applied sciences

MDPI

Article

Multimodal Sentiment Classifier Framework for Different Scene Contexts

Nelson Silva, Pedro J. S. Cardoso and João M. F. Rodrigues *

NOVA LINCS & ISE, Universidade do Algarve, 8005-139 Faro, Portugal; a60678@ualg.pt (N.S.); pcardoso@ualg.pt (P.J.S.C.)
* Correspondence: jrodrig@ualg.pt

Abstract: Sentiment analysis (SA) is an effective method for determining public opinion. Social media posts have been the subject of much research, due to the platforms' enormous and diversified user bases that regularly share thoughts on nearly any subject. However, on posts composed by a text–image pair, the written description may or may not convey the same sentiment as the image. The present study uses machine learning models for the automatic sentiment evaluation of pairs of text and image(s). The sentiments derived from the image and text are evaluated independently and merged (or not) to form the overall sentiment, returning the sentiment of the post and the discrepancy between the sentiments represented by the text–image pair. The image sentiment classification is divided into four categories—"indoor" (IND), "man-made outdoors" (OMM), "non-man-made outdoors" (ONMM), and "indoor/outdoor with persons in the background" (IOwPB)—and then ensembled into an image sentiment classification model (ISC), that can be compared with a holistic image sentiment classifier (HISC), showing that the ISC achieves better results than the HISC. For the Flickr sub-data set, the sentiment classification of images achieved an accuracy of 68.50% for IND, 83.20% for OMM, 84.50% for ONMM, 84.80% for IOwPB, and 76.45% for ISC, compared to 65.97% for the HISC. For the text sentiment classification, in a sub-data set of B-T4SA, an accuracy of 92.10% was achieved. Finally, the text–image combination, in the authors' private data set, achieved an accuracy of 78.84%.

Keywords: sentiment analysis; affective computing; human-centered AI; Multimodal Sentiment Classifier

Citation: Silva, N.; Cardoso, P.J.S.; Rodrigues, J.M.F. Multimodal Sentiment Classifier Framework for Different Scene Contexts. *Appl. Sci.* **2024**, *14*, 7065. https://doi.org/10.3390/app14167065

Academic Editors: Xiaoming Zhang, Francisco De Arriba-Pérez and Silvia García-Méndez

Received: 16 July 2024
Revised: 7 August 2024
Accepted: 8 August 2024
Published: 12 August 2024

1. Introduction

The goal of human-centered artificial intelligence (HCAI) is to create technologies that assist people in carrying out various daily tasks, while also advancing human values, such as rights, fairness, and dignity [1]. As an interdisciplinary area involving computer science, psychology, and neuroscience, HCAI aims to achieve a balance between human control and (complete) automation by improving people's autonomy, well-being, and influence over future technologies. Affective computing (AffC) is a related field that combines the fields of sentiment analysis and emotion recognition. AffC is backed by a variety of physical information types, including text, audio (speech), visual data (such as body posture, facial expression, or environment), and physiological signals (such as electroencephalography or electrocardiograms). AffC may be developed using either unimodal or multimodal data within this framework [2].

HCAI and AffC have a wide range of applications. For example, in lato sensu, a machine needs to be designed to cooperate with or learn how to function in interpersonal relationships with people. Emotions and sentiments are fundamental to human–machine relationships, and any robot communicating with humans must incorporate them. Conversely, social media platforms are becoming increasingly significant in today's digital marketing landscape, because they influence individuals to travel, look for and purchase goods, alter their lives, and alter their perspectives on many topics. The sheer number

of daily posts on various platforms has made it necessary and simultaneously possible to monitor, evaluate, and comprehend the mood, sentiments, and emotions that such messages convey.

Considerable progress has been achieved in the area of sentiment analysis [3–6], including, but not limited to, text, image, and text–image posts (even in cases when the image and words together may convey a different meaning), music, and video analysis, or even robots that can read facial emotions and improve human-robot relations.

The quickest and easiest kind of publications to get people to click, buy, and read about a specific topic or product include short texts, images, and/or videos. This explains why social media platforms like Instagram and X (formerly known as Twitter) have gained much traction in recent years.

There are two main categories of posts where the pair text–image is connected as follows:

(1) *Where the text is (clearly) complemented by an image(s) of a person or group of persons*. In those cases, the environment (scene) is the background and the person(s) are in the foreground, i.e., they are looking in a (semi-)frontal manner at the camera, so their facial expressions or body posture have a powerful influence on the sentiment carried [7].

(2) *Where the text may or may not be complemented by the image(s)*. In those cases, there are no persons in the scene, or if existing, they are the "background", and the scene is the "foreground". In those cases, color, texture, edges, line type, orientation, etc., are important features in the attraction and sentiment that the image carries [8].

This paper focuses on the second case, i.e., detecting sentiments in posts relating to pair text–image(s) without persons, or existing persons who are only in the "background". In this case, in the literature, there is currently no good sentiment classifier for this type of image and text combination (see next section). To the best of our knowledge, sentiment classifiers that deal with this type of image use a holistic approach, i.e., the models are trained with all types of images, without considering their specific characteristics. Here, the images are segmented into four categories as follows: "non-man-made outdoors" (ONMM), "man-made outdoors" (OMM), "indoor" (IND), and "indoor/outdoor with persons only in the background" (IOwPB). Each class has partial results that are combined, resulting in a final model, which replaces the holistic approach.

The present proposal for the image sentiment classifier (ISC) has been developed based on three deep learning (DL) models, which were fine-tuned from one developed for the class ONMM. This work was presented in an earlier authors' publication [3]. The principle of using this category to develop the baseline DL models was that it is a more generic category and that it is expected to return a good baseline for the remaining categories. If dedicated DL models are developed for each category, the final ISC model will achieve a higher accuracy. Nevertheless, the authors' investigation hypothesis is that when using the four models/categories to develop the ISC, even if those are not fine-tuned for each category, the final result (accuracy) will be better than using a single holistic model.

Considering the above principle, it was applied an ensemble of the DL models to achieve the final ISC_ONMM model, ISC_ONMM = E{DL#1, DL#2, DL#3}, with E denoting the ensemble. The same principle and models, without additional tuning, were trained and applied to the other three categories (OMM, IND, and IOwPB). The ensemble of the models, ISC = E{ISC_ONMM, ISC_OMM, ISC_IND, ISC_IOwPB}, was then combined with machine learning (ML) models for the text sentiment classification (TSC), see [2] for more details, returning the final multimodal sentiment classification (MSC) model.

In this context, where the text and image(s) may or may not convey the same sentiment, the information about the sentiment discrepancy between the text and image depends on the individual results (for each modality). This discrepancy is used to decide if the image(s) should complement the sentiment information or simply state that the text and image represent completely different sentiments. In those cases, probably, the intention of the user/poster is just to select an image to illustrate the post, not to reflect their sentiment, or to be sarcastic.

In summary, this work presents a multimodal text–image sentiment classifier framework. In the case of the image, the image sentiment classifier has several classifiers trained with different segments (four), that can be compared with a holistic image sentiment classifier (IHSC), trained with all images available. The sentiment results from the images are combined or not (depending on the discrepancy) with the text sentiment results, returning the sentiment classification and discrepancy between the image- and text-predicted sentiments.

The main contributions of this work include twofold as follows: (1) the (single hyperparameter) image classification sentiment model that works in different scenes/environments (ONMM, OMM, IND, and IOwPB), and (2) the framework that combines image and text sentiment classification, returning the multimodal sentiment classification attach with the discrepancy (text–image sentiment) metric.

The main contributions of this work include threefold as follows: (1) the (single hyperparameter) image classification sentiment model that works in different scenes/environments (ONMM, OMM, IND, and IOwPB); (2) the consideration of the discrepancy between the image and text sentiment, which is not present in the literature, and can be used to decide if the text or image should be used to complement the sentiment information or not (e.g., this is particularly relevant for a post that may be sarcastic); and (3) the framework that combines image and text sentiment classification, returning the multimodal sentiment classification attach with the discrepancy (text–image sentiment) metric.

The present section introduces the work's goal, Section 2 presents the contextualization and state of the art, Section 3 introduces the data sets used, Section 4 details the models, and Section 5 outlines the tests, results, and respective discussion. Finally, Section 6 presents the conclusions and future work.

2. Contextualization and State of the Art

The affiliation, warmth, friendliness, and dominance between two or more individuals are displayed when relationships are made, returned, or deepened [9]. Opposite, impersonal human–machine interactions impede more extensive communication and complicate the establishment of intimate or reciprocal relationships between people and machines, devices, or interfaces. Within this framework, automated data evaluation systems to determine the conveyed emotion are known as automatic emotion analysis approaches [10] (categorized, e.g., as happiness, sadness, fear, surprise, disgust, anger, or neutral) or sentiment [11,12] (typically limited to positive, negative, and neutral).

Considerable advancements have been achieved in the field of sentiment categorization in recent years. Although sentiments and emotions are distinct entities, it should be noticed that they are interconnected [12,13], i.e., emotions affect sentiments and sentiments affect emotions. The sentiment may be impacted by a variety of elements including attitudes, opinions, emotions, prior experiences, cultural background, personal views, age, or even gender. It is a mental attitude that is connected to a good, negative, or neutral evaluation or thinking about anything [13].

As an example, color can be considered an important feature for sentiment classification [8,14–16]. Typically, images of beaches or oceans predominantly feature blue tones, evoking a sense of calm. In contrast, an image of a forest will highlight green tones, which are associated with harmony. Nevertheless, the meaning of a color can change depending on the context in which it is used; for example, the color red can sometimes represent anger, love, or frustration. Also, different individuals may perceive color differently on an emotional level, just like they do with music.

Different authors categorize emotions in a variety of ways and break them down into levels and sublevels [13]. Six fundamental emotions, which are typically accompanied by the neutral emotion, is the classification used by the renowned psychologist Paul Ekman [10]. This group of emotions is commonly used for facial emotion classification. Other authors have also put forward different categories. For example, based on biological mechanisms, Robert Plutchik [17] defined the following eight basic/primary emotions:

joy, trust, fear, surprise, sorrow, disgust, anger, and anticipation. Plutchik created a color wheel, known as Plutchik's wheel, to symbolize feelings, with a particular color assigned to each. Emotions in this instance may be categorized according to their intensities and combinations; that is, primary emotions are coupled in various ways to generate secondary and tertiary emotions, which are symbolized by various color tones and hues, for a total of 24 emotions.

In summary, Plutchik's wheel categorizes emotions into the following two primary sentiments: positive and negative. Joy, trust, anticipation, and surprise are examples of good sentiments; on the other hand, sorrow, contempt, fear, and rage are examples of negative sentiments. As it was discussed before [16], the division of emotions into positive and negative sentiments can be subjective and based on individual and cultural variables. Compared to emotion, which can shift quickly in reaction to shifting circumstances and stimuli, sentiment is often longer and more consistent [18].

Ortis et al. [13] provided a summary of sentiment analysis in images. The authors outlined the prospects and difficulties in the field and discussed the main problems. To classify the content of composite comments on social media, a multimodal sentiment analysis (text and image) model was published by Gaspar and Alexandre [19]. The three primary components of the technique are an image classifier, a text analyzer, and a method that examines an image's class content, determining the likelihood that it falls into one of the potential classes. By combining deep features with semantic information obtained from the scene characteristics, the authors also assess how classification and cross-data set generalization performance might be enhanced. The authors used the T4SA data set [20], as the source of their study, which consists of three million tweets—text and photos—divided into three sentiment categories (positive, negative, and neutral).

In [8], a color cross-correlation neural network for sentiment analysis of images was introduced. The architecture considers the relationships between contents and colors in addition to utilizing them concurrently. The authors collected color features from several color spaces using a pretrained convolutional neural network to extract content characteristics and color moments. Then, using a sequence convolution and attention mechanism, they present a cross-correlation method to model the relationships between content and color features. This method integrates these two types of information for improved outcomes by enhancing the sentiment that content and color express.

In [21], the authors suggest a system to categorize the tone of outdoor photos that people post on social media. They examine the differences in performance between the most advanced ConvNet topologies and one created especially for sentiment analysis. The authors also assess how classification and cross-data set generalization performance might be enhanced by combining deep features with semantic information obtained from the scene characteristics. Finally, they note that the accuracy of all the ConvNet designs under study is enhanced by the integration of knowledge about semantic characteristics.

A deep-learning architecture for sentiment analysis on 2D photos of indoor and outdoor environments was presented by Chatzistavros et al. [22]. The emotion derived from catastrophe photographs on social media is examined in [23]. A multimodal (text and image) sentiment classification model based on a gated attention mechanism is provided in [24]. In the latter, the attention mechanism uses the image feature to highlight the text segment, allowing the machine to concentrate on the text that influences the sentiment polarity. Furthermore, the gating method allows the model to ignore the noise created during the fusion of picture and text, retaining valuable image information.

More examples can be found in [25–27] (see also Table 1) and in the very recent work presented in [28], which presents the Controllable Multimodal Feedback Synthesis (CM-Feed) data set, that enables, according to the authors, the generation of sentiment-controlled feedback from multimodal inputs. The data set contains images, text, human comments, comments' metadata, and sentiment labels. The authors propose a benchmark feedback synthesis system comprising encoder, decoder, and controllability modules. It employs

transformer and Faster R-CNN networks to extract features and generate sentiment-specific feedback, achieving a sentiment classification accuracy of 77.23%.

Focusing on the large language model (LLM), other recent models exist. For example, in [29], the authors use transformers and LLM for sentiment analysis of foreign languages (Arabic, Chinese, etc.) by translating them into a base language—English. The authors start by using the translation models LibreTranslate and Google Translate, and the resulting sentences were then analyzed for sentiment using an ensemble of pretrained sentiment analysis models, like Twitter-Roberta-Base-Sentiment-Latest, Bert-base-multilingual-uncased-sentiment, and GPT-3. A 2024 survey about LLM and multimodal sentiment analysis can be found in [30].

Furthermore, recent multimodal methods, such as CLIP, BLIP, and VisualBERT (see for instance [31] or [32] for details), achieve excellent results in handling multimodal data. Nevertheless, some studies, like the one from Mao et al. [32], also suggest that using different pretrained models for prompt-based sentiment analysis introduces biases, which can impact the performance of the model. Deng et al. [31] also addressed those models (CLIP, BLIP, and VisualBERT), validating that they are excellent models, but mentioning also, as drawbacks, that they frequently have a lot of parameters and need image–text pairings as inputs, which limits their adaptability. The latter authors, Deng et al., implemented MuAL, which utilizes pretrained models as encoders. A cross-modal attention is used to extract and fuse sentiment information embedded in both images and text with a difference loss incorporated into the model, to discern similar samples in the embedding space. Finally, a token (*cls*) was introduced preceding each modality's data to represent overall sentiment information. In a vision language pretraining model based on cross-attention (VLPCA) [33], a multihead cross-attention to capture both textual and visual elements was used to improve the representation of visual–language interactions. In addition, to improve the performance of the model, the author created two subtasks and suggested a new unsupervised joint training strategy based on contrastive learning.

It is crucial to emphasize that there are a huge number of models available for text analysis [8], with the typical steps being as follows:

(1) *Processing text*: Using methods like tokenization, stop word removal, stemming, lemmatization, emoticon and emoji conversion, and deleting superfluous material, resulting in a "clean" text to improve the sentiment prediction accuracy.

(2) *Extraction of Features*: Words, in this context, are relevant features that are extracted from the preprocessed text. The most popular method for doing this is to use strategies like n-grams and bag-of-words.

(3) *Model Development*: Develop a DL model or a machine learning model (such as a random forest or decision tree) to learn from data and subsequently accurately classify newly unknown text sentiment.

(4) *Sentiment Classification Evaluation* occurs when sentiment analysis is performed. To achieve the highest performance possible, the combination of several models and model adjustments is also common usage.

Table 1 summarizes different approaches/models to multimodal sentiment analysis. In parentheses is a small citation that points out the model. As can be seen, most of the models use different data sets, but all consider that text and images always have the "same" sentiment, not considering that the text can be a specific sentiment, and the image can be a different one, or is only there to illustrate the post. Saying that, some of the models process the image and text separately, but in the end, join both (text and image) without any added consideration.

Also, it is possible to verify that, despite not being comparable, once there are different data sets and or different ways to use the (sub-)data sets, the accuracy (accr.) of the models is around 70 to 80%. As a note, in the last model presented in the table, the authors do not present the accuracy, only precision (P) and recall (R).

In the present paper, ensemble/stacking modeling is suggested, which often enables data mining and predictive analytics applications to become more accurate. The process

of running two or more related but independent analytical models and then integrating the results into a single score or spread is known as ensemble modeling, or fusion, in this context. In this instance, we can associate various outcomes from various/supplementary models to maximize accuracy or to provide complementary data. You may find examples of these methods, e.g., in [18,24]. The preprocessing procedures used on the data before analysis, as well as the sub-data sets generated to evaluate the model, will be briefly discussed in the next section.

Table 1. Summary of the multimodal approaches and respective accuracy.

Model (Brief Text Citation)	*Ref.*	*Year*	*Data Set*	*Type*	*Accr.*
Deep Model Fusion ("... reduces the text analysis dependency on this kind of classification giving more importance to the image content...")	[19]	2019	B-T4SA	Text-img.	60.42%
Gated Attention Fusion Network (GAFN) ("... image feature is used to emphasize the text segment by the attention mechanism...")	[24]	2022	Yelp restaurant review data set	Text-img.	60.10%
Textual Visual Multimodal Fusion (TVMF) ("... explore the internal correlation between textual and visual features...")	[25]	2023	Assamese news corpus	Text-img.	67.46%
Hybrid Fusion Based on Information Relevance (HFIR) ("... mid-level representation extracted by a visual sentiment concept classifier is used to determine information relevance, with the integration of other features, including attended textual and visual features ...")	[26]	2023	Authors' data set	Text-img.	74.65%
Deep Multi-Level Attentive Network (DMLANet) ("... correlation between the image regions and semantics of the word by extracting the textual features related to the bi-attentive visual features...")	[27]	2023	MVSA multiple Flickr	Text-img.	77.89% 89.30%
Transformer and Faster R-CNN Networks ("... controllable feedback synthesis to generate context-aware feedback aligned with the desired sentiment...")	[28]	2024	CMFeed	Text-img.	77.23%
Ensemble Model of Transformers and LLM ("... sentences were then analyzed for sentiment using an ensemble of pre-trained sentiment analysis models...")	[29]	2024	Compilation of several data sets	Text—"foreign languages"	86.71%
Multimodal sentiment analysis approach (MuAL) ("... cross-modal attention is used to integrate information from two modalities, and difference loss is utilized to minimize the gap between image and text information...")	[31]	2024	MVSA single, MVSA multiple, Hateful Memes, Twitter2015, Twitter2017	Text-img.	80.78% 77.77% 79.15% 79.34% 80.39%
Vision-Language Pre-training model based on cross-attention (VLPCA) ("... multi-head cross attention to capture textual and visual features for better representation of visual-language interactions...")	[33]	2024	Twitter2015, Twitter2017	Text-img.	(P & R) 71.20% 72.80%, 73.40% 74.00%

Lastly, to the best of our knowledge, no framework or model exists in the literature that considers the possibility that the post may elicit different reactions from readers to the image and text, regardless of the author's motivation for doing so. Also, assessing images of environments according to classes and integrating that data with the text seems to be missing from literature. The next section describes the data sets used for the implementation of the proposed models and frameworks.

3. Data Set

In the present study, the following three main groups of data sets were used: (sub-)data sets to develop and validate the Image Sentiment Classifier models (see Section 4.1), a data set to build and validate the Text Sentiment Classifier models (see Section 4.2), and data sets used to develop and test the integrated model, i.e., the Multimodal Sentiment Classifier model (see Section 4.3).

3.1. Image Sentiment Data Sets

In the image sentiment study, we used three image data sets, one being further divided into sub-data sets. In this context, from the (i) **Flickr data set** (Flickr data set available at: https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html, accessed on 1 August 2024) [34], four sub-data sets are extracted. The reason is that the Flickr data set includes outdoor, indoor, man-made scenes, scenes with persons, etc., all mixed together, and it is intended to be used to train four distinct ISC models (see Section 4.1). To do this subdivision, a scene detection algorithm (code available at: https://github.com/AMANVerma28/Indoor-Outdoor-scene-classification, accessed on 1 August 2024) was used, which classifies whether an image is outdoor or indoor, and some other attributes, such as man-made, natural light, no horizon, enclosed area, etc. This algorithm, out of the focus of this paper, was used to speed up the segmentation of the initial data set into the four sub-data set. The resulting subsets of images were then visually validated by the authors. These sub-data sets were used to train and test the ISC models (see Section 4.1). Furthermore, although in Flickr's original data set, the sentiment polarity of the images ranges from −2 to 2, the present models have only three sentiment classes: *positive*, *neutral*, and *negative*. This was achieved by defining a negative sentiment as between −2 and −1, a neutral sentiment as between −1 and 1, and a positive sentiment as between 1 and 2.

In more detail, the Flickr data set was created from the computed Columbia Multilingual Visual Sentiment Ontology (MVSO) framework. This database covers a visual sentiment ontology consisting of 3244 adjective–noun pairs and SentiBank, which is a set of 1200 trained visual concept detectors, providing a midlevel representation of sentiment, associated training images acquired from Flickr, and a benchmark containing 603 photo tweets covering a diverse set of 21 topics. With ~461 k images, the Flickr data set is not balanced, so, ~179 k balanced images were randomly selected from those 461 k images (see Table 2, column "balanced"). From those, the images were subdivided into four sub-data sets (as mentioned above), one for each of the four environment categories as follows: (i.1) **Flickr_ONMM**; (i.2) **Flickr_OMM**; (i.3) **Flickr_IND**; and (i.4) **Flickr_IOwPB**. Since those sub-data sets were not balanced, for training and testing, when needed, they were balanced using data augmentation, namely, small rotations (5 degrees left/right), horizontal flips and zooms in (10% to 30%) on the images. Table 2 shows the number of images of the data set as well as the sub-data sets and Figure 1 shows examples of images existing in the data set.

Table 2. Summary of the Flickr sub-data sets for development of the image sentiment classifiers.

Sentiment	*Flickr*		*Sub-Data Sets (Flickr)*			
	Original	**Balanced**	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	280,157	**59,794**	32,878	28,653	12,664	57,639
Negative (−)	121,377	**59,794**	32,878	28,653	12,664	57,639
Neutral (=)	59,794	**59,794**	32,878	28,653	12,664	57,639
Total	461,328	**179,382**	98,634	85,959	37,992	172,917

Figure 1. Left to **right**, examples of images for the four categories extracted from the Flickr data set, i.e., ONMM, OMM, IND, and IOwPB. **Top** to **bottom**, examples of images with positive, neutral, and negative sentiments.

(ii) The **Simula Image Sentiment data set (SIS)** (SIS data set is available at: https://datasets.simula.no/image-sentiment/, accessed on 1 August 2024) [23] is a disaster-related data set with ~3.7 k images that was created by five different people. The ground truth classification was performed by humans and varies between 1 (highly negative) and 9 (highly positive), with values between 1–3 being considered as negative sentiment, 4–6 as neutral sentiment, and 7–9 as positive sentiment. It is important to stress it is only a disaster-related data set.

(iii) The **Image Sentiment Polarity data set (ISP)** (ISP data set is available at: https://data.world/crowdflower/image-sentiment-polarity, accessed on 1 August 2024) [35] contains over 12,000 sentiment-scored images where the ground truth was at least approved by one human. There were five ground truth options: images classified as "Negative" or "Highly Negative" were associated with the negative sentiment, "Neutral" with the neutral sentiment, and "Positive" or "Highly Positive" with the positive sentiment. Table 3 summarizes the number of samples in the SIS & ISP data sets and sub-data sets, and Figure 2 shows examples of those samples.

Table 3. Sub-data sets are used for testing the inference of the ISC model.

Sentiment	*SIS & ISP*	*Sub-Data Sets (SIS & ISP)*			
	Original	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	8828	1908	2696	1552	2672
Negative (−)	3996	242	1327	581	1846
Neutral (=)	2963	273	951	709	1030
Total	**15,787**	**2423**	**2842**	**4974**	**5548**

It is important to stress that SIS & ISP data sets are used solely to infer the results of the ISC model, as they are the only data sets from the presented group that have been validated with human-verified ground truth. It is also important to emphasize that these two data sets are not balanced.

Figure 2. Left to **right**, examples of images for the four categories (ONMM, OMM, IND, and IOwPB); **top** to **bottom**, positive, neutral, and negative sentiments for the SIS & ISP data sets.

3.2. Text Sentiment Data Set

For the text sentiment study, **the T4SA data set** (T4SA is available at: http://www.t4sa.it/#dataset, accessed on 1 August 2024) [20] was used. The data set contains ~3.4 M tweets, corresponding to ~4 M images, as each tweet may have more than one image. Each tweet, text and associated image(s) has been labeled according only to the sentiment polarity of the text, namely negative (−1), neutral (0), or positive (1). The data set's authors removed corrupted and near-duplicate images and selected a balanced subset of images, named B-T4SA (more details in [3,20]). From that data set, consisting of ~379 k unbalanced samples, a sub-data set **B-T4SAtext**, which has 50 k balanced samples, was randomly selected for the TSC models (see Section 4.2). Table 4 shows the distribution of B-T4SAtext samples per sentiment category.

Table 4. Summary of the B-T4SAtext sub-data sets.

Sentiment	***B-T4SA***	***B-T4SAtext***
	Original	**Balanced**
Positive (+)	127,086	16,667
Negative (−)	21,643	16,667
Neutral (=)	203,471	16,667
Total	389,200	50,001

To this sub-data set was applied the following preprocessing steps for text analysis: (a) the replacement of emojis/emoticons for words (e.g., 😻 was replaced by "smiling face with hearts"); (b) convert all text to lowercase; (c) remove stop words (e.g., "i", "me", "after", "moreover"), which proves to be useful because stop words do not contribute to sentiment analysis and their exclusion avoids unnecessary computations; (d) removed HTML tags, images, mentions, links, punctuation etc., because they do not carry sentiments; (e) apply a lemmatizer, which removes inflectional endings from a token to turn it into the base word lemma (e.g., the word "dancing" would be lemmatized to "dance"); and (f) apply stemming, which is the process of removing suffices from words to obtain their root form (e.g., the word "dancing" would be stemmed to "danc"). Both stemming and lemmatizations serve the purpose of reducing word forms to their base or root forms to generalize the words, resulting in more accurate predictions in sentiment detection (see also [3]). Table 5 shows examples of some texts from the **B-T4SAtext** data set, alongside the same texts preprocessed according to the previously explained method.

Table 5. Examples of preprocessed text used for the model's development. On the left is the sentiment, in the middle is the post, and on the right is the preprocessed text used as input for the TSC model.

Sentiment	Text	Preprocessed Text
Negative (−)	RT@Reuters: Special Report: Iraq militia massacre worse than U.S. acknowledged. https://t.co/KhzqwloPam (accessed on 1 August 2024)"	special report iraq militia massacr wors u acknowledg
Positive (+)	Have a look at the most spectacular #NationalParks in #California. #traveler https://t.co/wiIfw7nnH7 https://t.co/sxnJnQH0cC (accessed on 1 August 2024)	look spectacular nationalpark california travel
Neutral (=)	Live Next Door To Parents https://t.co/FfJrf5n8D1 https://t.co/2bZRdnkPQY (accessed on 1 August 2024)	live next door parent

3.3. Multimodal Sentiment Data Set (Image + Text)

The **B-T4SAmultimodal** sub-data set was extracted from the T4SA data set [20], being used in the MSC model (see Section 4.3). For this data set, 1000 text posts have been randomly selected, carefully chosen by sentiment class, to ensure that the text samples are balanced. The text sentiment label was directly retrieved from the B-T4SA data set, but the T4SA did not provide sentiment labeling (ground truth) for the images.

From each post, a single image per post was selected, presented and classified by a group of 10 persons, 6 male and 4 female, aged between 20 and 53 years, all with Portuguese nationality. The data set then was filtered to include only those cases where a minimum of 6 out of the 10 individuals unanimously agreed on the sentiment classification (positive, negative, or neutral). This approach led to an unbalanced text and image (sub-)data set, see Table 6, column "Used". This imbalance was due not only to the filtering process but also to other factors, such as the diversity of images/text sentiments pairing.

Table 6. Summary of the B-T4Smultimodal sub-data set.

Sentiment	*B-T4SA ("TSC")*	*B-T4Smultimodal*					
		Text		*Images*			
	Original	**Preclassification**	**Used**	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	127,086	333	195	200	117	53	25
Negative (−)	21,643	334	177	18	45	19	20
Neutral (=)	203,471	333	255	3	26	84	17
Total	**389,200**	1000	**627**	**221**	**188**	**156**	**62**

For example, images classified as positive, neutral, or negative were sometimes paired with texts that had different sentiment classifications (negative, neutral, or positive). Consequently, after this image classification, from the initial 1000 posts samples, only 627 text–image samples remained. Table 6 shows the distribution of samples per sentiment used for training and testing the Multimodal Sentiment Classifier model.

In summary, the sub-data sets Flickr_ONMM, Flickr_OMM, Flickr_IND, and Flickr_IOwPB were used to train and test the ISC models; the SIS&ISP_ONMM, SIS&ISP_OMM, SIS&ISP_IND, and SIS&ISP_IOwPB were used only to test the ISC models; the (sub-)data set B-T4SAtext was used to train and test the TSC models; and the B-T4Smultimodal_text, B-T4Smultimodal_ONMM, B-T4Smultimodal_OMM, B-T4Smultimodal_IND, and B-T4Smultimodal_IOwPB to train and test the MSC models. In the next section the Multimodal Sentiment Classification Framework and respective sub-modules are presented.

4. Multimodal Sentiment Classification Framework

As already mentioned, the Multimodal Sentiment Classification model (MSC) is used to extract sentiments from images and texts, following the principles presented in [3] and [19]. However, the present model exhibits distinctions, while [3], the previous work of the present authors, focused only on posts (text–image) associated with landscapes. In [19], the authors achieved 60.42% accuracy on a test set of 51 k samples from a B-T4SA image-balanced sub-data set, with the currently considered three classes (negative, neutral, and positive). Nevertheless, in [19], the authors did not consider that a post can have more than one associated image and the fact that image and text can reflect the same or different sentiments. Additionally, a post with multiple images may present different sentiments, which may or may not coincide with the text sentiment. The present work focuses on improving accuracy, but also, most importantly on differentiating-marking posts images and texts that reflect different sentiments.

In these works, ensembles are employed to improve the framework's accuracy. Namely, image and text sentiment classification results from the combining of various methods, i.e., components are combined into an ensemble to yield the ultimate result.

In this context, the possible outputs are as follows (see Figure 3): (i) the sentiment (*isc*) resulting from the image (ISC), generated by the Image Sentiment Classification block; (ii) the sentiment (*tsc*) resulting from the text (TSC), generated by the Text Sentiment Classification block; the (iii) sentiment (*msc*) resulting from the combination of image and text (MSC), generated by the Multimodal Sentiment Classifier block; and (iv) the discrepancy (*dis.*) between image and text. So, for each pair (text–image), the model returns the following sentiment classifier vector SCv = [*isc*, *tsc*, *msc*, *dis.*].

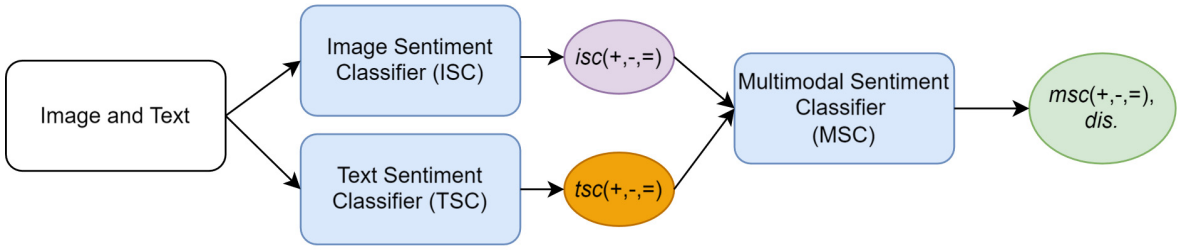

Figure 3. Multimodal Sentiment Classification Framework.

Figure 3 shows a simplified diagram block of the model, where "+" represents a positive sentiment, "−" a negative sentiment, "=" the neutral sentiment, and *dis.* is the discrepancy between the sentiment returned from the image and the text, for each pair of text–image presented in the input.

4.1. Image Sentiment Classification

The Image Sentiment Classification model combines several individual image sentiment classifier models into an ensemble that predicts the final sentiment. In this context (see Figure 4 top), the ISC is made up of four blocks that correspond to each class of sentiment classifier, i.e., the ISC is the result of the ensemble of the results of the ISC for non-man-made outdoors (ISC_ONMM), for man-made outdoors (ISC_OMM), for indoor (ISC_IND), and for indoor/outdoor with persons in the background (ISC_IOwPB).

Each sentiment classifier, ISC_*class*, with *class* ∈ {ONMM; OMM; IND; IOwPB}, returns the probabilities of an image carrying a negative, neutral, or positive sentiment. The outputs of the four class sentiment classifier blocks are then ensembled using a random forest (ISC_RF) and a neural network (ISC_NN). Each of these individual blocks, ISC_*class*, return results for each category (*class*), again doing an ensemble of the three DL models that will be presented below, this is illustrated in Figure 4 bottom. This last step follows the author's previous work carried out for ONMM [3].

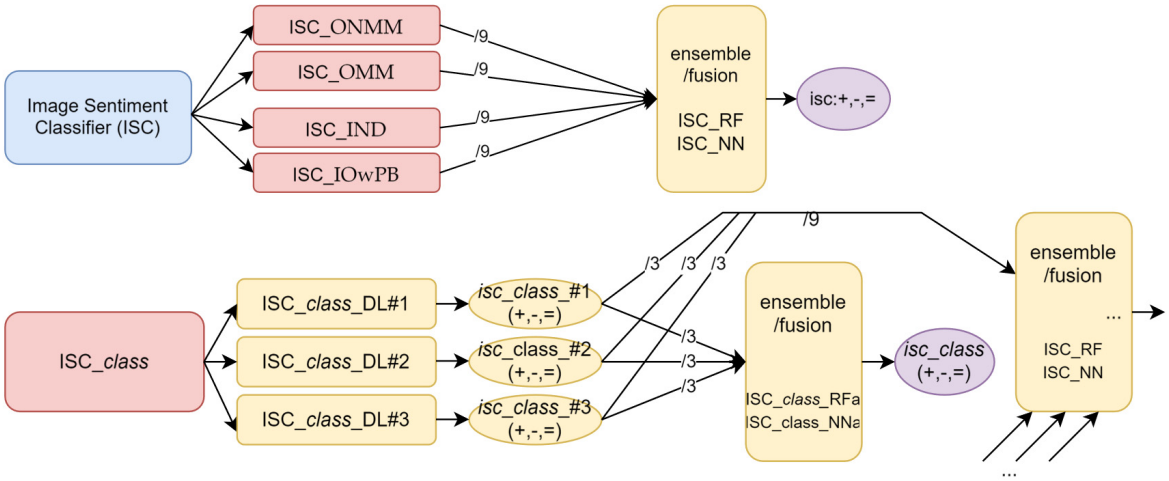

Figure 4. (**Top**) the ISC model, (**bottom**) the ISC sub-block specification, with *class* ∈ {NMMO; MMO; IND; IOPB}.

In summary, ISC_RF or ISC_NN fuses the 36 probabilities given by the four blocks (nine answers for each block resulting from the three probabilities outputs by each of the three individual models) to finally decide the final sentiment of an image (ISC). The number of individual models is a hyperparameter that can be tuned.

The abovementioned three models, for each ISC_*class*, have DL architectures. The architectures are based on backbones from known architectures followed by a handcrafted network head, based on fully connected layers. In this scenario, three distinct backbones were used to extract different types of features, namely, EfficientNetB0 [34], Xception [36], and ResNet-50 [37]. It is worth noting that all backbones were trained using the well-known ImageNet data set.

Different strategies of transfer learning were used for the models, including different numbers of fully connected layers at the networks' heads. Nevertheless, all ISC_*class* blocks have the same three individual models, with the same architecture, and hyperparameter, that were only optimized for the ONMM *class*. What differentiates an individual classifier from the others is the class category of the sub-data set images (Flickr_ONMM, Flickr_OMM, Flickr_IND and Flickr_IOwPB) that were used to train the individual and ensemble models. The deep learning models are as follows:

(a) **Model DL#1:** the backbone is an EfficientNetB0 (237 layers). Furthermore, let $L_{i,j}$ represent a dense layer, where i is the layer number and j is the number of units. Then, in ISC_*class*_DL#1, the head has 5 layers. In the initial layer, L_{in}, the number of units equals the number of outputs of the backbone. The second dense layer has n units with X_2 activation function ($L_{2,n}$), a dense layer $L_{3,m}$, and a dense layer with 24 units ($L_{4,24}$), all with X_i activation functions (see Table 7). The last layer has 3 units with a softmax activation function (Ls).

(b) **Model DL#2:** the backbone is Xception (71 layers) and the head also has 5 layers. In the ISC_*class*_DL#2, the first dense layer has L_{in} units, equal to the number of outputs of the backbone. Then, it has a dense layer with n units ($L_{2,n}$). The third layer has also m units and $L_{3,m}$. Then a dense layer with 24 units ($L_{4,24}$). All layers have X_i activation functions. The last layer is the Ls.

(c) **Model DL#3:** the backbone is ResNet-50 (50 Layers) and the head has 4 layers. In the ISC_*class*_DL#3, the first dense layer is L_{in}, the second layer $L_{2,n}$, the third layer $L_{3,24}$, all with X_i activation functions, and the last layer is the Ls.

These architectures were tuned using the sub-data set ONMM. In this context, several hyperparameters were tested, such as the number of units, batch size, number of epochs, etc. (see Section 5 for the hyperparameters). The only fixed values were the penultimate (with 24 units) and the last (with 3 units) layers, and the softmax activation function. The reason

for the layer of 24 units is based on the authors' hypothesis derived from Plutchik's wheel of emotions. The authors hypothesize that due to the importance of color in sentiment analysis and the relation between emotion and sentiment, and once there are 24 emotions in Plutchik's wheel, it is expected that a layer of 24 units can help the network to learn those emotions and relate them with the three sentiments, which appears in the last layer. That is, it will allow the image classification of the sentiment into positive, negative, or neutral (reason for three units in the last layer) according to the responses of those "emotions".

It should be noted that although there are only three models here, more than 100 models have been tested for *class* ONMM, with different hyperparameters, optimizations, and backbones. No drop-out layers were used in these three models either, but they were also tested.

In the next step, the results from the three models are ensembled to obtain a final classification. The inputs of the ensemble sub-block will be the predictions made by the individual models, resulting from the softmax function, and the output will be the final image sentiment prediction. Ensembling leverages the idea that combining the strengths of multiple models can result in improved overall performance and accuracy, compared to using a single model, as well as a better generalization (reacts better to unseen data than single models), reduces "overfitting", and increases the robustness of the results obtained. For the aggregation of sentiment classification, the ensemble sub-block, was used:

(i) ***Random Forest* (ISC_*class*_RFa):** k estimators (see Section 5), Gini impurity as criteria function to measure the quality of split, and the minimum number of samples required to split an internal node was 2. The rest of the hyperparameters were set to the default values of the scikit-learn library (https://scikit-learn.org/, accessed on 1 August 2024) (v. 1.5).

(ii) ***Neural Network* (ISC_*class*_NNa):** three dense layers, where the first layer has nine units (L_{in}), then a n layer with m units ($L_{n,m}$) with X_i activation functions, and the third layer is the Ls. A search tuner function was used to find the best (hyper-)parameters for the proposed model. The only layer not found by the search tuner is the last, Ls layer, of three neurons, which uses the softmax activation function to obtain the probability of sentiment.

4.2. Text Sentiment Classification

In the *Text Sentiment Classification* block, the first step consists of converting the text (see Section 2) into structured data, in a way that can be used by ML methods. To achieve this, a Bag of Words was applied, see details in [3], and the ML models used are:

(a) ***Random Forest (TSC_RFc)***: created with k estimators and Gini impurity as criteria function to measure the quality of the splits. The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5).

(b) ***Neural Network (TSC_NNc)***: four dense layers where the first has 5000 units (L_{in}), then two layers with n units ($L_{2,n}$) units and m units ($L_{3,m}$) are added, with X_i activation functions. The last layer uses three units with a softmax activation function (Ls), used to predict the probability of a text carrying a positive, negative, or neutral sentiment.

(c) ***Natural Language Toolkit*** (TSC_NLTK): this method was used as a third model (NLTK available at: https://www.nltk.org/, accessed on 1 August 2024).

The block diagram of the proposed TSC is presented in Figure 5. For the aggregation of text sentiment classification (the ensemble sub-block) was used:

(i) ***Random Forest (TSC_RFt)*:** k estimators and "Gini criteria" function (as before). The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5);

(ii) ***Neural Network (TSC_NNt)*:** four dense layers, the first layer has nine units (L_{in}), then a dense layer with three units ($L_{2,n}$), a third layer with m units ($L_{3,m}$) all with X_i activation function, and the fourth layer is the Ls.

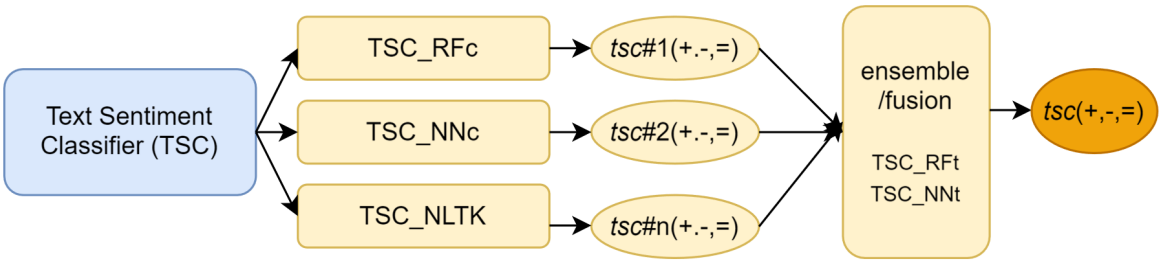

Figure 5. Block diagram of the Text Sentiment Classifier.

As mentioned, a search tuner function was used to find the best (hyper-)parameters for the proposed models (see Section 5).

4.3. Multimodal Sentiment Classification

The Multimodal Sentiment Classifier block, as largely mentioned in the text, is based on classifications resulting from text and image. Going to the block diagram in Figure 6, from the ISC model there are 36 output probabilities (4 *classes* × 3 individual models × 3 sentiments), from which the 9 output probabilities (3 individual models × 3 sentiments) are used from the individual models corresponding to the *class* of the image that is being analyzed. From the TSC, the 9 output probabilities (3 individual models × 3 sentiments) are also used. Also, *isc* and *tsc* sentiment classification are used to compute the *discrepancy*, which acts as a selector to compute the *msc* based on the ISC and TSC, as we will see next.

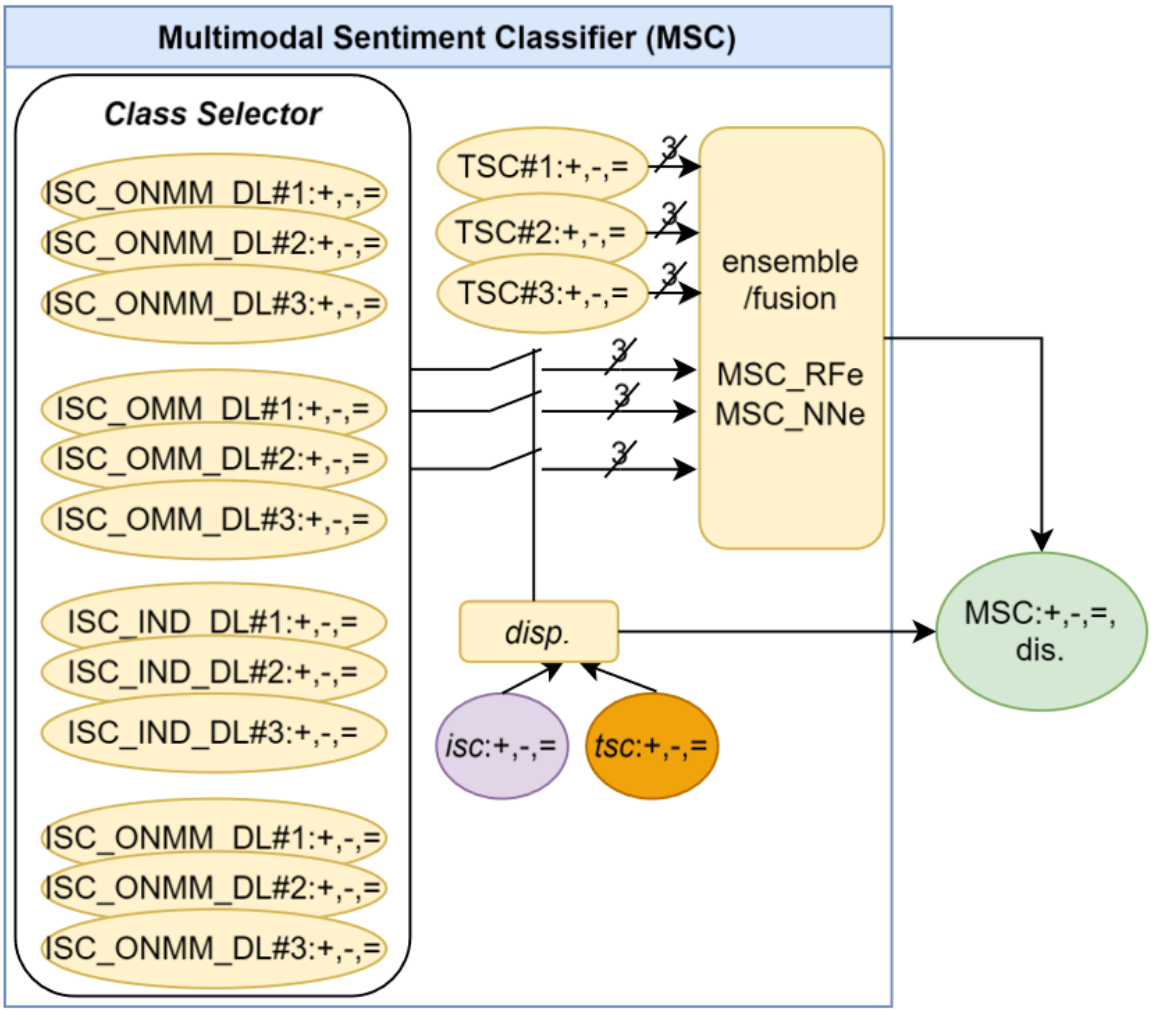

Figure 6. Block diagram of the Multimodal Sentiment Classifier.

The *discrepancy (dis.)* varies between 0 and 100, with 0 corresponding to the sentiment between the image and text being the same, and 100 for the sentiment between the image and text being different, and it is computed as follows:

(a) ***Similar sentiments***, i.e., *isc* = *tsc*, then *msc* = *isc* = *tsc*, and SCv = [*tsc*, *tsc*, *tsc*, *dis.* = 0] is the resulting output.

(b) ***Hypothetic different sentiments***, i.e., $isc \neq tsc$, then considering μ_I the average results (between 0–100) from all ISC models that returned the predicted sentiment, μ_T the average of all TSC models that returned the predicted sentiment, and a threshold $\mu_t = 85$ (empirically calculated), two cases can occur:

(i) ***The image and text have clearly different sentiments:*** If $\mu_I \geq \mu_t$ and $\mu_T \geq \mu_t$ then both ISC and TSC are considered to be certain of the predicted sentiment and, in this case, most probably the person who posted the text–image "intended" different sentiments. The ensemble block is not computed, resulting SCv = [*isc*, *tsc*, ×, *dis.* = 100].

(ii) ***The image and text have indeterminate sentiments***: The remaining cases. It means that the ensemble between text sentiments and image sentiments must be computed, once there is no certainty about what the person intended to post. Resulting in SCv = [*isc*, *tsc*, *msc*, *dis.*], where the $dis. = 100 - (\mu_I - \mu_T)/2$ if *tsc* < 50 or *isc* < 50, or $dis. = (\mu_I - \mu_T)/2$ for the remaining situations.

Because there is no balance in the sentiment classes data set and the number of samples is relatively small, once the B-T4Smultimodal (sub-)data set is used, the Stratified K-Fold cross-validation technique is used to train the MSC model (K = 5). The ensemble block is computed following the same principles presented before, namely:

(i) ***Random Forest (MSC_RFe)*:** k estimators, l minimum samples in a leaf, m minimum samples required to split the internal node, and Gini as criteria function. The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5);

(ii) ***Neural Network (MSC_NNe)*:** four dense layers, where the first layer has 18 units (L_{in}), a dense layer with n units ($L_{2,n}$), a third layer with m units ($L_{3,m}$), with X_i activation function, and the fourth layer has the Ls with softmax as the activation function.

The following section details the tests conducted and their results, followed by a discussion of the work undertaken.

5. Tests, Results, and Discussion

The tests and discussion are organized into four sections as follows: one focusing on the ISC, another on the TSC (presenting some results achieved in [3], to better understand the present paper), a section on the combination of text–image (post) following the MSC, and a final discussion.

5.1. Image Sentiment Classification

The procedure implemented to evaluate ISC_*class* (individual) models is divided into two steps as follows: (1) evaluate image sentiment classification on test data of Flickr sub-data sets and (2) evaluate the models in the SIS & ISP data sets. The models were trained and tested using the Kaggle's platform (with 14.8 GB RAM e GPU T4 x2).

All ISC_*class* (individual) models were trained with 70% of the samples, being from the remaining 10% to validate the model, and 10% for testing. For the ensemble model evaluation, 30% of the previous samples were considered, the ones not used for training the individual (*class*) models (namely, 10% from the validation, 10% from the test, and the 10% remaining data). From these 30%, 80% of the samples were used for ensemble training and 20% for testing.

To validate the hypothesis that four ISC models, one per class, work better than a single model trained with all images, a Holistic Image Sentiment Classifier (HISC) was trained using the Flickr sub-data set listed as "balanced" in Table 2 (70% of samples for training, 10% for validation, and 20% for testing). The HISC uses the 3 DL blocks/models

and the same previously stated ensemble strategies to predict the sentiment of an image. The difference lies in the training samples as follows: ISC models divide the samples by classes, whereas HISC uses all samples together.

Table 7 shows the model's hyperparameters. It is important to emphasize that the classifier models for each *class* and the HISC use the same hyperparameters. In more detail, the first column of the table shows the used models, and the second column summarizes the number of units used in the layers (resulting from the optimizations), as well as the number of estimators. The remaining columns present how many layers the backbone was trained with, as well as the hyperparameters used to train with the new data. There is only one exception—for the ensemble models that were built using a neural network, a search tuner was used to choose the network settings, and hyperparameters, to obtain the best possible result.

Table 7. ISC_*class*, HSC, and ISC individual and ensemble models backbones and hyperparameters.

Model	Number Units/Estimators	Backbone (Layers Trained)	Hyperparameters	Activation Function
Model DL#1	n = 1024 m = 512	EfficientNetB0 (none)	Opt: Adam (1×10^{-4}) Epochs: 20 Batch size: 32	Xi = ReLU i = {1, . . ., 5}
Model DL#2	n = 1024 m = 512	Xception (8)	Opt: Adam (1×10^{-4}) Epochs: 20 Batch size: 32	Xi = ReLU i = {1, . . ., 5}
Model DL#3	n = 512	ResNet50 (12)	Opt: Adam (1×10^{-4}) Epochs: 20 Batch size: 32	Xi = ReLU i = {1, . . ., 4}
RFa	k = 100 (est.)	-	-	-
NNa	decided by the optimizer	-	Batch size: 32	-

For the random forest ensemble models, the Gini impurity was the criteria function used to measure the quality of split and the minimum number of samples required to split an internal node was two (the rest of the hyperparameters were set to the default values of the scikit-learn library). The model was initially trained in the following two ways: (1) by using the sentiment values of −1 (negative), 0 (neutral), and 1 (positive), predicted by the individual models; and (2) by using the sentiment-predicted probabilities from the individual models. After analysis, option (2) was chosen, because the results indicated a significant improvement when comparing to option (1), as we can see in Figure 7. The figure shows ISC_OMM models' confusion matrices. On the left, the model uses three inputs (option 1) from the direct sentiments of the individual models, achieving an accuracy of 80.81%, and on the right, the model uses nine inputs corresponding to the probabilities of the sentiment (2), achieving an accuracy of 82.21%.

The accuracies for the different sub-data sets and ensemble are presented in Table 8, where the first column indicates the class, the second the model used, and the remaining columns the accuracy for the Flickr sub-data sets and SIS & ISP sub-data sets. When evaluating the models, the ensemble of individual models (ISC) presented better results than the holistic model (HISC) for the Flickr and SIS & ISP sub-data sets, as marked in gray in the table. For Flickr sub-data sets, ISC achieved 76.45% compared to HISC's 65.97%. Similarly, for the SIS & ISP sub-data set, ISC achieved 53.31% compared to HISC's 47.80%.

Going down to the individual models, the results for the Flickr sub-data sets are all above 60%, except for two cases of model DL#3. For the SIS & ISP data sets, the results are less favorable, because there is some accuracy below 50% (in blue), but always above 44%. It is important to remember that humans did the labeling of SIS & ISP, unlike Flickr. For the Flickr sub-data sets, the individual ISC models achieved accuracies of 84.54%, 83.21%, 68.53%, and 84.79% for the classes ONMM, OMM, IND, and IOwPB, respectively. For

SIS & ISP, the accuracies were 61.53%, 56.92%, 51.62%, and 49.06% for the same classes, respectively. The best result for each ISC class, HISC and ISC is marked in bold.

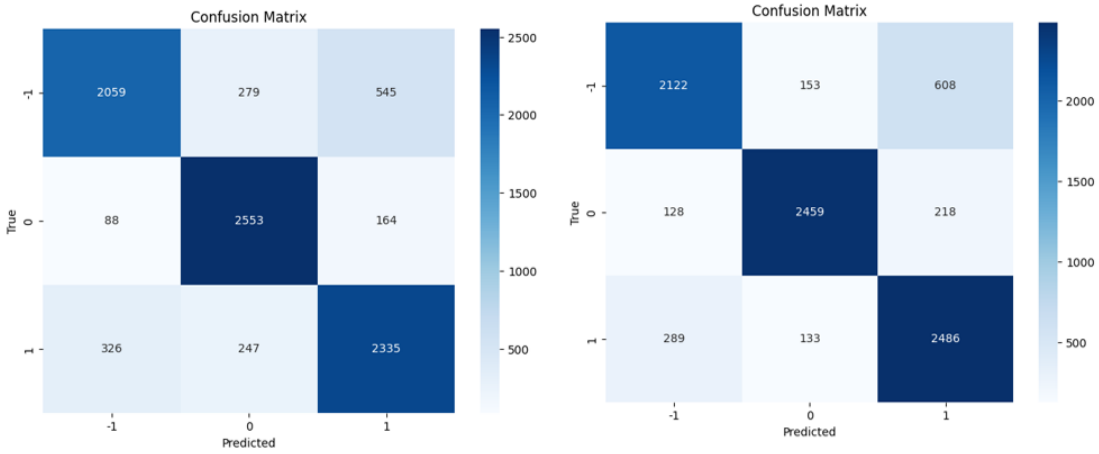

Figure 7. ISC_OMM models' confusion matrices: on the **left**, the model uses three inputs (which are the direct sentiments of the individual models), while the model on the **right** uses nine inputs (corresponding to the probabilities of the predicted sentiment).

Table 8. ISC_*class*, ISC, and HISC individual and ensemble models accuracy.

Class	*Model*	*Flickr Sub-Data Sets Accuracy*	*SIS & ISP Accuracy*
ISC_ONMM	Model DL#1	80.30%	53.82%
	Model DL#2	77.87%	53.61%
	Model DL#3	71.09%	50.23%
	RFa	**84.54%**	61.16%
	NNa	83.34%	**61.53%**
ISC_OMM	Model DL#1	79.31%	50.36%
	Model DL#2	76.29%	53.40%
	Model DL#3	55.26%	47.29%
	RFa	**83.21%**	55.93%
	NNa	83.09%	**56.92%**
ISC_IND	Model DL#1	64.53%	46.31%
	Model DL#2	61.37%	49.65%
	Model DL#3	55.47%	42.93%
	RFa	**68.53%**	50.21%
	NNa	67.47%	**51.62%**
ISC_IOwPB	Model DL#1	81.46%	44.85%
	Model DL#2	77.65%	46.99%
	Model DL#3	64.75%	46.41%
	RFa	**84.79%**	**49.06%**
	NNa	84.52%	48.77%

Table 8. *Cont.*

Class	*Model*	*Flickr Sub-Data Sets Accuracy*	*SIS & ISP Accuracy*
HISC	Model DL#1	63.46%	**45.25%**
	Model DL#2	61.85%	**44.31%**
	Model DL#3	54.74%	**41.97%**
	RFa	**65.97%**	47.21%
	NNa	64.45%	**47.80%**
ISC	RFa	**76.45%**	**53.31%**
	NNa	69.92%	48.51%

Another strong indicator is that the test accuracies are very close for the three individual models per class, and this consistency suggests that they are all effectively capturing the underlying patterns in the data, leading to a somehow similar performance. Additionally, the ensemble models play an important role because they enhance the accuracy of the best individual models by ~4% to ~5%, consistently in all classes. The same occurs for the global models. In addition, the ensembles offer stability in their responses as they are based on the multiple analyses of the other (individual) models. Figure 8 highlights the close test accuracies of the individual models, using the example of ISC_NMMO confusion matrices for models DL#1, DL#2, and DL#3 (top line, from left to right), and ensembles RFa and NNa (bottom line, left to right).

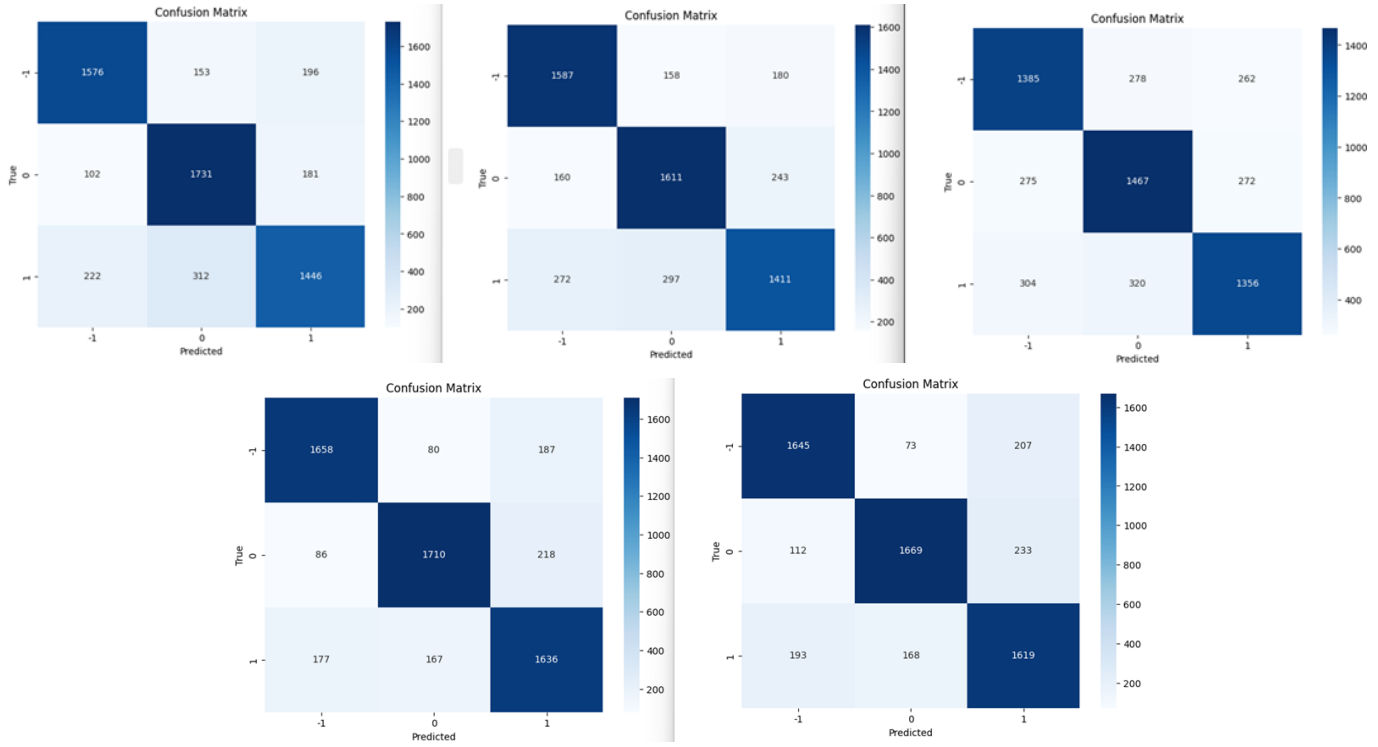

Figure 8. ISC_NMMO confusion matrices for models DL#1, DL#2, and DL#3 (**top** line, from **left** to **right**), and ensembles RFa and NNa (**bottom** line, **left** to **right**).

Once again, it is important to stress that the same backbones, hyperparameters, and parameters (configurations) were used for all individual models and ensembles, ISC_*class*, ISC, and HISC model. Fine-tuning the models will achieve (for sure) better performance (accuracy); however, that was not the goal of this study, because the aim was to compare the models using consistent configurations.

5.2. Text Sentiment Classification

For the development of the individual and ensemble TSC models, the Colab platform was used with 12.7 GB of RAM and 107.7 GB of disk space. Table 9 shows the configurations and accuracy of the individual and ensemble models for TSC that were trained using the B-T4SAtext data set. It also displays the accuracy of the individual models while highlighting the effectiveness of ensemble models in stabilizing and enhancing their accuracy. For more details about this section, please see [3]. A final observation, it is important to highlight that the TSC model, utilizing neural networks, achieves the best result with an accuracy of 92.10% (marked in bold).

Table 9. TSC parameters, hyperparameters, and accuracy of the models.

Model	Number Units/Estimators	Hyperparameters + Activation Function		Accuracy
TSC#1 (RF)	k = 100 (est.)	—	—	90.10%
TSC#2 (NN)	n = 300 m = 100	Opt: SGD (1×10^{-2}) Epochs: 10 Batch size: 8	X_i = ReLU i = {1, . . ., 3}	88.55%
TSC#3 (NLTK)	—	—	—	84.34%
TSC–RFt	k = 100 (est.)	—	—	91.95%
TSC–NNt	n = 100 m = 20	Opt: SGD (1×10^{-2}) Epochs: 10 Batch size: 2	X_i = ReLU i = {1, . . ., 3}	**92.10%**

5.3. Multimodal Sentiment Classification

The MSC was trained in the Kaggle's platform (with 14.8 GB RAM e GPU T4 x2) using B-T4SAmultimodal and a stratified K-fold cross-validation technique, focusing only on samples where the image and the text share the same sentiment (273 samples). Before entering more details about MSC results, characterizing the samples per discrepancy is important.

From the B-T4SAmultimodal data set (classified by humans), the MSC framework using the specification stated in Section 4.3 separates samples into the following three categories: those where the image and text represent the same sentiment (38.60%), those where there is uncertainty about their alignment (55.66%), and those where the sentiments are clearly different (see Table 10).

Table 10. Number of samples per sentiment discrepancy.

Discrepancy	Number of Samples	Percentage of Samples
dis. = 0	242	38.60%
0 < *dis.* < 100	**349**	**55.66%**
dis. = 100	36	5.74%

This means that 38.60% (242 from 627 samples) plus 5.74% (36 from 627 samples) are already classified (no additional processing is required). The former because *isc* equals *tsc*, and consequently *msc* equals *tsc* returning SCv = [*tsc*, *tsc*, *tsc*, 0]; the latter because there is no possible classification for *msc*, once the model is completely sure of the sentiment for ISC and for TSC, and they are different, meaning SCv = [isc, tsc, ×, 100]. The framework only needs to do the inference for the samples that have 0 < *dis.* < 100 (marked in bold), in the present data set 55.66% (i.e., 349 from the 627 samples).

Table 11 shows the parameters, hyperparameters, and accuracies per discrepancy of the MSC model. In more details, if *dis.* = 0 (ISC and TSC report the same sentiment), then the framework reports a 100% accuracy, i.e., these samples need no further computation, once

isc equals *tsc*, so it just checks the sentiments against the ground truth (no inference was carried out). For $0 < dis. < 100$, the best result was achieved for the ensemble using random forest, with an accuracy of 64.18% (using NN, the result is similar with a difference ~4%).

Table 11. MSC parameters, hyperparameters, and accuracy.

Model MSC		Number Units/Estimators	Hyperparameters + Activation Function		Accuracy (Samples)
RFe	$dis. = 0$		—		100.00% (242/242)
	$0 < dis. < 100$	k = 150 (est.) l = 4 (samples) m = 5 (samples)	—	—	64.18% (224/349)
				MSC Model (using RFe)	**78.84% (446/591)**
NNe	$dis. = 0$		—		100.00% (242/242)
	$0 < dis. < 100$	n = 466 m = 24	Opt: Adam (7.88×10^{-4}) Epochs: 20 Batch size: 4	X_i = ReLU i = {1, . . ., 3}	60.17% (210/349)
				MSC Model (using NNe)	76.48% (452/591)
N.A.	$dis. = 100$		—		Not applicable

At this point, it is again important to stress how the accuracy of MSC is computed, because it aggregates results from the ISC model and TSC model, and checks these predictions against the ground truth, once the image and text reflect or could reflect different sentiments. Following this, the good results in the accuracy of the MSC model (above 78%) are justified by the fact that (1) the final accuracy is computed between the samples for $dis. = 0$ (100% accuracy; no ensemble model was applied) and the samples from $0 < dis. < 100$ (64.18% accuracy), which returns the final accuracy of the model of 78.84%; (2) they are exclusively employed to determine and enhance sentiment accuracy based on the results of the 12 individual models (3 TSC and 3 ISC_*class*); and (3) for the remaining samples, when $dis. = 100$, the ensemble is not applied, once it is not computed the *msc*, i.e., SCv = [*isc*, *tsc*, ×, 100].

It is important to note that the number of samples existing in B-T4SAmultimodal to train and test the MSC model is very low, and this can bias the results. Nevertheless, the results prove that it is possible to detect if an image and text share the same sentiment, as well as to identify instances where they have completely different sentiments. Furthermore, the framework effectively combines images and text that may or may not have different sentiments, achieving a good result when comparing with the classification performed by humans.

Despite this, it is mandatory for future works to exponentially increase the number of samples classified by humans (a second version of the B-T4SAmultimodal) and make it public for other authors to test the results against these initial baseline results. The present data set will be available at https://osf.io/institutions/ualg (accessed on 1 August 2024).

5.4. Discussion

One of the difficulties of this research is the nonexistence of a sentiment data set classified by humans that can validate the main research goals. Specifically, these objectives are (1) developing a single hyperparameter image classification sentiment model that performs well across different environments, including validating that the classification accuracy supported on four categories is better than using a single category (including all images); and (2) the development of a framework that combines image and text sentiment classification. The framework must return a multimodal sentiment classification along with the discrepancy metric, which includes the idea that text and image can only be combined if both return the same sentiment or if both return uncertain sentiments.

In the latter case, one text/image can complement the other to achieve the final sentiment classification. When both return different sentiments, they should not be joined. That is, when conflicting sentiments occur, empirically, it is possible to say that the person posted the text with a sentiment and used the image only for illustration purposes, or the opposite, they posted the image with sentiment and the text is only to "frame" the image. This needs further research, including how these posts can/should be used by managers who receive this information to manage their companies or platforms.

Returning to the nonexistence of a data set with ground-truth sentiment for text and images from posts, the solution adopted in this paper to mitigate this problem was to use multiple data sets and sub-data sets. While this is not an ideal solution, it effectively supports the scientific goals of the paper.

For the TSC, a single data set was used that filled all the requirements. For the ISC, the initial data set, classified automatically, was divided into five classes. From these five classes, the one depicting in the foreground (semi-)frontal humans was discarded due to the following three main reasons: (1) this class is certainly the most analyzed in the literature, presenting excellent models that validate the sentiment, with most of the cases using the human face; (2) text and image sentiment usually, in this case, are very similar; and (3) in a post, indoor and outdoor scenes many times are used with different purposes, such as transmitting a sentiment, being ironic, for illustration, etc. Usually, images with faces transmit a specific emotion in posts. For the MSC, it was not possible to find a data set that had human-annotated ground-truth classification for the image, the text, and the combined image–text.

When developing the authors' data set, it was validated that these three possibilities are very crucial. For example, a person might classify a text with a positive sentiment and an image with a negative sentiment. However, when viewing both the image and the text together, the overall sentiment might be different, such as neutral. In reality, the last one is the one that is meaningful for (training) the MSC.

In consequence, despite the authors' data set having a limited number of samples, there were enough to validate the planted concept. In the authors' data set, 55.66% of the samples fall in the mentioned situations, which is not an insignificant number. In summary, there are (sub-)data sets that fill the requirements to validate the paper's goals. Nevertheless, for future work, all the data sets need more samples classified by humans. In addition, the feedback of the post sentiment to the managers who receive the information to manage their companies or platforms should be the image sentiment, the text sentiment, the combined (text–image) sentiment if it exists, and the discrepancy between the image and the text sentiment (SCv = [*isc*, *tsc*, *msc*, *dis.*].

Also related to the data set are the failures of the models, which are detected mostly in the ISC. The classification of the images where the ISC models detect positive sentiment and the image is negative or select negative and the image is positive (see Figures 7 and 8) is mostly due to the data sets not being balanced and the difficulty of having a clear classification of the "neutral" image. Figure 9 shows examples of images where different persons gave different classifications, going from the positive, to the neutral, to the negative. Take the example on the right; for someone who appreciates history, the knight's armor evokes a positive sentiment. For others, the complete scene might be perceived as neutral. However, for different individuals, the image might conjure thoughts of war and medieval times, leading to a negative sentiment. So, to develop a more robust model, the data set must have not only the human classification, but also the characteristics of the human classifier, and the model should also account for that information. For more information about this subject, please see [38].

Similar issues arise in the MSC due to the lack of a comprehensive data set. When evaluating both text and images together, human classifiers sometimes base their decisions on group considerations or personal biases—such as a preference for images over text, or vice versa—leading to inconsistent classifications for the same text–image pair. The solution to this problem would be to have a large number of samples for training the

classifier. However, such a data set is not currently available and should be addressed in future work.

Figure 9. Examples of images where the human classification presented more doubts.

All the above leads us to compare the current framework with state-of-art models. In terms of results, the present framework achieves 78.84% accuracy, which places it among the top results presented in Table 1. It is possible to say that the present framework is at the top of the results, or at least it is in line with the state-of-the-art results. However, a direct comparison is not a completely fair one for the present framework or for the models presented in the table, because (1) all use different data sets or sub-data sets; (2) some are trying to achieve the best result possible, while the present framework is proving/showing/presenting a concept/idea/goal; (3) only some models are trained with human-classified data; and (4) some models train image and text classifiers separately and then combine them, without using data where both the image and text are classified together by humans. The bottom line is, at the moment, the results in this specific area cannot be comparable until the authors from different publications use the same procedure. Nevertheless, as mentioned, it is possible to validate that the present results are completely in line with the best state-of-the-art results.

Examining the results of the individual models in detail, Tables 8 and 9 present the results of the ISC and TSC individual models, and respective ensembles. It can be seen (as already discussed) that the ensembles return better results than the individual models (which was one of the research questions). Not mentioned, but also important to stress, is that each (individual) ISC model uses different backbones, and, consequently, extracts different features from the image. The same occurs for the text, i.e., TSC_RFc and TSC_NNc use the same features, which are different from TSC_NLTK.

Another important point to reinforce is the training of the models. As mentioned, the ISC_ONMM was tuned using more than 100 variations of hyperparameters of the network's head. Each training procedure took around 4 h, for each variation of parameter, in the Kaggle's platform. Having this classifier fine-tuned and considering this category the more generic one, it was considered (hypothesized) that the same parameters could be used for all the categories, reducing the time used for fine-tuning each category (ISC_*class*). Nevertheless, it is clear that if all the models were all fine-tuned, then better results would have been achieved. Within this principle, the framework presents low-complexity models for which it is easy to specify the hyperparameters (all are the same between different ISC_*class*) and, possibly, have a faster training procedure than other more complex models presented in the state of the art (see Section 2). In terms of the training time, for the class ISC_ONMM, the train takes approximately 4 h. In categories with fewer samples, it takes less time to train, and more samples take more time, which varies from around 1 h to around 7 h (for the HISC model). The training of the ensembles is quite fast, taking a few minutes.

Finally, it is crucial to emphasize (as already mentioned) that tests and results are reported for each (individual) model/classifier—the module in the overall framework (see Tables 8 and 9) as well as for each ensemble classifier module. Nevertheless, no ablation study was conducted, because the focus was on the comparison of the models and the ensemble models. This means, for instance, that the impact of employing combinations of two (individual) models for the ensemble rather than the three models was not investigated

(for each ensemble). In future work, we intend to conduct a full ablation study when utilizing increasingly complex individual modules.

6. Conclusions

This work presented research on text and image sentiment classification, especially in social media posts related to "indoor", "man-made outdoors", "non-man-made outdoors", and "indoor/outdoor with persons" environments. The study and analysis demonstrated the effectiveness of the proposed class-specific and holistic image classifiers and text classifiers in predicting sentiments, highlighting their potential applications and implications.

The preprocessing techniques, the incorporation of deep learning models, and the advanced feature extraction techniques used led the Multimodal Sentiment Classifier framework to obtain a high accuracy in sentiment classification from text–image posts, as well as a very consistent prediction of image and text representing the same sentiment and indeterminate sentiments.

Experiments on detecting sentiments in images showed promising results, demonstrating that the system can classify sentiments based on objects, colors, and other aspects present in an image. Additionally, scene-specific image models obtained higher accuracies due to their ability to capture specific details in contexts, while the holistic model offers a lower accuracy but a higher versatility, once it is not needed to use preclassification techniques to classify the images (the last being class segmentation, which is out of the focus of this paper).

Finally, the ensemble models allowed the system to leverage the complementary information provided by the textual and visual models, which leads to better performance. This is very significant when a multiple model approach is used, in this case in sentiment analysis tasks.

In summary, this study contributed to the understanding of sentiment detection from data present in text and images. Although the accuracy of the system can be improved, the potential of the models has been demonstrated. More significantly, it introduced a framework that has not yet been published in the literature. The framework utilizes separate sentiment models for text and image; these models are only combined at the end if they convey the same sentiment or if there are uncertainties about the sentiment, allowing one to enhance the other. In the cases where the image and language clearly convey different sentiments, they should not be merged. An empirical conclusion is that the user only intended to illustrate the text or vice versa, that is, the text was only used to frame the image. This paper presents an initial approach to the discussion of this problem, that in future works needs to be deepened. Continuing to improve and refine benchmark sentiment classification can open new possibilities to enable more sophisticated and nuanced sentiment analysis in interfaces and/or robots.

Looking ahead, there are several directions for future research. The focus should be on improving the sentiment detection model in images and obtaining more and better image sentiment classification data sets. We also intend to train and test models for images of scenes not mentioned during this research, to further increase the effectiveness of the final model that contains the responses from each environment-specific classifier.

Author Contributions: Conceptualization, N.S., P.J.S.C. and J.M.F.R.; methodology, N.S.; software, N.S.; validation, N.S., P.J.S.C. and J.M.F.R.; formal analysis, N.S., P.J.S.C. and J.M.F.R.; investigation, N.S.; resources, N.S., P.J.S.C. and J.M.F.R.; data curation, N.S., P.J.S.C. and J.M.F.R.; writing—original draft preparation, N.S.; writing—review and editing, N.S., P.J.S.C. and J.M.F.R.; visualization, N.S., P.J.S.C. and J.M.F.R.; supervision, P.J.S.C. and J.M.F.R.; project administration, J.M.F.R.; funding acquisition, P.J.S.C. and J.M.F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020) and ref. UIDP/04516/2020 (https://doi.org/10.54499/UIDP/04516/2020) with the financial support of FCT.IP.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available data sets were analyzed in this study. This data can be found here (accessed on 27 June 2024): Flickr data set: https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html; SIS data set: https://datasets.simula.no/image-sentiment/; ISP data set: https://data.world/crowdflower/image-sentiment-polarity; T4SA: http://www.t4sa.it/#dataset. There are some private data for part of the B-T4SAmultimodal sub-data set, that will be public during the next year at: https://osf.io/institutions/ualg/.

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

References

1. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
2. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83*, 19–52. [CrossRef]
3. Silva, N.; Cardoso, P.J.S.; Rodrigues, J.M.F. Sentiment Classification Model for Landscapes. In Proceedings of the 18th International Conference on Universal Access in Human-Computer Interaction, Part of HCI International, Washington, DC, USA, 29 June–4 July 2024.
4. Ramos, C.M.Q.; Cardoso, P.J.S.; Fernandes, H.C.L.; Rodrigues, J.M.F. A Decision-Support System to Analyse Customer Satisfaction Applied to a Tourism Transport Service. *Multimodal Technol. Interact.* **2023**, *7*, 5. [CrossRef]
5. Cardoso, P.J.S.; Rodrigues, J.M.F.; Novais, R. Multimodal Emotion Classification Supported in the Aggregation of Pre-trained Classification Models. In *Computational Science*; Mikyška, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M., Eds.; Springer: Cham, Switzerland, 2023; LNCS Volume 10477. [CrossRef]
6. Novais, R.; Cardoso, P.J.S.; Rodrigues, J.M.F. Emotion Classification from Speech by an Ensemble Strategy. In Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, Lisbon, Portugal, 31 August–2 September 2022; Association for Computing Machinery: New York, NY, USA, 2023; pp. 85–90. [CrossRef]
7. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**, *3045*, 1195–1215. [CrossRef]
8. Ruan, S.; Zhang, K.; Wu, L.; Xu, T.; Liu, Q.; Chen, E. Color Enhanced Cross Correlation Net for Image Sentiment Analysis. *IEEE Trans. Multimed.* **2021**, *26*, 4097–4109. [CrossRef]
9. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From Facial Expression Recognition to Interpersonal Relation Prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569. [CrossRef]
10. Ekman, P. Are there basic emotions? *Psychol. Rev.* **1992**, *99*, 550–553. [CrossRef] [PubMed]
11. Noroozi, F.; Corneanu, C.A.; Kaminska, D.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on Emotional Body Gesture Recognition. *IEEE Trans. Affect. Comput.* **2021**, *12*, 505–523. [CrossRef]
12. Nandwani, P.; Verma, R. A Review on Sentiment Analysis and Emotion Detection from Text. In *Social Network Analysis and Mining*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 11. [CrossRef]
13. Ortis, A.; Farinella, G.M.; Battiato, S. An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. In Proceedings of the 16th International Joint Conference on e-Business and Telecommunications, Prague, Czech Republic, 26–28 July 2019; SciTePress: Setúbal, Portugal, 2019; pp. 296–306. [CrossRef]
14. Fugate, J.M.B.; Franco, C.L. What Color is Your Anger? Assessing Col-or-Emotion Pairings in English Speakers. *Front. Psychol.* **2019**, *10*, 206. [CrossRef] [PubMed]
15. Amencherla, M.; Varshney, L.R. Color-Based Visual Sentiment for Social Communication. In Proceedings of the 15th Canadian Workshop on Information Theory (CWIT), Quebec City, QC, Canada, 11–14 June 2017. [CrossRef]
16. Peng, Y.F.; Chou, T.R. Automatic Color Palette Design Using Color Image and Sentiment Analysis. In Proceedings of the IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019. [CrossRef]
17. Plutchik, R. Chapter 1—A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion*; Plutchik, R., Kellerman, H., Eds.; Academic Press: Cambridge, MA, USA, 1980; pp. 3–33. [CrossRef]
18. Munezero, M.; 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]
19. Gaspar, A.; Alexandre, L.A. A multimodal approach to image sentiment analysis. In Proceedings of the Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, 14–16 November 2019; Proceedings, Part I 20. pp. 302–309.
20. Vadicamo, L.; Carrara, F.; Cimino, A.; Cresci, S.; Dell'Orletta, F.; Falchi, F.; Tesconi, M. Cross-media learning for image sentiment analysis in the wild. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 308–317. Available online: http://www.t4sa.it (accessed on 1 August 2024).
21. De Oliveira, W.B.; Dorini, L.B.; Minetto, R.; Silva, T.H. OutdoorSent: Sentiment Analysis of Urban Outdoor Images by Using Semantic and Deep Features. *ACM Trans. Inf. Syst.* **2020**, *38*, 23. [CrossRef]

22. Chatzistavros, K.; Pistola, T.; Diplaris, S.; Ioannidis, K.; Vrochidis, S.; Kompatsiaris, I. Sentiment Analysis on 2D Images of Urban and Indoor Spaces Using Deep Learning Architectures. 2022. Available online: https://www.mturk.com/ (accessed on 1 August 2024).
23. 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. [CrossRef] [PubMed]
24. Du, Y.; Liu, Y.; Peng, Z.; Jin, X. Gated Attention Fusion Network for Multimodal Sentiment Classification. *Knowl.-Based Syst.* **2022**, *240*, 108107. [CrossRef]
25. Das, R.; Singh, T.D. Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 161. [CrossRef]
26. Chen, D.; Su, W.; Wu, P.; Hua, B. Joint multimodal sentiment analysis based on information relevance. *Inf. Process. Manag.* **2023**, *60*, 103193. [CrossRef]
27. Yadav, A.; Vishwakarma, D.K. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 15. [CrossRef]
28. Kumar, P.; Malik, S.; Raman, B.; Li, X. CMFeed: A Benchmark Dataset for Controllable Multimodal Feedback Synthesis. *arXiv* **2024**, arXiv:2402.07640.
29. Miah, M.S.U.; Kabir, M.M.; Sarwar, T.B.; Safran, M.; Alfarhood, S.; Mridha, M.F. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci. Rep.* **2024**, *14*, 9603. [CrossRef] [PubMed]
30. Yang, H.; Zhao, Y.; Wu, Y.; Wang, S.; Zheng, T.; Zhang, H.; Che, W.; Qin, B. Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. *arXiv* **2024**, arXiv:2406.08068.
31. Deng, Y.; Li, Y.; Xian, S.; Li, L.; Qiu, H. MuAL: Enhancing multimodal sentiment analysis with cross-modal attention and difference loss. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 31. [CrossRef]
32. Mao, R.; Liu, Q.; He, K.; Li, W.; Cambria, E. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Trans. Affect. Comput.* **2022**, *14*, 1743–1753. [CrossRef]
33. Hu, H. A Vision-Language Pre-training model based on Cross Attention for Multimodal Aspect-based Sentiment Analysis. In Proceedings of the 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 19–21 April 2024; pp. 370–375. [CrossRef]
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
35. CrowdFlower. Image Sentiment Polarity. 2015. Available online: https://data.world/crowdflower/image-sentiment-polarity (accessed on 10 May 2024).
36. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
37. 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.
38. Rodrigues, J.M.F.; Cardoso, P.J.S. Body-Focused Expression Analysis: A Conceptual Framework. In *Universal Access in Human-Computer Interaction*; HCII 2023. Lecture Notes in Computer Science; Antona, M., Stephanidis, C., Eds.; Springer: Cham, Switzerland, 2023; Volume 14021. [CrossRef]