

5.5. Lessons Learned

This initial work on visual sentiment analysis has revealed a number of challenges, showing us all the different facets of such a complex research domain. We have summarized the main points hereafter:

- Sentiment analysis aims to extract people's perceptions of the images; thus, crowd-sourcing seems a suitable option for collecting training and ground truth datasets. However, choosing labels/tags for conducting a successful crowd-sourcing study is not straightforward.
- The most commonly used three tags, namely positive, negative, and neutral, are not enough to fully exploit the potential of visual sentiment analysis in applications such as disaster analysis. The complexity of the task increases as we go deeper into the sentiment/emotion hierarchy.
- The majority of the disaster-related images in social media represent negative (i.e., sad, horror, pain, anger, and fear) sentiments; however, we noticed that there exists a number of samples able to evoke positive emotions, such as joy and relief.
- Disaster-related images from social media exhibit sufficient features to evoke human emotions. The objects in images (gadgets, clothes, broken houses, scene-level (i.e., background, landmarks)), color/contrast, and human expressions, gestures, and poses provide crucial cues in the visual sentiment analysis of disaster-related images. This can be a valuable aspect to be considered to represent people's emotions and sentiments.
- Human emotions and sentiment tags are correlated, as can also be noticed from the statistics of the crowd-sourcing study. Thus, a multi-label framework is likely to be the most promising research direction.

6. Conclusions and Future Research Directions

In this article, we focused on the emerging concept of visual sentiment analysis, and showed how natural disaster-related images evoke people's emotions and sentiments. To this aim, we proposed a pipeline starting from data collection and annotation via a crowd-sourcing study and conclude with the development and training of deep learning models for multi-label classification of sentiments. In the crowd-sourcing study, we analyzed and annotated 4003 images with four different sets of tags resulting in four different datasets with different hierarchies of sentiments. The resulted dataset is expected to provide a benchmark for future research on the topic. It could also be utilized in other domains through transfer learning. Based on our analysis, we believe visual sentiment analysis in general and the analysis of natural disaster-related content, in particular, is an exciting research domain that will benefit users and the community in a diversified set of applications. The current literature shows a tendency towards visual sentiment analysis of general images shared on social media by deploying deep learning techniques to extract object and facial expression-based visual cues. However, we believe, as also demonstrated in this work, visual sentiment analysis can be extended to more complex images where several types of image features and information, such as object and scene-level features, human expressions, gestures, and poses, could be jointly utilized. All this can be helpful to introduce new applications and services.

We believe there is a lot to be explored yet in this direction, and this work provides a baseline for future work in the domain. For example, in the current implementation, we utilize only objects and image backgrounds. Moreover, we did not consider analyzing the impact of annotators' demographics on the performance of the proposed sentiment analyzer. We believe this could be interesting to be explored in the future by conducting a crowd-sourcing study with a particular focus on this aspect by putting some constraints on the annotators. Unfortunately, the majority of the annotators on crowd-sourcing platforms are from certain regions, which could result in longer delays in conducting crowd-sourcing studies. In the future, we also plan to extend our framework to other visual cues, such as color schemes, textures, and facial expressions. We would also like to collect a multi-modal dataset, where the text associated with images complements visual features leading to improved visual sentiment analysis.

Author Contributions: Conceptualization, K.A. and M.R.; Data curation, S.Z.H. and S.H.; Formal analysis, S.H.; Investigation, K.A. and M.R.; Methodology, K.A.; Project administration, N.C. and M.R.; Resources, P.H. and A.A.-F.; Software, S.Z.H. and S.H.; Supervision, P.H., N.C. and A.A.-F.; Validation, S.Z.H. and M.R.; Writing—original draft, K.A.; Writing—review and editing, P.H., N.C., A.A.-F. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset collected this work is publicly available, and can be found here: <https://datasets.simula.no/image-sentiment/>.

Acknowledgments: We are thankful to the Qatar National Library (QNL) for supporting the publication charges of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Öztürk, N.; Ayvaz, S. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telemat. Inform.* **2018**, *35*, 136–147. [\[CrossRef\]](#)
- Kušen, E.; Strembeck, M. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian presidential elections. *Online Soc. Netw. Media* **2018**, *5*, 37–50. [\[CrossRef\]](#)
- Sadr, H.; Pedram, M.M.; Teshnehlab, M. A Robust Sentiment Analysis Method Based on Sequential Combination of Convolutional and Recursive Neural Networks. *Neural Process. Lett.* **2019**, *50*, 2745–2761. [\[CrossRef\]](#)
- Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **2019**, *20*, 1–68. [\[CrossRef\]](#)
- Poria, S.; Majumder, N.; Hazarika, D.; Cambria, E.; Gelbukh, A.; Hussain, A. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intell. Syst.* **2018**, *33*, 17–25. [\[CrossRef\]](#)
- Said, N.; Ahmad, K.; Riegler, M.; Pogorelov, K.; Hassan, L.; Ahmad, N.; Conci, N. Natural disasters detection in social media and satellite imagery: A survey. *Multimed. Tools Appl.* **2019**, *78*, 31267–31302. [\[CrossRef\]](#)
- Imran, M.; Ofli, F.; Caragea, D.; Torralba, A. Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Inf. Process. Manag.* **2020**, *57*, 102261. [\[CrossRef\]](#)
- Ahmad, K.; Pogorelov, K.; Riegler, M.; Conci, N.; Halvorsen, P. Social media and satellites. *Multimed. Tools Appl.* **2018**, *78*, 2837–2875. [\[CrossRef\]](#)
- Hassan, S.Z.; Ahmad, K.; Al-Fuqaha, A.; Conci, N. Sentiment analysis from images of natural disasters. In Proceedings of the International Conference on Image Analysis and Processing, Trento, Italy, 9–13 September 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 104–113.
- Chua, T.H.H.; Chang, L. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Comput. Hum. Behav.* **2016**, *55*, 190–197. [\[CrossRef\]](#)
- Munezero, M.D.; Montero, C.S.; Sutinen, E.; Pajunen, J. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.* **2014**, *5*, 101–111. [\[CrossRef\]](#)
- Kim, H.R.; Kim, Y.S.; Kim, S.J.; Lee, I.K. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Trans. Multimed.* **2018**, *20*, 2980–2992. [\[CrossRef\]](#)
- Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [\[CrossRef\]](#)
- Khan, K.; Khan, R.U.; Ahmad, K.; Ali, F.; Kwak, K.S. Face Segmentation: A Journey From Classical to Deep Learning Paradigm, Approaches, Trends, and Directions. *IEEE Access* **2020**, *8*, 58683–58699. [\[CrossRef\]](#)
- Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 759–760. [\[CrossRef\]](#)
- Araque, O.; Gatti, L.; Staiano, J.; Guerini, M. DepecheMood++: A Bilingual Emotion Lexicon Built through Simple Yet Powerful Techniques. *IEEE Trans. Affect. Comput.* **2019**. [\[CrossRef\]](#)
- Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [\[CrossRef\]](#)
- Ortiz, A.; Farinella, G.; Battiato, S. Survey on Visual Sentiment Analysis. *IET Image Process.* **2020**, *14*, 1440–1456. [\[CrossRef\]](#)
- Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 83–92.
- Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 223–232. [\[CrossRef\]](#)

21. Chen, T.; Borth, D.; Darrell, T.; Chang, S.F. DeepSentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv* **2014**, arXiv:1410.8586.
22. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
23. Chandrasekaran, G.; Hemanth, D.J. Efficient Visual Sentiment Prediction Approaches Using Deep Learning Models. In Proceedings of the Iberoamerican Knowledge Graphs and Semantic Web Conference, Kingsville, TX, USA, 22–24 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 260–272.
24. Pournaras, A.; Gkalelis, N.; Galanopoulos, D.; Mezaris, V. Exploiting Out-of-Domain Datasets and Visual Representations for Image Sentiment Classification. In Proceedings of the 2021 16th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP), Corfu, Greece, 4–5 November 2021; pp. 1–6.
25. Al-Halah, Z.; Aitken, A.P.; Shi, W.; Caballero, J. Smile, Be Happy :) Emoji Embedding for Visual Sentiment Analysis. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 4491–4500.
26. Huang, F.; Wei, K.; Weng, J.; Li, Z. Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 1–19. [[CrossRef](#)]
27. Gelli, F.; Uricchio, T.; He, X.; Del Bimbo, A.; Chua, T.S. Learning subjective attributes of images from auxiliary sources. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2263–2271.
28. You, Q.; Jin, H.; Luo, J. Visual sentiment analysis by attending on local image regions. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
29. Ou, H.; Qing, C.; Xu, X.; Jin, J. Multi-Level Context Pyramid Network for Visual Sentiment Analysis. *Sensors* **2021**, *21*, 2136. [[CrossRef](#)]
30. Wu, L.; Zhang, H.; Deng, S.; Shi, G.; Liu, X. Discovering Sentimental Interaction via Graph Convolutional Network for Visual Sentiment Prediction. *Appl. Sci.* **2021**, *11*, 1404. [[CrossRef](#)]
31. Yadav, A.; Vishwakarma, D.K. A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimed. Syst.* **2020**, *26*, 431–451. [[CrossRef](#)]
32. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
33. Wang, X.; Jia, J.; Yin, J.; Cai, L. Interpretable aesthetic features for affective image classification. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, VIC, Australia, 15–18 September 2013; pp. 3230–3234.
34. Ortis, A.; Farinella, G.M.; Torrisi, G.; Battiato, S. Exploiting objective text description of images for visual sentiment analysis. *Multimed. Tools Appl.* **2021**, *80*, 22323–22346. [[CrossRef](#)]
35. Katsurai, M.; Satoh, S. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2837–2841.
36. Wang, J.; Fu, J.; Xu, Y.; Mei, T. Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), New York, NY, USA, 9–15 July 2016; pp. 3484–3490.
37. Peng, K.C.; Sadovnik, A.; Gallagher, A.; Chen, T. Where do emotions come from? Predicting the emotion stimuli map. In Proceedings of the 2016 IEEE international conference on image processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 614–618.
38. Cowen, A.S.; Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E7900–E7909. [[CrossRef](#)] [[PubMed](#)]
39. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 487–495.
40. Ahmad, K.; Conci, N. How Deep Features Have Improved Event Recognition in Multimedia: A Survey. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2019**, *15*, 39. [[CrossRef](#)]
41. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

-
45. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
 46. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
 47. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.